HENRY

Introducción a Machine Learning

Data Science

Agenda

- Introducción al Machine Learning
- Tipos de aprendizaje
- Flujo de Trabajo en ML
- Transformación de Datos
- Scikit-Learn
- Correlación
- Regresión Lineal

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

- ➜ Entender el concepto de Aprendizaje supervisado Vs. Aprendizaje no supervisado.
- ➜ Conocer el Flujo de Trabajo en ML.
- ➜ Identificar falacias en los Datos.
- ➜ Conocer qué Transformaciones se pueden hacer sobre los Datos.
- ➜ Conocer Scikit-Learn.
- ➜ Comprender las ventajas de Reescalar los Datos.
- ➜ Comprender el concepto de Correlación.
- ➜ Comprender el funcionamiento del algoritmo Regresión Lineal.

Al **finalizar** cada uno de los temas, tendremos un **espacio de consultas**.

Hay un **mentor** asignado para responder el **Q&A**.

¡Pregunta, pregunta, pregunta! :D

Introducción a **Machine Learning**

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.

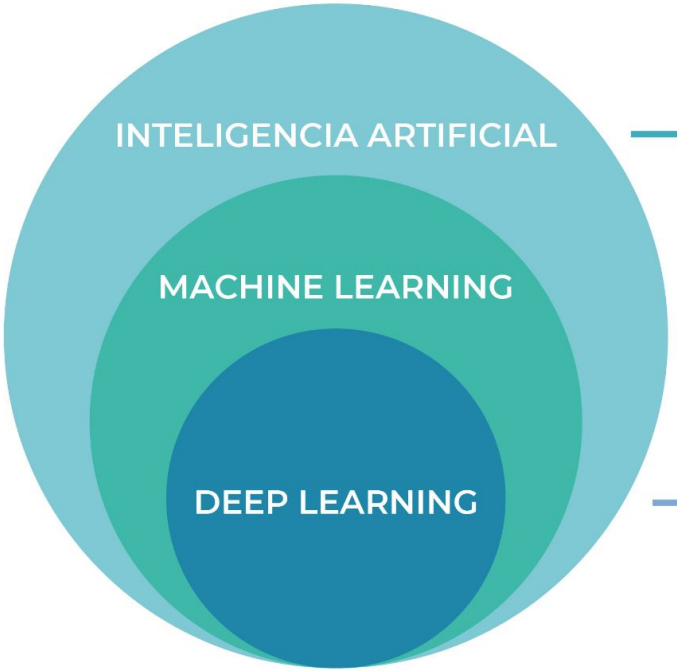
Inteligencia artificial

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



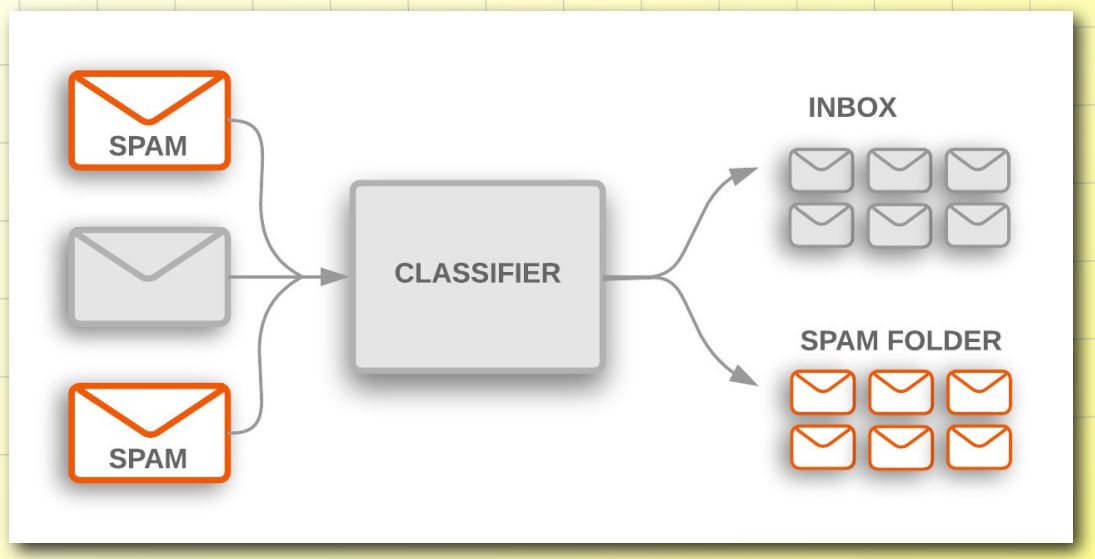
Cualquier técnica que permite a los ordenadores imitar el comportamiento humano.

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.

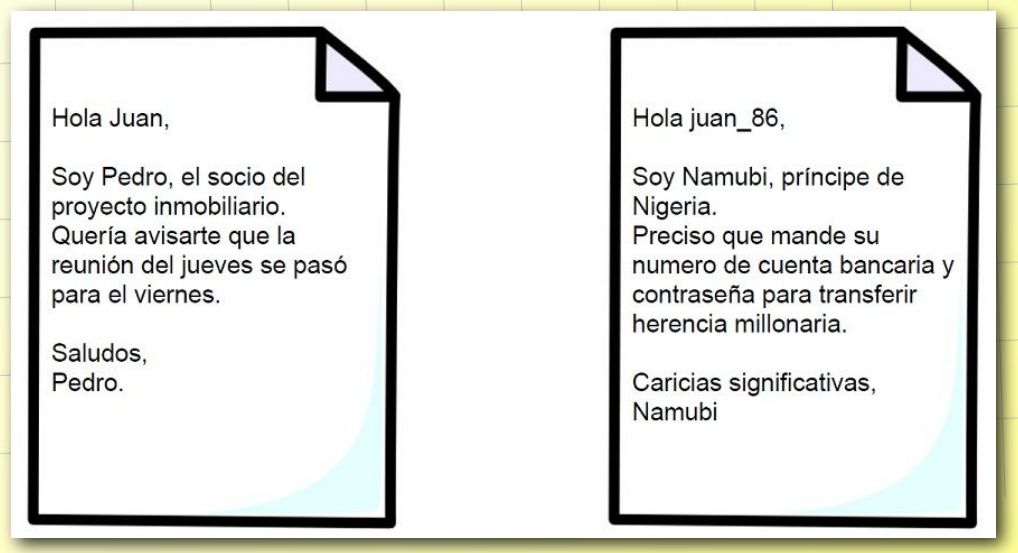
Ejemplo clásico

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

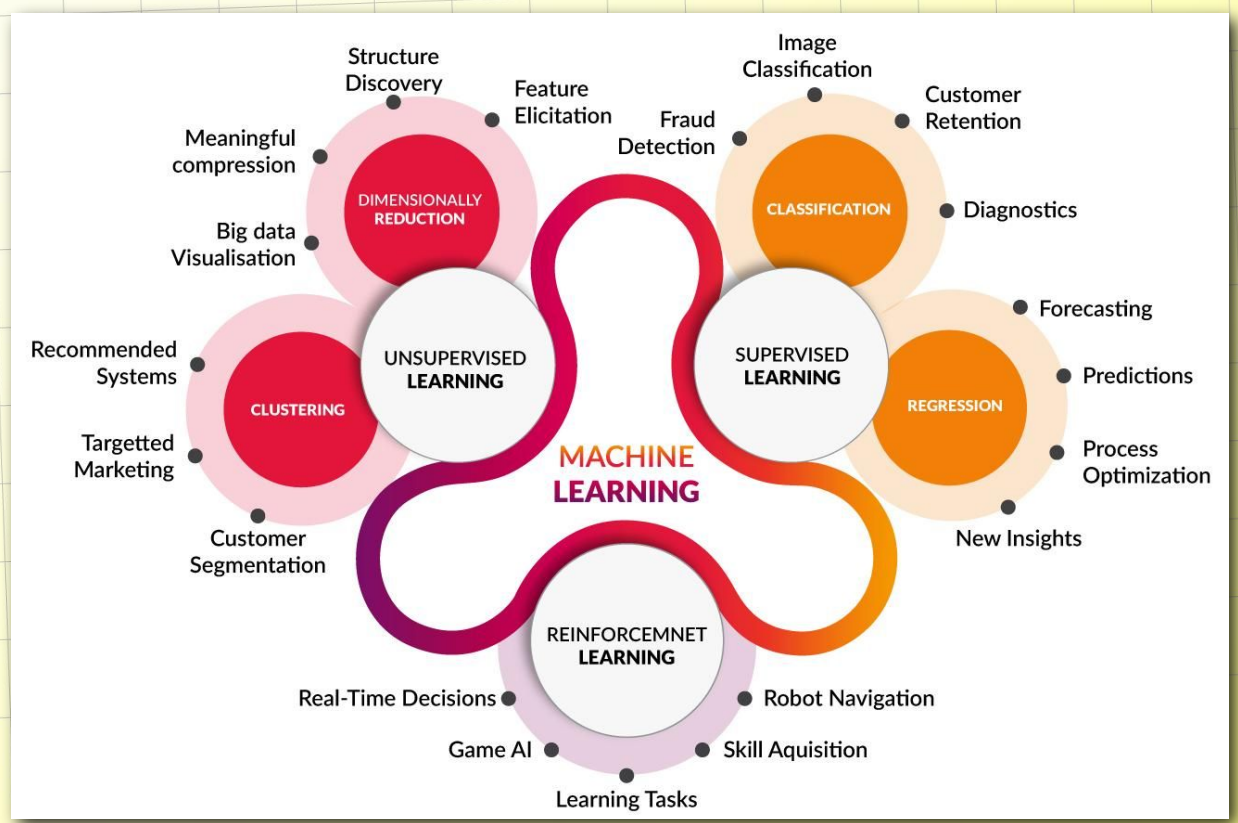


Ejemplo clásico

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

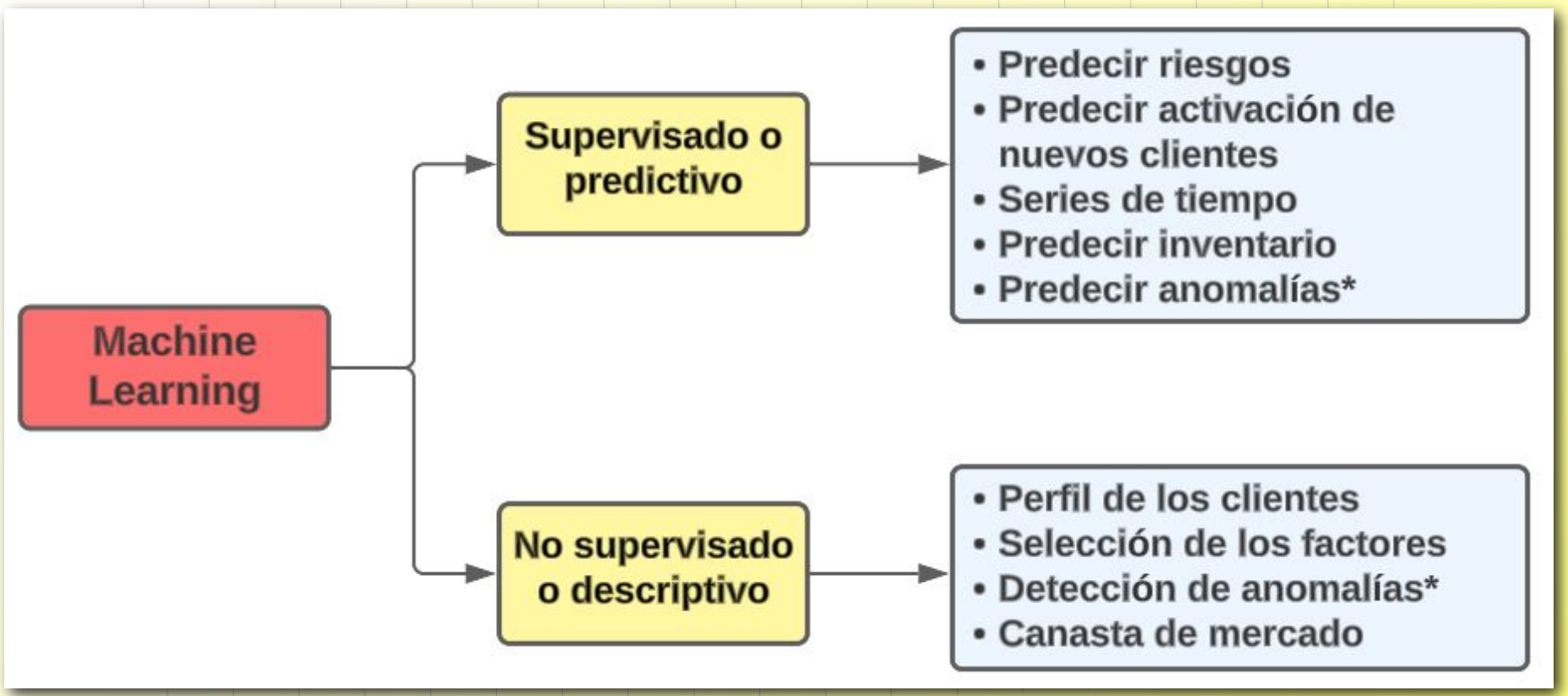


Esquema de ML



Metodología Crisp-DM

Tipos de **Aprendizaje**


Machine Learning
Supervisado o predictivo
• Predecir riesgos
• Predecir activación de nuevos clientes
• Series de tiempo
• Predecir inventario
• Predecir anomalías*
No supervisado o descriptivo
• Perfil de los clientes
• Selección de los factores
• Detección de anomalías*
• Canasta de mercado

Aprendizaje supervisado

El aprendizaje supervisado permite modelar la relación entre las características medidas de los datos y alguna etiqueta asociada con ellos.

Es decir, podremos predecir **y** para nuevos datos **X** de los cuales no conozcamos la salida.

Aprendizaje supervisado

De acuerdo al tipo de etiquetas que asociamos a los datos, el modelo puede realizar dos tipos de tareas:

- ➜ Clasificación: las etiquetas son categorías. Ejemplo: enfermo/sano, gato/perro/pájaro, spam/no spam.
- ➜ Regresión: la variable de salida es un valor numérico. Ejemplo: precio, cantidad, temperatura.

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

La variable que queremos predecir es una clase (categoría)

Clasificación

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

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

Regresión

Aprendizaje no supervisado

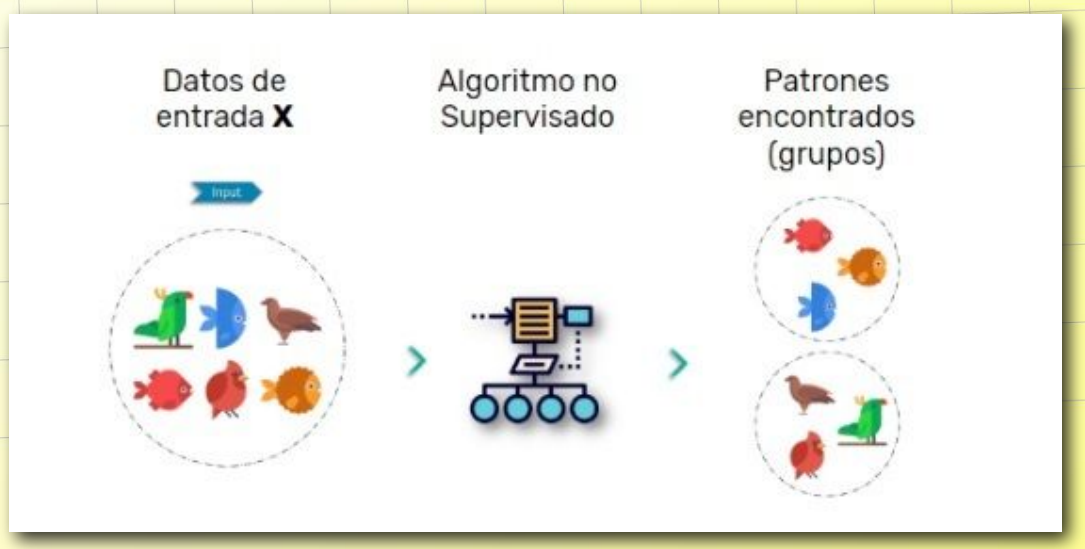
En este caso, se deja que el conjunto de datos hable por sí mismo. Este modelo tiene datos de entrada, pero no se busca una salida en particular.

Implicar modelar las características de un conjunto de datos sin referencia a ninguna etiqueta.

La función de este tipo de algoritmos es encontrar patrones de similaridad.

Aprendizaje no supervisado

Por ejemplo, en clustering, busca identificar distintos grupos de datos:



Flujo de **trabajo**

¿Cuál es?

- Definición: Se definen las preguntas que queremos responder. ¿Qué datos necesitamos para responder esas preguntas?
- Investigación: Se obtienen los datos, se "limpian" y se procede a explorarlos.
- Análisis: Los datos obtenidos se analizan con modelos (estadísticos, Machine Learning, etc.). Interpretamos los resultados y transformamos datos en información.
- Presentación: Presentamos los resultados obtenidos y las conclusiones a las que llegamos. La información se transforma en Conocimiento. Puesta en producción.

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

¿Qué es?

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.

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

Sus posibles valores son categorías pero sí hay una relación de orden. A pesar de que pueden ser números, ¡no se deben sumar!

Ejemplos:

- ➜ Tamaño de una prenda de ropa: XS, S, M, L, XL
- ➜ Tipo de Nafta por octanaje: 95, 98, más de 98
- ➜ Rangos etarios: bebé, niño/a, adolescente, adulto/a, anciano/a

Tratamiento sobre Variables Cualitativas Ordinales

Podemos hacer una asignación a números enteros manteniendo el orden:

S → 0 M → 1 L → 2

Este es uno de los tipos de encoding más comunes que se realizan.

Por ejemplo, llevar al género, valores male y female, a 0 y 1. Lo importante es no perder cuál es cuál.

Esto, en Pandas, se puede hacer con la función map(). Este tipo de encoding se denomina Label_enconding.

Tratamiento sobre Variables Cualitativas Nominales

Sus posibles valores pertenecen a una o varias categorías. Las categorías no siguen una relación de orden. Ninguna es mayor o menor que otra.

Ejemplos:

- 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 Cualitativas Nominales

Una variable dummy toma como valor 0 o 1 para indicar la presencia o ausencia de algún atributo categórico. La función get_dummies() hace automáticamente esto en un dataframe sobre las columnas indicadas.

```
df = pd.concat([df, pd.get_dummies(df['Sex'])], axis=1)
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age_labels	Sex_Map	female	male
	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	adulto	0	0	1
	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	adulto	1	1	0
	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	adulto	1	1	0
	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	adulto	1	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.

Ejemplos:

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

Tratamiento sobre Variables Cuantitativas

En general, este tipo de variables ya vienen en un formato "cómodo" para trabajar, pero a veces queremos agruparlas según grupos o rangos.

Por ejemplo, agrupar edades en rangos (bebés, niños, adolescentes, adultos, ancianos), esto se denomina Discretización y Binning.

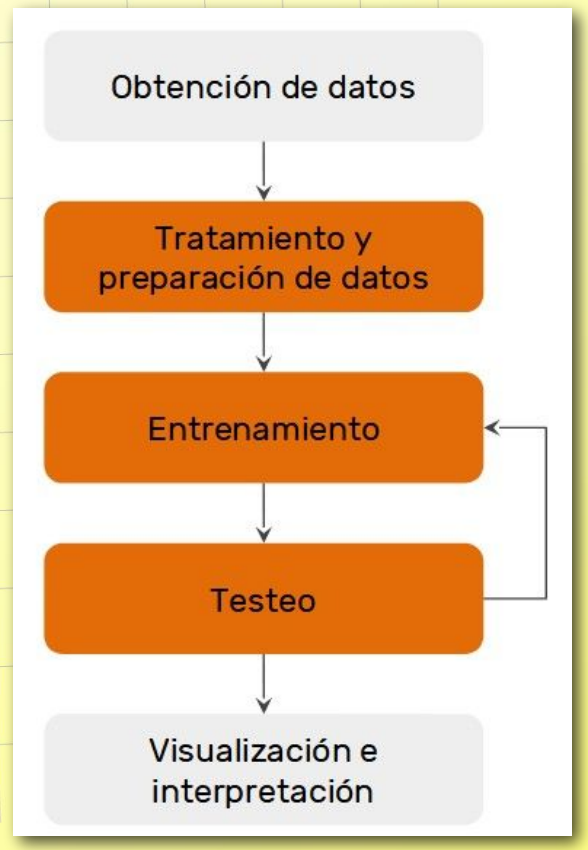
Scikit-Learn

¿Qué es?

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

Se utiliza para:

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



¿Como trabaja?

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()

Herramientas

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.

Herramientas

- 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.

Se utilizan dos métodos: MinMax Scaler y Standard Scaler.

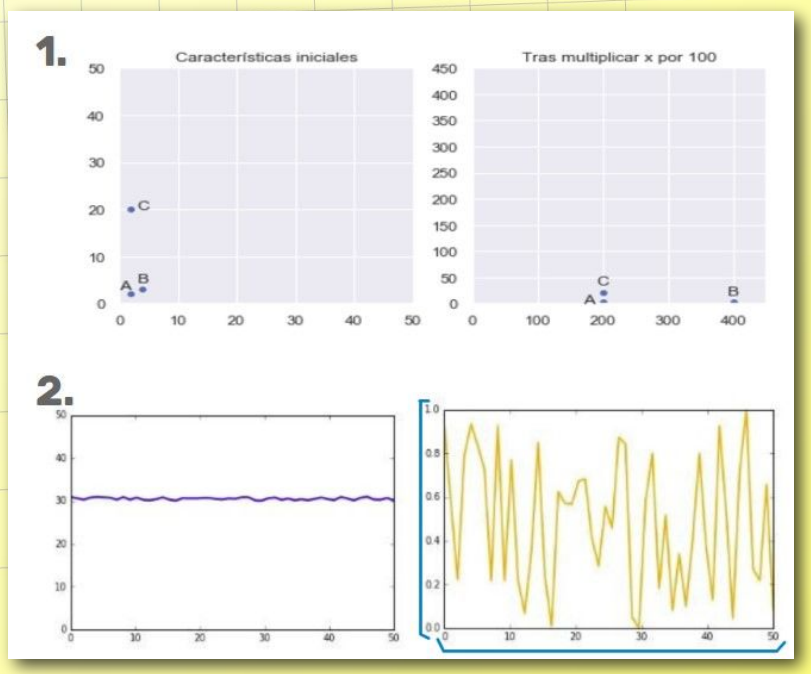
Reescalar los datos: MinMax Scaler

Las entradas se normalizan entre dos límites definidos:

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

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.

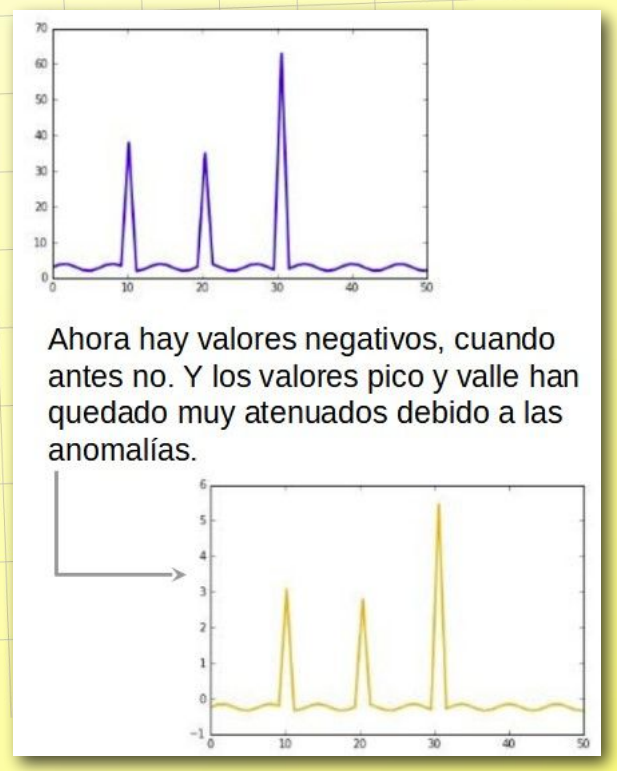


Reescalar los datos: Standard Scaler

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

¿Qué es?

Eventualmente vamos a querer conocer si existe una variación conjunta entre dos variables. Si éste 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:

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

Coeficiente de correlación

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:

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

- Sxy es la covarianza entre x e y.
- Sx y Sy son la desviación estándar de x e y.
- rxy es el coeficiente de correlación.

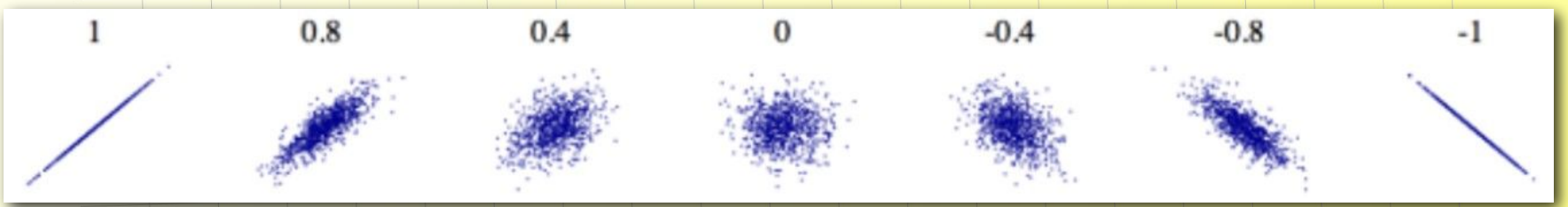
Coeficiente de correlación

Este coeficiente se denomina **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.

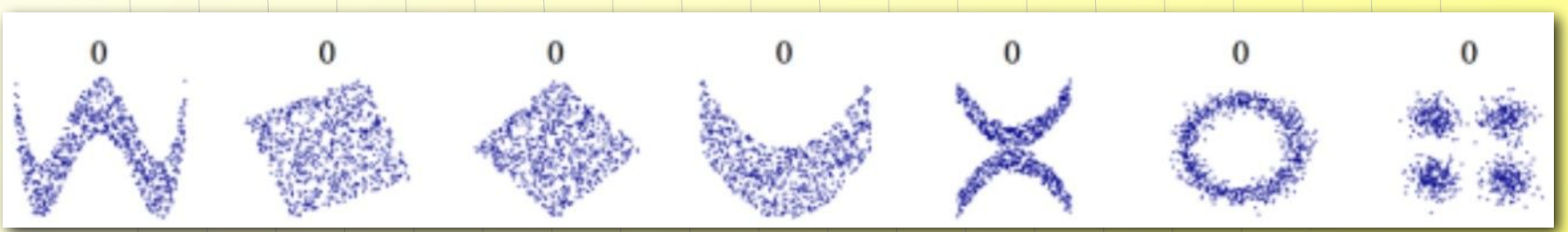
Coeficiente de correlación

- Si el coeficiente de correlación es 0, significa que no existe una relación lineal entre las variables, sin embargo, podría existir otra relación funcional.



Coeficiente de correlación

- Si no hay ninguna relación entre dos variables, entonces el coeficiente de correlación 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.

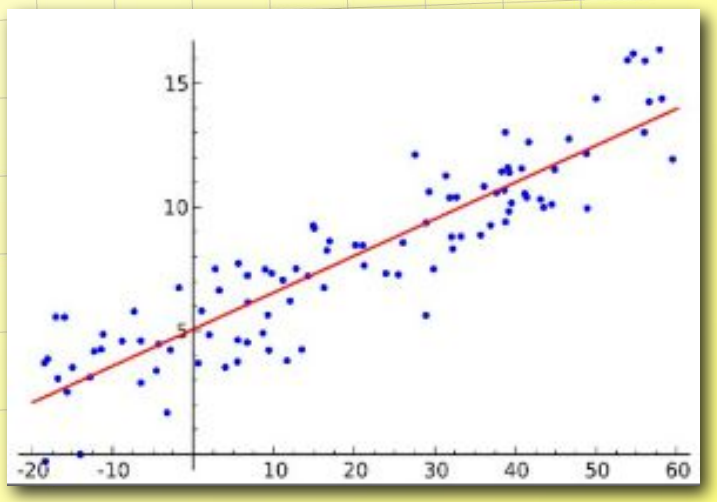
$$Y \approx f(X1, X2, ..., Xp)$$

Se busca $Y = mX + b$ que mejor ajuste a los datos, donde:

m: pendiente

b: ordenada al origen

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



Evaluación del modelo

Evaluar el modelo es cuantificar la performance (calidad de las predicciones).

Algunas de la métricas son:

➜ MAE (Mean Absolute Error)

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

y: $valor\ real$
$\hat{y}$: $valor\ predicho$
N: $n°\ muestra$

Nota: mayor error cuantas más muestras haya.

Evaluación del modelo

Más sensibles a outliers!!

➜ MSE (Mean Squared Error)

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

y: valor real
$\hat{y}$: valor predicho
N: n° muestra

Nota: al estar elevado al cuadrado, no tiene la misma unidad de medida que y.

➜ RMSE (Root Mean Squared Error)

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

y: valor real
$\hat{y}$: valor predicho
N: n° muestra

Nota: Muy utilizado.

Evaluación del modelo

➜ R^2 (Coeficiente de Determinación)

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

y: $valor\ real$
$\hat{y}$: $valor\ predicho$
$\bar{y}$: $valor\ medio\ (promedio)$
T: $número\ de\ observaciones$

Notas:

- Toma valores de 0 a 1.
- Cuanto más cercano a 1, mejor se ajusta.
- Valores cercanos a 0, el modelo es poco fiable.

¿PREGUNTAS?

¿Alguien dijo Homework?

HENRY

Próxima lecture

Modelos de clasificación

¡Feedback!

Click on me

Dispones de un formulario en:

- Homeworks
- Guías de clase
- Slack


HENRY