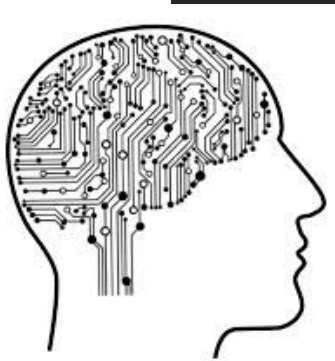
Machine Learning



EN ESTA CLASE

01 ¿Qué es Machine Learning?

Definiciones: ML, DL, aprendizaje, atributos, observaciones, target o label

02 El proceso de ML

Feature engineering, modelos supervisados y no supervisados, train/test split, cross validation, evaluación de modelos

03 Casos de Aplicación

Fraude, churn, upselling, cross-selling, segmentación, clustering geográfico.

04 Etapas de un proyecto de ML

Framing del problema de negocio, integración de datos, modelado, puesta en producción y monitoreo.

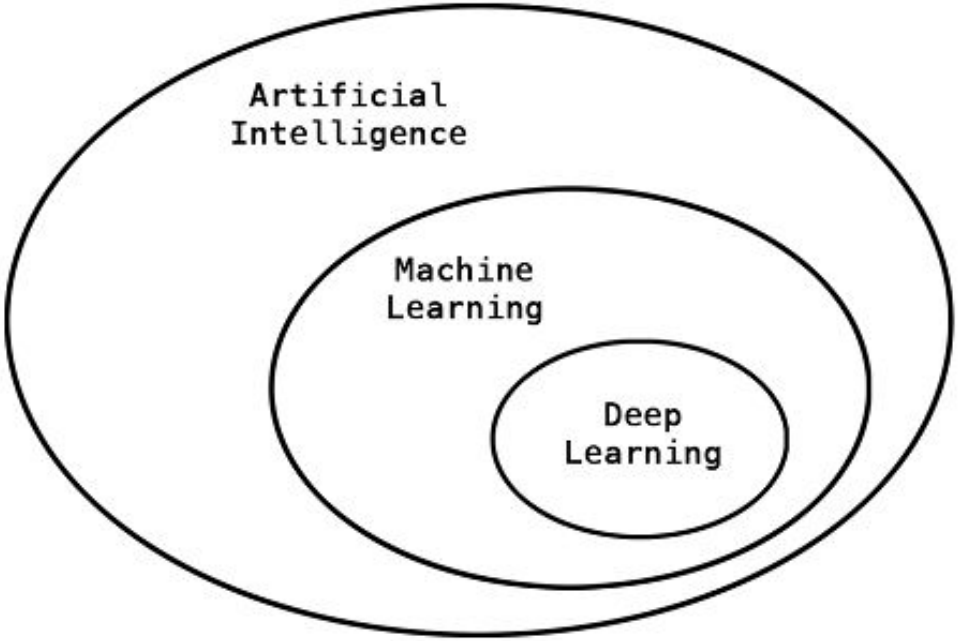
"Machine learning es la ciencia de hacer que las computadoras actúen sin ser programadas explícitamente." - Andrew Ng

01

¿QUÉ ES ML?

Algunas definiciones básicas

INTELIGENCIA ARTIFICIAL



La “Inteligencia Artificial Simbólica” dominó el paradigma de IA desde 1950 hasta la parte de los ‘80. Su pico de popularidad: boom de los “sistemas expertos”.

MACHINE LEARNING

Programación clásica se ingresan reglas explícitamente para procesar el input.

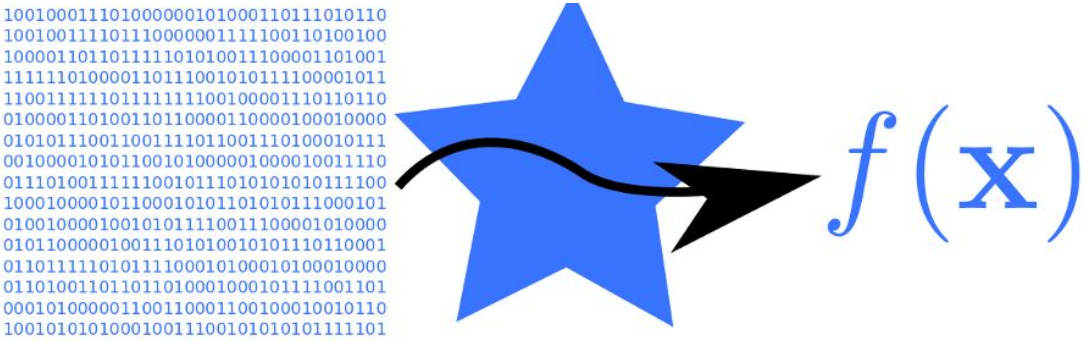
Machine Learning: los humanos ingresan datos como input además de las respuestas esperadas, y las reglas surgen como output. Estas reglas pueden luego ser aplicadas a nuevos datos para producir respuestas originales.

DATOS, ALGORITMO Y MODELO

Datos

Algoritmo

Modelo



$f(\mathbf{x})$

CONCEPTOS BÁSICOS

La información de organiza en observaciones con atributos. En algunos casos, estas observaciones tienen una variable target que queremos predecir.

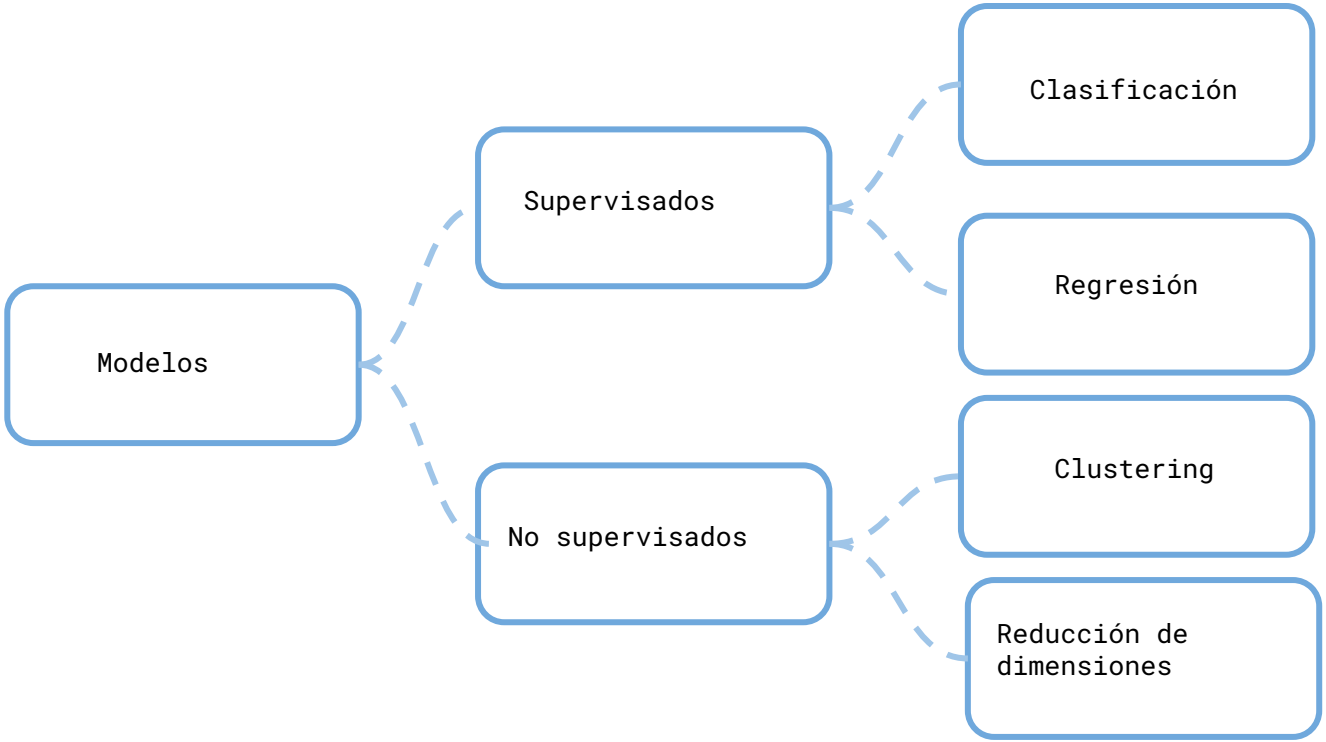
Atributos

Target

Observación

upsell	user_id	mes	page_views	tiempo	compras previas
1	144332	5	23	15	3
0	634631	5	14	10	1
0	123126	5	10	8	0

TIPOS DE ALGORITMOS



02
OVERVIEW DEL PROCESO

Cómo entrenar un modelo

TRANSFORMACIÓN DE LOS DATOS

Los modelos de ML sólo son capaces de aprender de representaciones numéricas

¿Cómo podremos entonces trabajar con texto, categorías o imágenes?

CATEGORÍAS

One hot encoding o variables Dummy

Legible para un humano

Pet
Cat
Dog
Turtle
Fish
Cat

Legible para un algoritmo

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

TEXTO

Bag of words

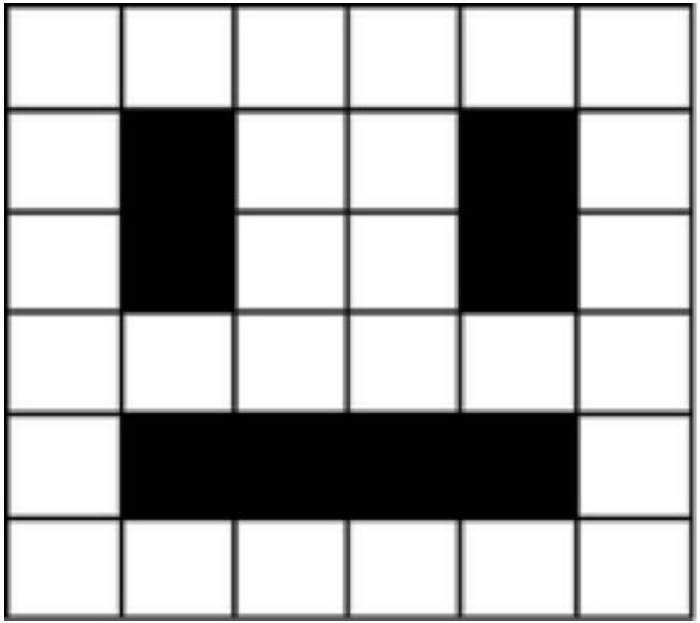
Legible para un humano

Legible para un algoritmo

IMÁGENES

Brillo en los canales R,G,B

Legible para un humano



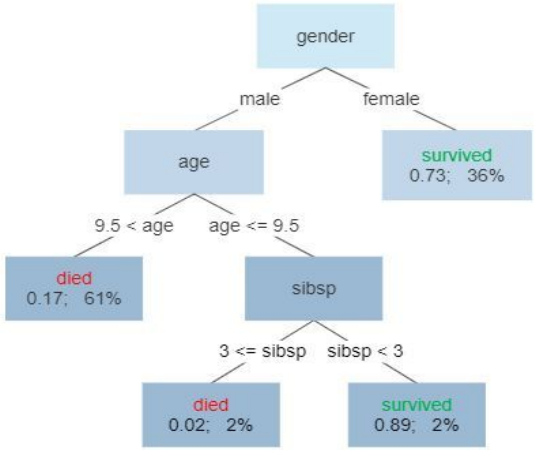
Legible para un algoritmo

0	0	0	0	0	0
0	1	0	0	1	0
0	1	0	0	1	0
0	0	0	0	0	0
0	1	1	1	1	0
0	0	0	0	0	0

Ya tengo mis datos representados como vectores numéricos o “features” y etiquetas o “labels”

¿Cómo entrenar un modelo que capture las relaciones entre los mismos?

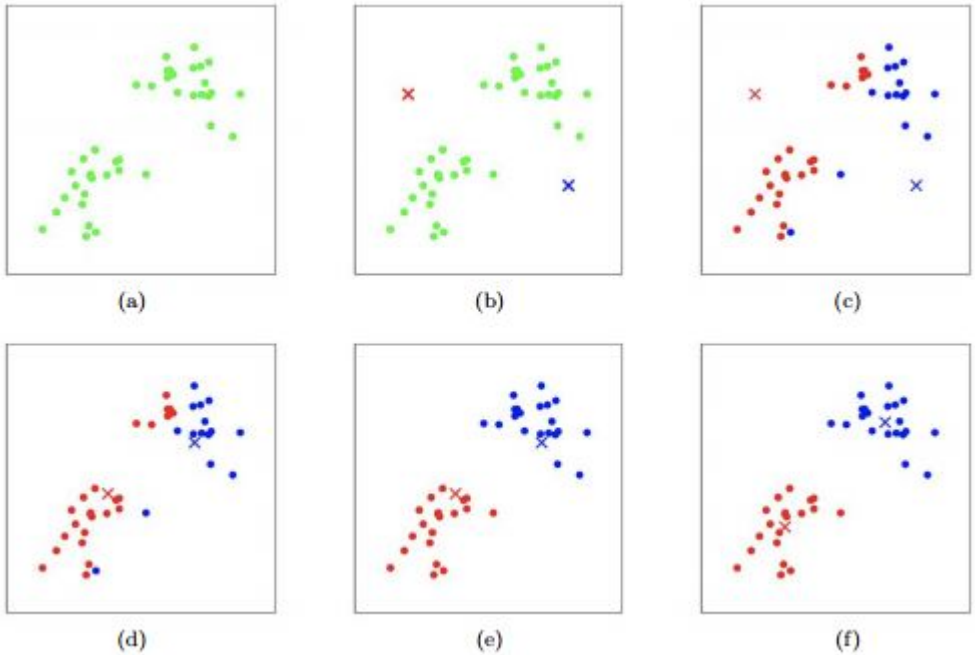
MODELOS SUPERVISADOS | ÁRBOLES



Los modelos de Machine Learning más performantes, se basan en "ensamblar" muchos árboles simples. Sirven tanto para **clasificación** como para **regresión**.

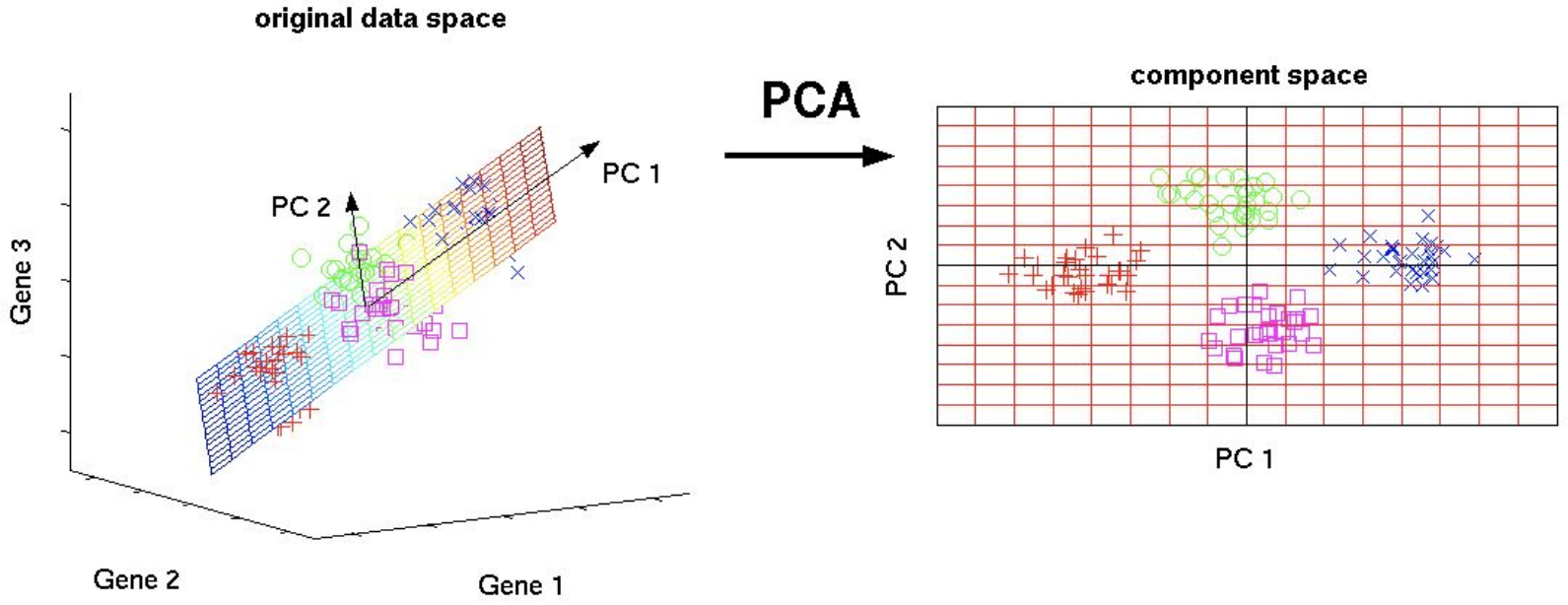
- **Random Forest**
- **XGBoost**
- **LightGBM**
- **Tree Gradient Boosting**

MODELOS NO SUPERVISADOS | K-MEANS



1. Asignar centroides al azar
1. Calcular qué punto pertenece a cada cluster
1. “Promediar” todos los puntos en cada dimensión para recalcular el centroide
1. Repetir hasta que los puntos dejen de cambiar de cluster

MODELOS NO SUPERVISADOS | PCA



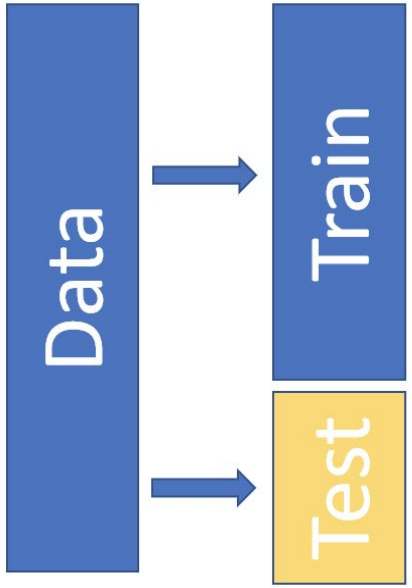
MODELOS DE ML

Ya tengo un modelo que me permite predecir, clusterizar o reducir dimensiones

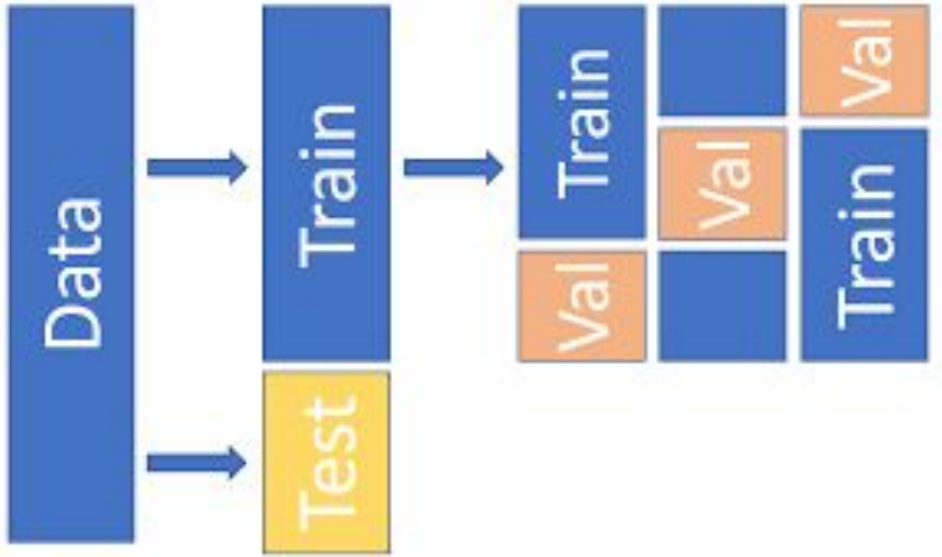
¿Cuán bueno es mi modelo? ¿Cómo funcionará sobre datos nuevos?

HOLD OUT SETS PARA EVALUACIÓN

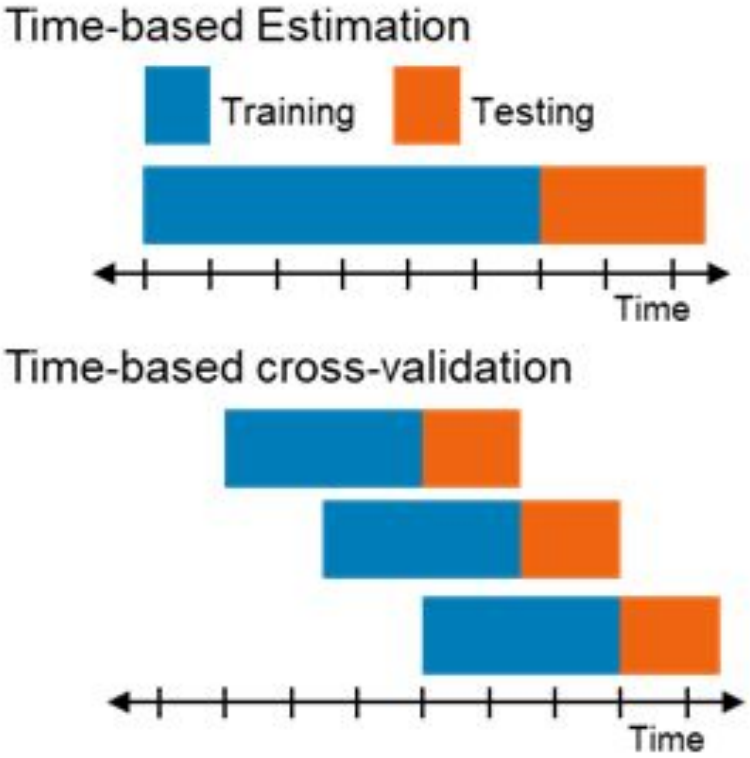
Train/Test Split



Cross Validation



EVALUACIÓN EN SERIES DE TIEMPO



¡Cuidado! Es imposible usar el futuro para predecir el presente.

Cuando los modelos incorporan información que naturalmente no debería estar disponible este problema se llama “**information leak**”.

Aquí un caso famoso de information leak que tomó conocimiento público: https://liaa.dc.uba.ar/es/sobre-la-prediccion-automatica-de-embarazos-adolescentes/

MÉTRICAS PARA CLASIFICACIÓN

Matriz de confusión

	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

$$Accuracy = \frac{Casos\ Correctamente\ clasificados}{Casos\ Totales}$$

$$Recall = TP / TP + FN$$

$$Precision = TP / TP + FP$$

$$F\,score = 2\,\frac{precision * recall}{precision + recall}$$

MÉTRICAS PARA REGRESIÓN

Mean Absolut Error (Error absoluto medio)

$$MAE = \frac{1}{n}\sum\left|y - y_{predicha}\right|$$

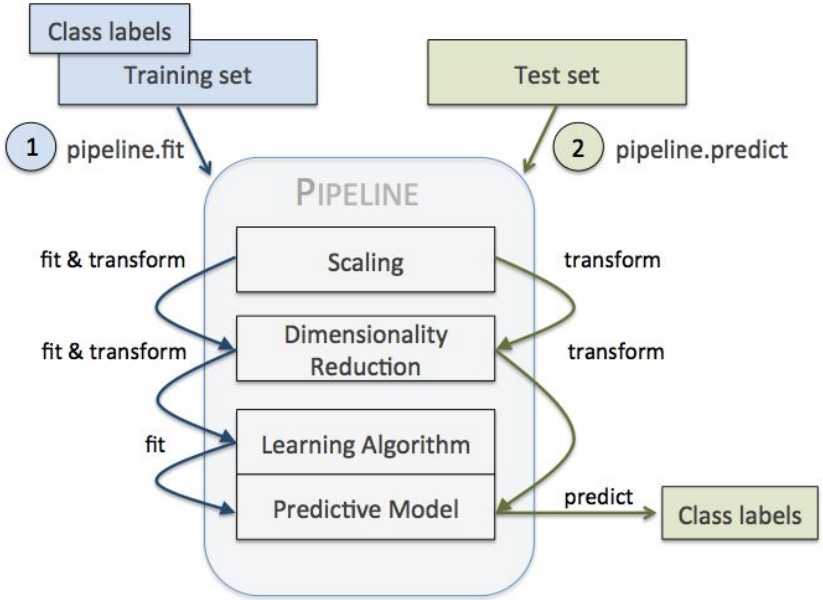
Mean Squared Error (Error cuadrático medio)

$$MSE = \frac{1}{n}\sum(y - y_{predicha})^2$$

R cuadrado

$$R^2 = \sum\frac{(y-y_{predicha})^2}{(y-y_{promedio})^2}$$

PIPELINES


Class labels
Training set
Test set
1
pipeline.fit
2
pipeline.predict
PIPELINE
fit & transform
Scaling
transform
fit & transform
Dimensionality Reduction
transform
fit
Learning Algorithm
Predictive Model
predict
Class labels

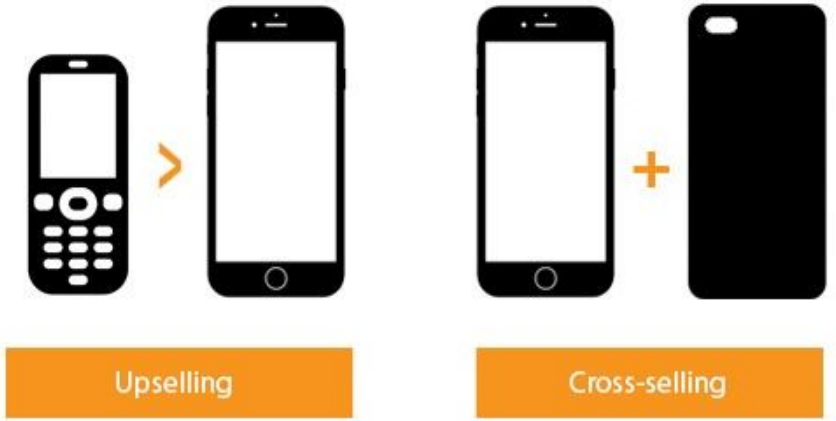
03
CASOS DE APLICACIÓN

¿Qué se hace hoy con ML?

PROPENSIÓN A LA CONVERSIÓN ONLINE

- **Objetivo**: predecir qué tan probable es que un visitante a una app realice una conversión. Ésta puede ser una compra online, un clic, completar un formulario, etc.
- Esto se usa para decidir cuánto pagar en una subasta, asignar un precio o producto dinámicamente, etc.

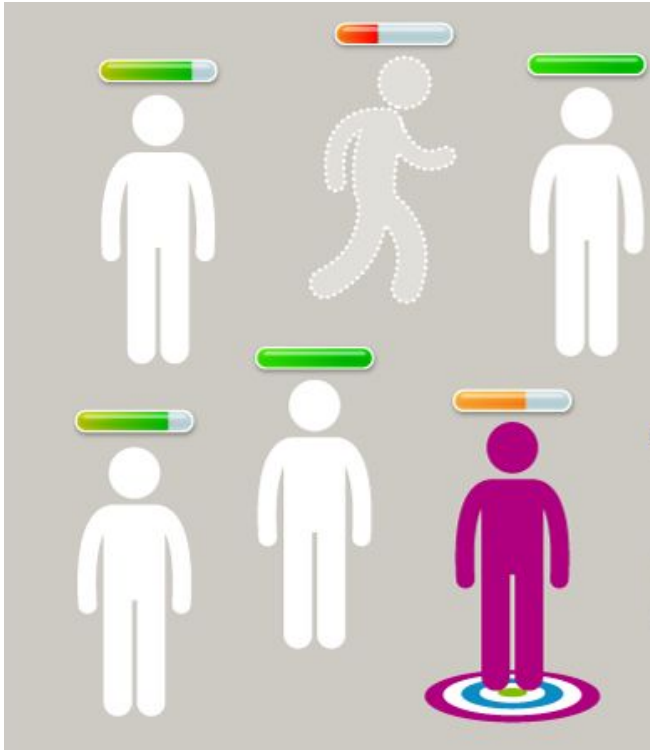
UPSELLING Y CROSS SELLING



- **Objetivo:** predecir la probabilidad de que un cliente compre un producto más caro (upselling) o complementario (cross-selling).

- Esto se puede integrar a un CRM o una herramienta de marketing para realizar campañas automáticamente.

PREDICCIÓN DE CHURN



- **Objetivo:** predecir la probabilidad de que un cliente se dé de baja en determinado período.

- Sabiendo quiénes son los más propensos podemos generar un incentivo para impedirlo.

SISTEMAS DE RECOMENDACIÓN

- **Objetivo:** optimizar qué productos ofrecer en una plataforma de ventas online

04
EL PROCESO COMPLETO

¿Qué hay que tener en cuenta para desplegar un modelo?

EL PROCESO REAL DE LOS DATOS EN ML