Introducción a Machine Learning

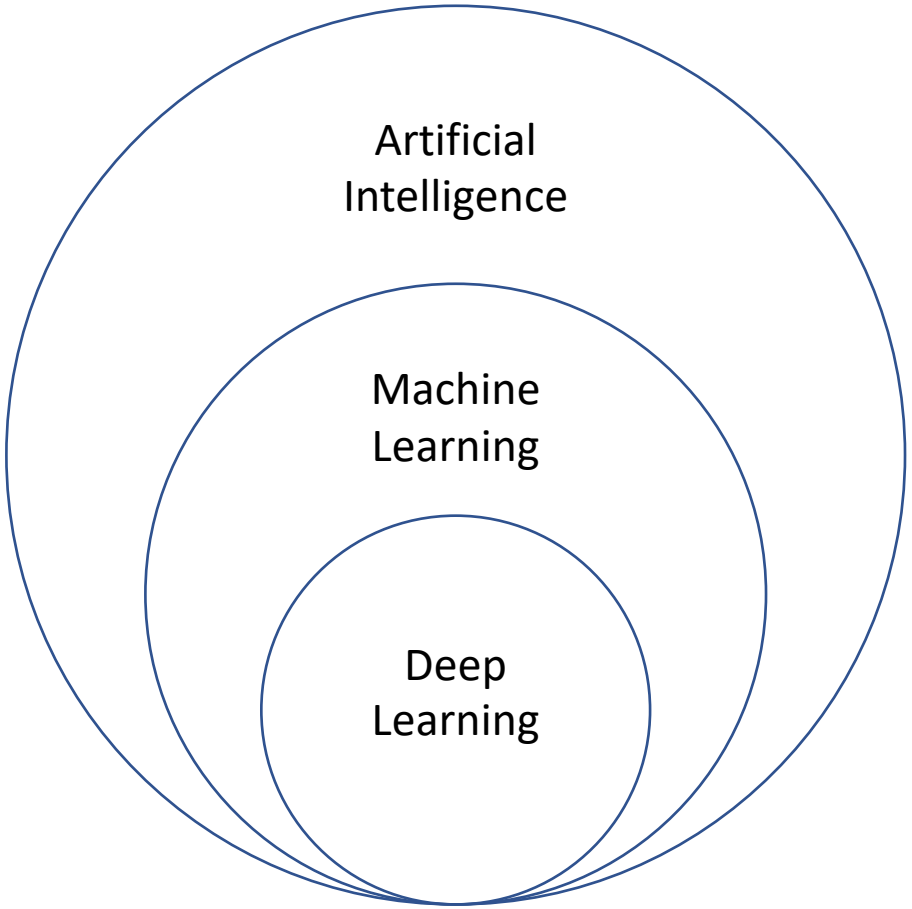
Bootcamp Python – Febrero 2023

Visión General

Inteligencia Artificial
Emular el comportamiento humano

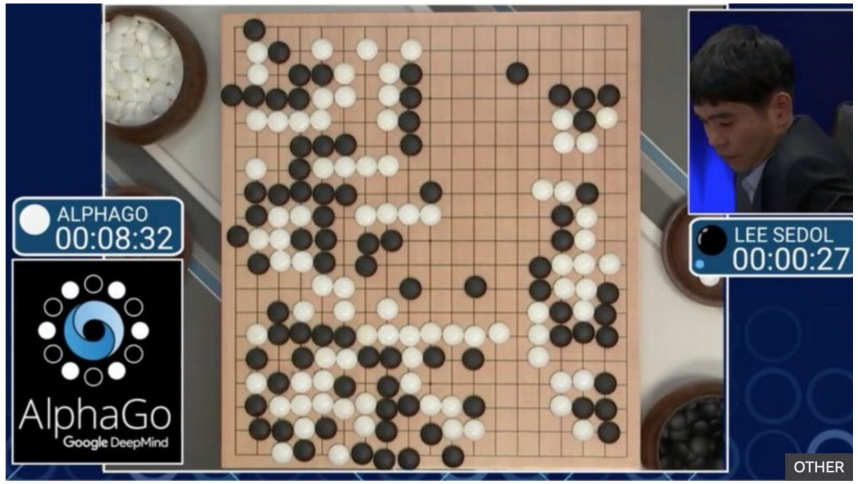
Machine Learning
Aprender a realizar tareas a partir de experiencias previas

Deep Learning
Mejorar el aprendizaje y representación de los datos



Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

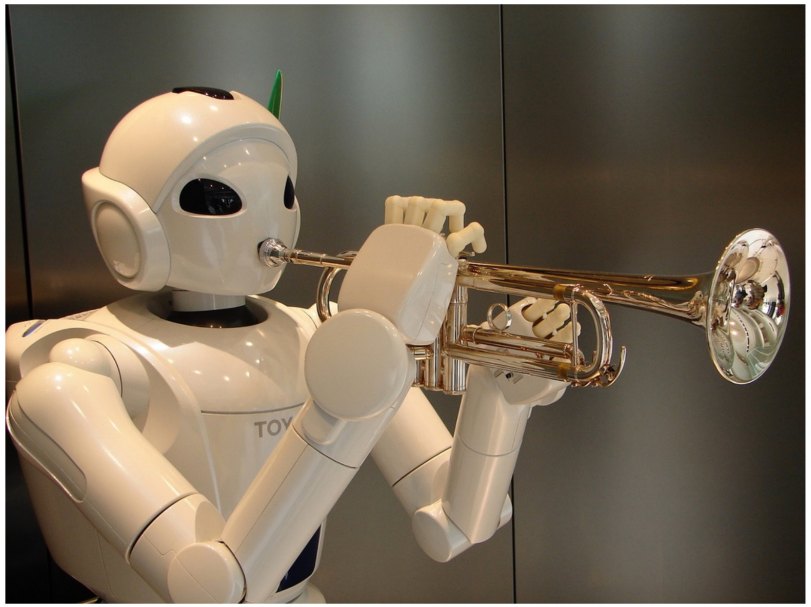
12 March 2016



https://www.bbc.com/news/technology-35785875

Sony Releases Beatles-Inspired Song Made with A.I. Software

written by Ville Iso-Ahola | October 4, 2016



https://geekinsider.com/sony-releases-beatles-inspired-song-created-artificai-intelligence-software/

Documentales y películas

¿Qué es machine learning?

Es un campo de estudio que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas

(Arthur Samuel – 1959)

Aprendizaje supervisado

El conjunto de entrenamiento consiste en las variables y una **etiqueta**

Se entrena el modelo para **predecir** las etiquetas en un conjunto de datos nuevo

Ejemplo: predicción de compra de un producto

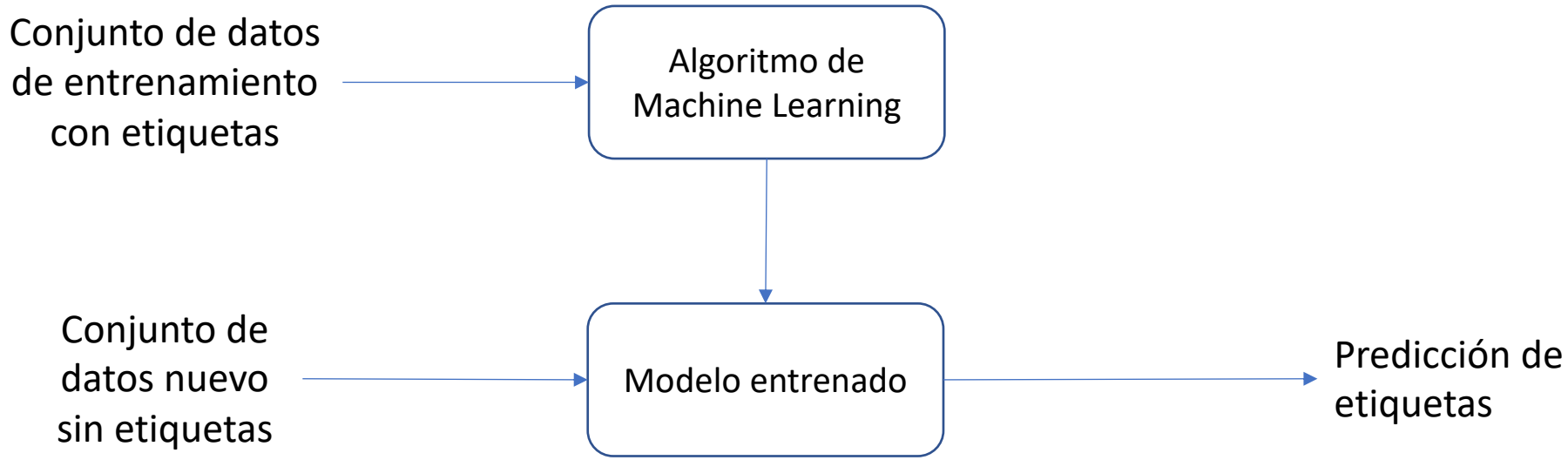
Aprendizaje no supervisado

El conjunto de entrenamiento no tiene etiquetas

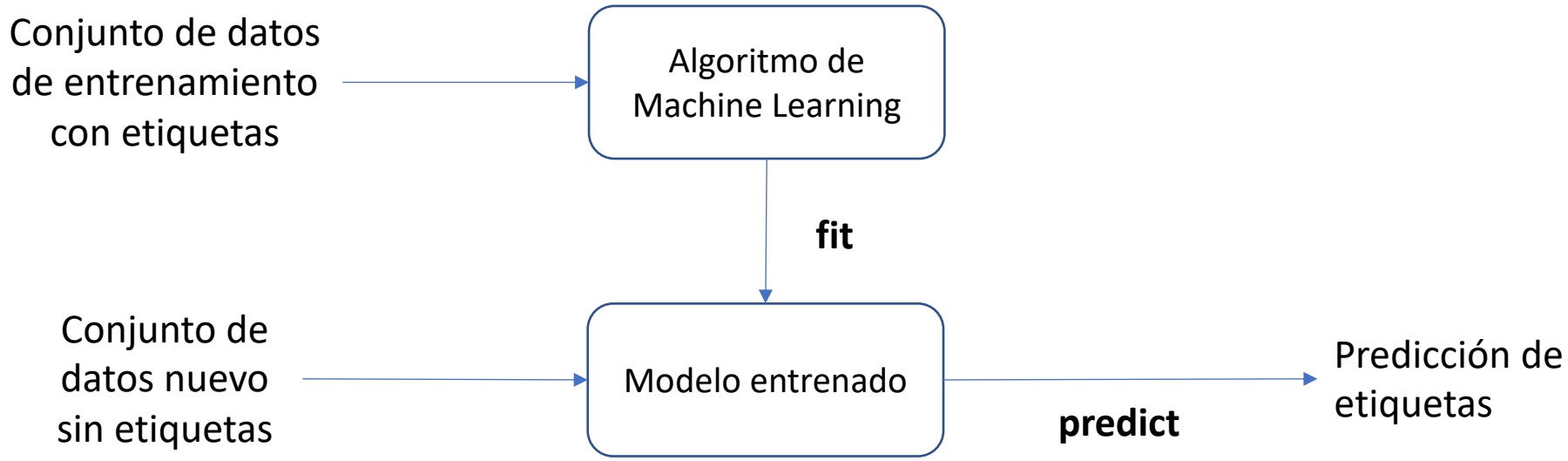
Se entrena el modelo para **encontrar patrones** en la data

Ejemplo: segmentación de clientes

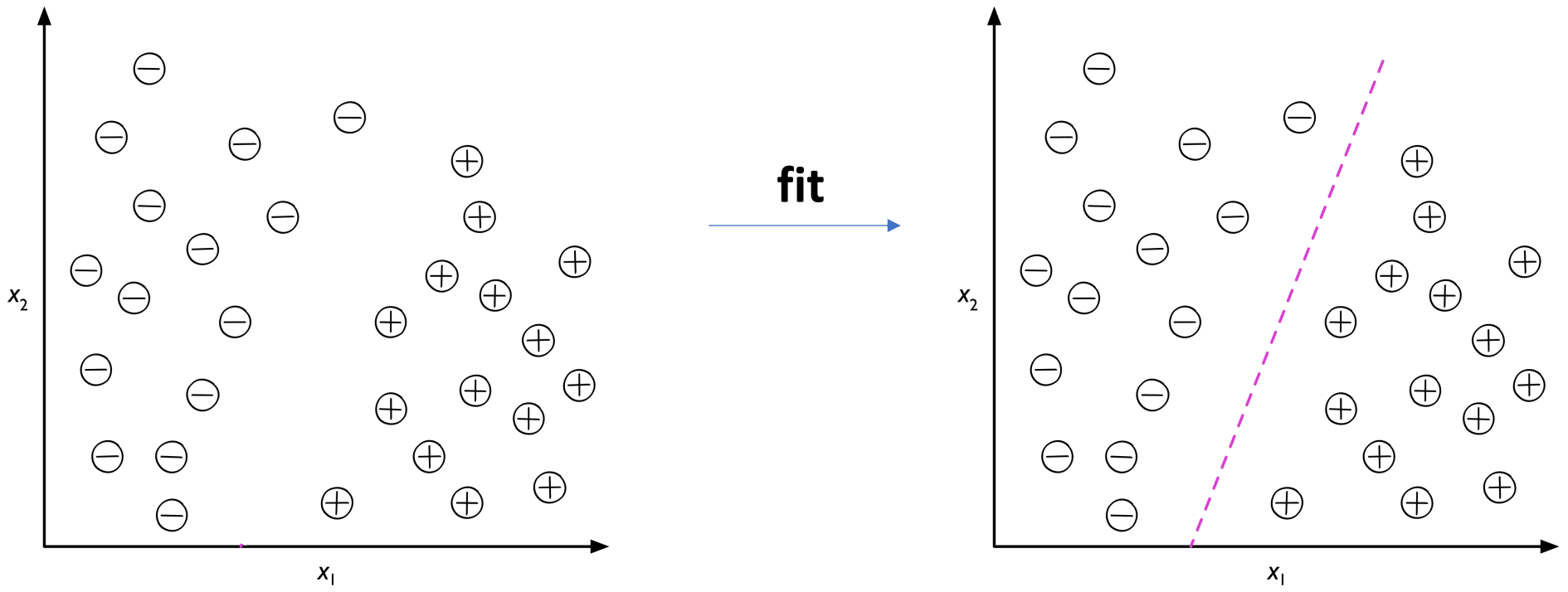
Aprendizaje supervisado: haciendo predicciones sobre el futuro



Aprendizaje supervisado: haciendo predicciones sobre el futuro



Aprendizaje supervisado: clasificación para predecir etiquetas



Aprendizaje supervisado: Terminología

Sueldo	Antiguedad en sistema financiero	Edad	Capacidad de endeudamiento	Pagó el préstamo
				Sí
				Sí
				Sí
				No
				No

Cantidad de muestras

Características
(**Features**, atributos, dimensiones)

Etiqueta
(Label / **Target**)

Aprendizaje supervisado: Conjuntos de entrenamiento y prueba

Sueldo	…	…	…	Pagó
				Sí
				Sí
				Sí
				No
				No

Conjunto de entrenamiento **(Train)**

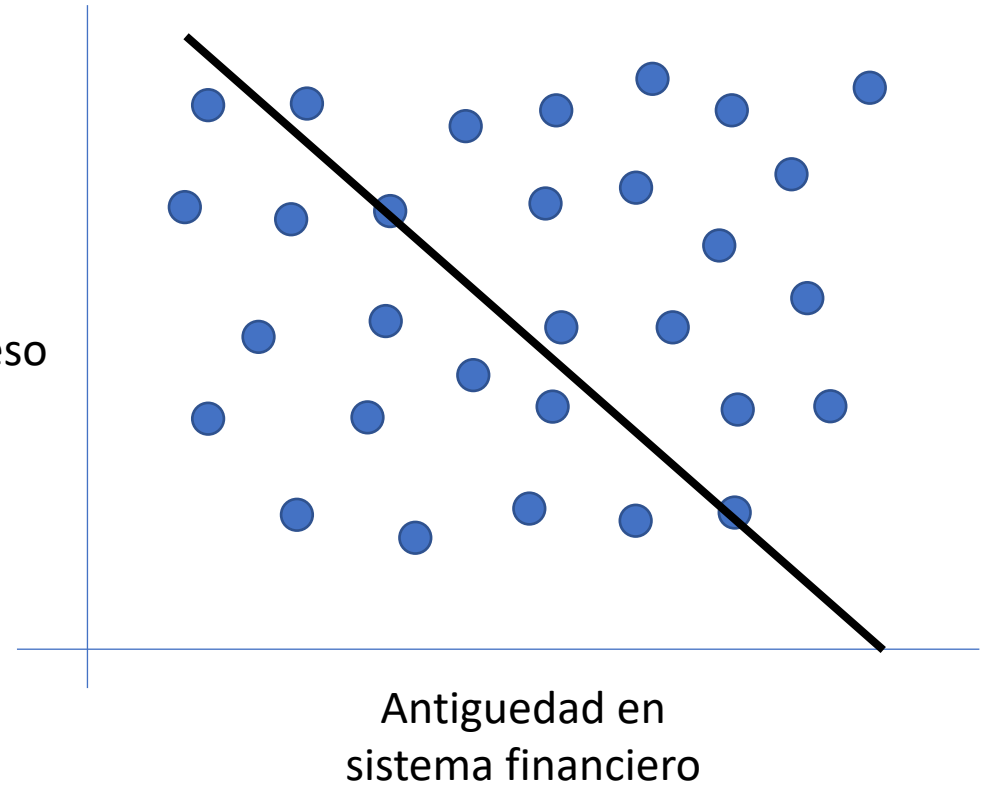
Sueldo	…	…	…	Pagó
				?
				?
				?
				?
				?

Conjunto de prueba **(Test)**

Clasificación: Predicción de probabilidades

- En los modelos de predicción, muchas veces no se predice directamente una clase, sino se predice una probabilidad.

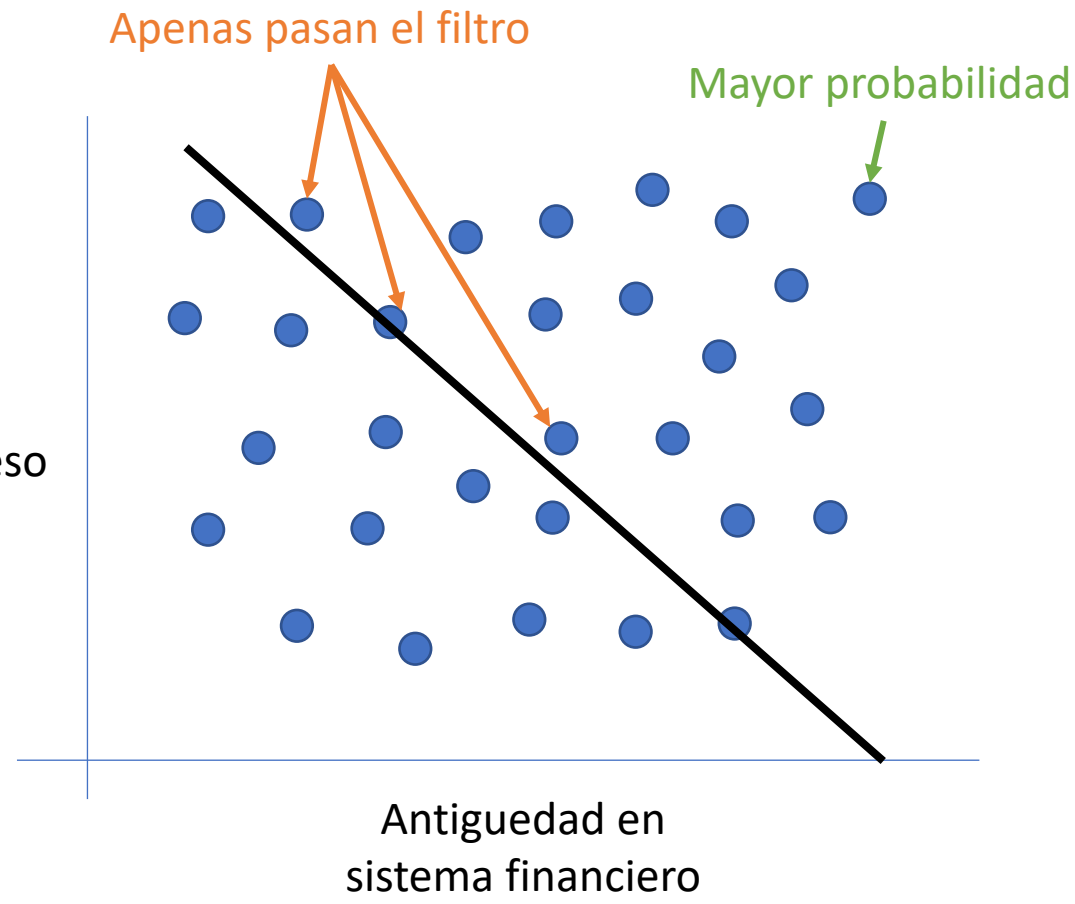
- ¿Qué tan probable es que un cliente me pague un crédito?



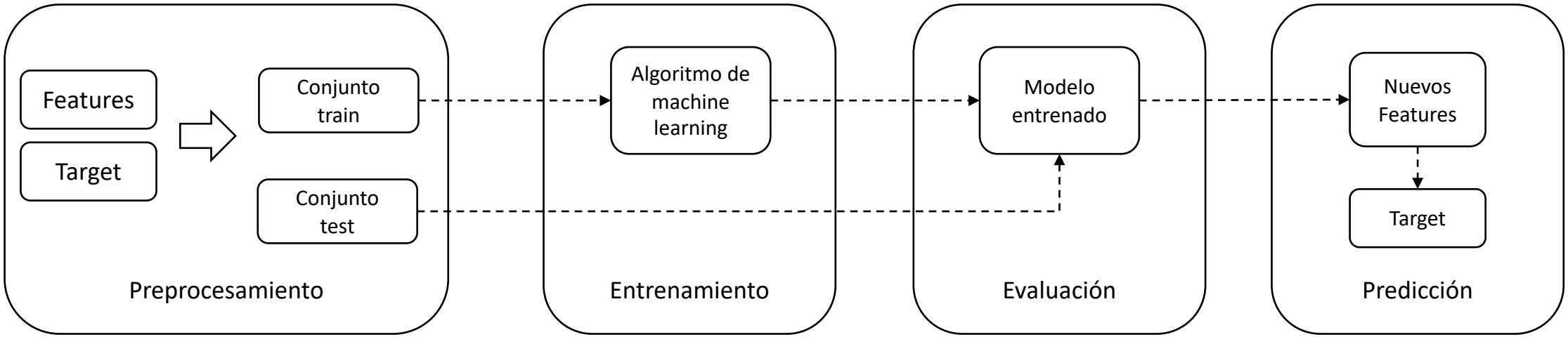
Clasificación: Predicción de probabilidades

- En los modelos de predicción, muchas veces no se predice directamente una clase, sino se predice una probabilidad.

- ¿Qué tan probable es que un cliente me pague un crédito?



Aprendizaje supervisado: Proceso de modelamiento



Muestreo
Tratamiento de nulos y valores atípicos
Escalamiento

Selección de modelo
Regularización
Cross validation
Optimización de parámetros

Métricas de evaluación
Análisis de error

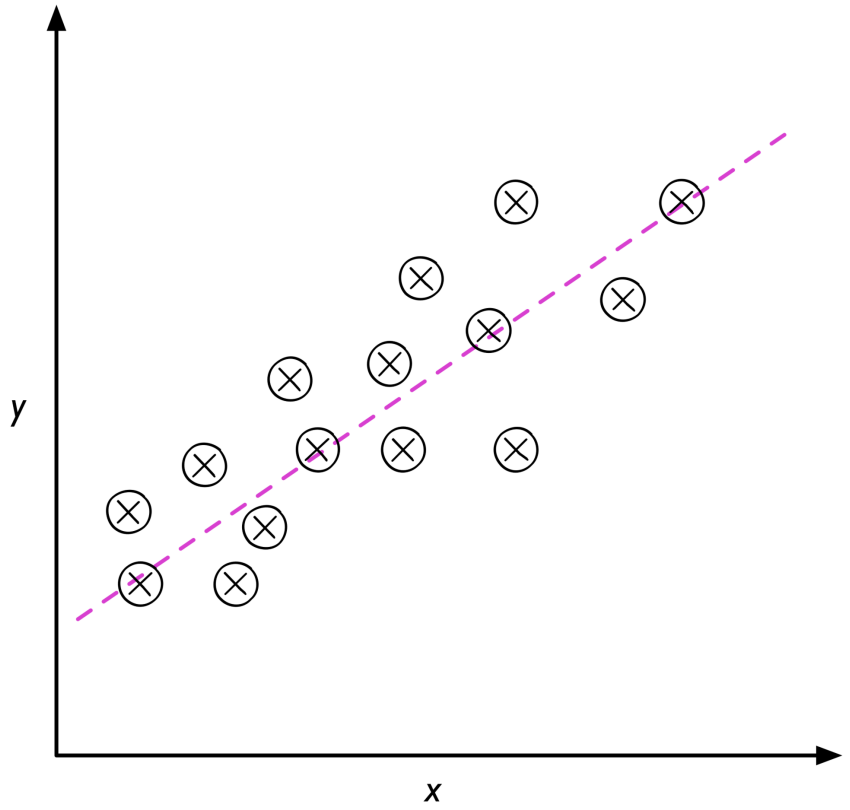
Métricas de evaluación
Análisis de error

Aprendizaje supervisado: regresión para predecir valores continuos

En un modelo de regresión queremos predecir un valor continuo de target.

En la imagen, dada una variable x intentamos estimar el valor de la variable y

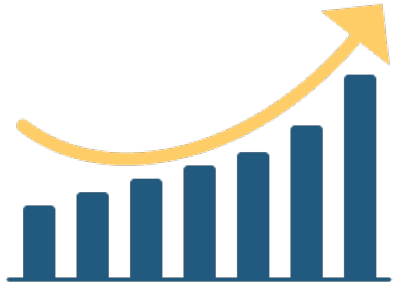
La línea elegida (el modelo) es resultado de ajustar un modelo en el que se intenta minimizar la distancia entre los valores reales y la predicción del modelo.



Aprendizaje supervisado en la industria

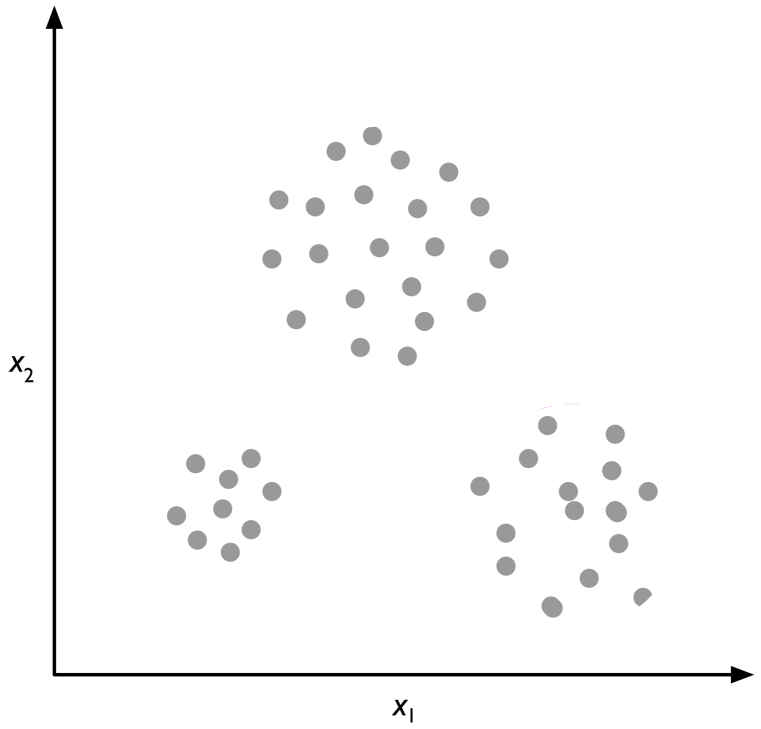
Propensión para campañas de ventas

Predicción de churn
Retención de clientes

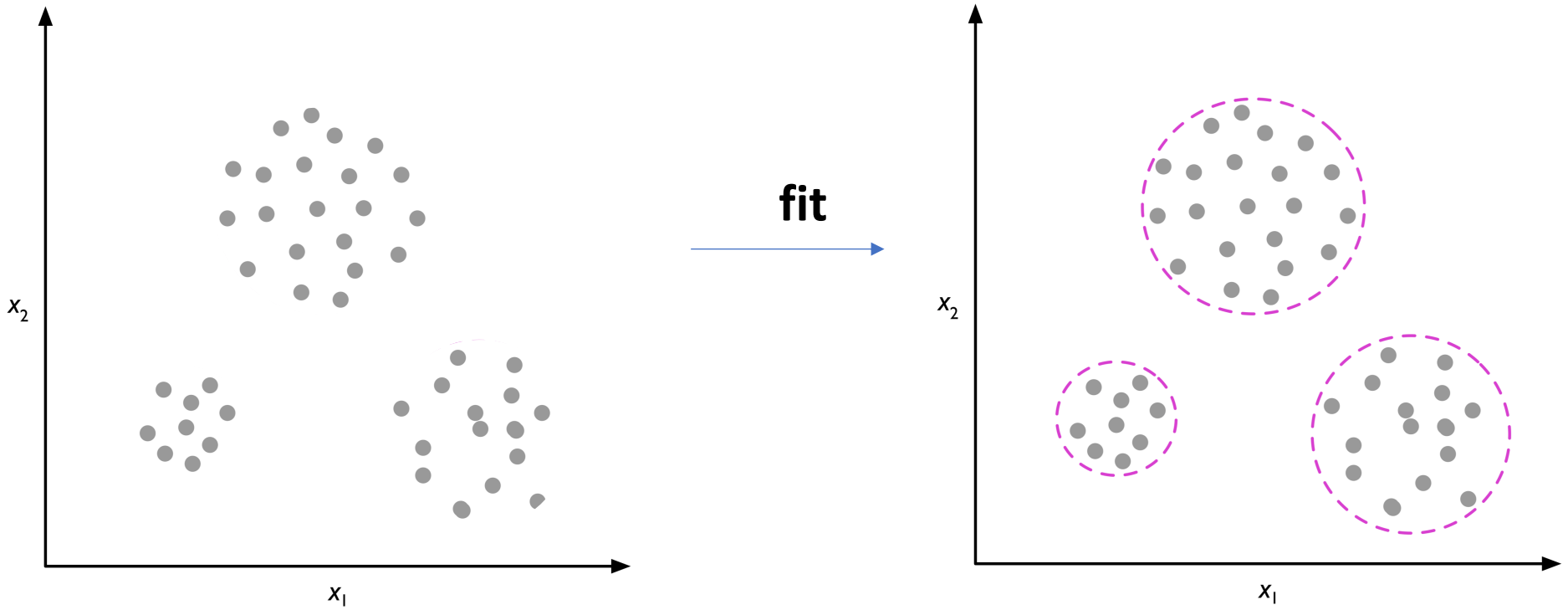


Forecasting
Estimación de demanda

Aprendizaje no supervisado: encontrando subgrupos con clustering



Aprendizaje no supervisado: encontrando subgrupos con clustering



Técnicas de clustering: https://scikit-learn.org/stable/modules/clustering.html

Aprendizaje no supervisado en la industria

Segmentación de clientes

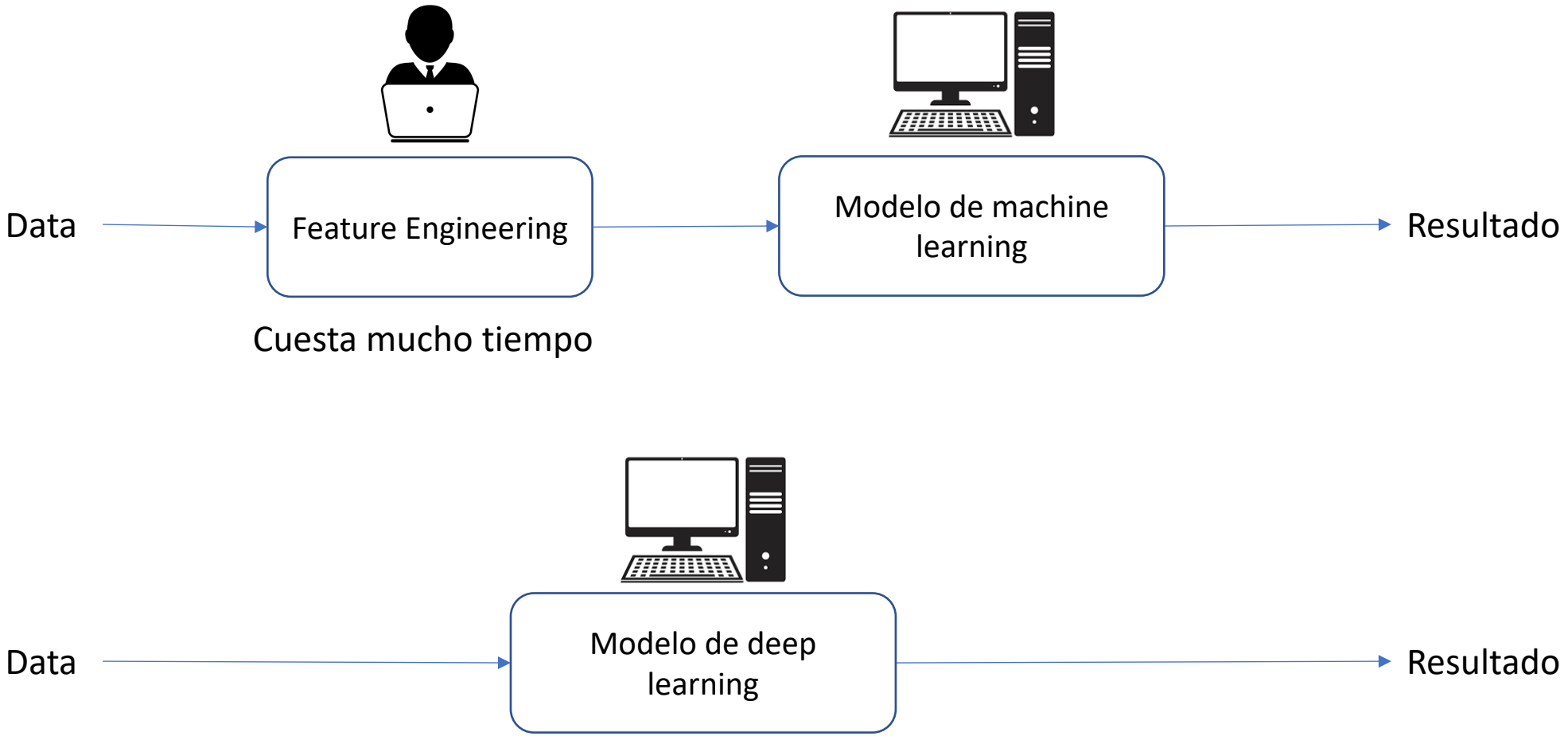
Análisis de redes transaccionales

Análisis de canasta de mercado
Sistemas de Recomendación

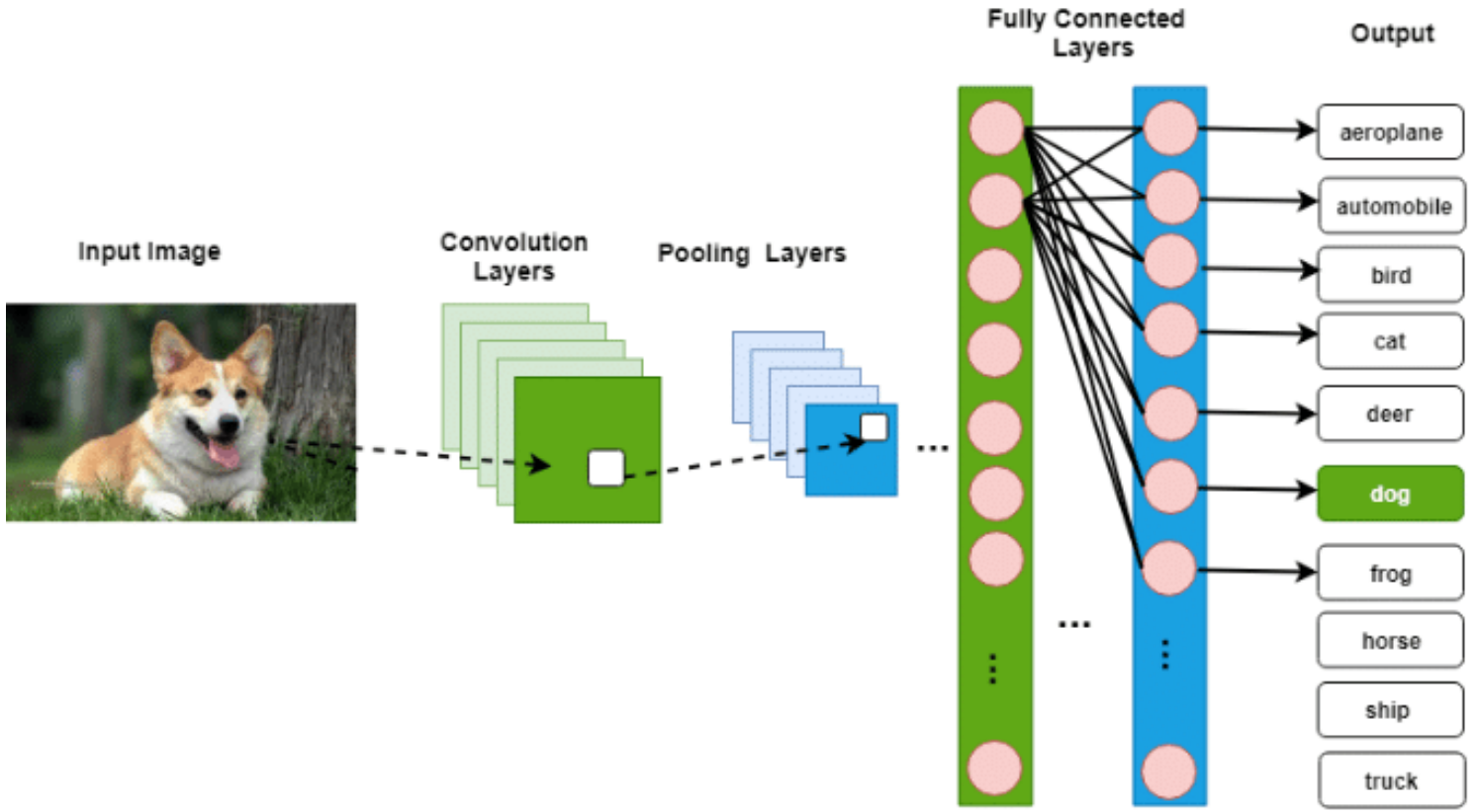
¿Qué es Deep Learning?

Es una nueva área de Machine Learning que fue introducida con el fin de acercar al Machine Learning a uno de sus objetivos originales: la inteligencia artificial

(deeplearning.net)


Data
Feature Engineering
Modelo de machine learning
Resultado
Cuesta mucho tiempo
Data
Modelo de deep learning
Resultado

¿Qué es Deep Learning?

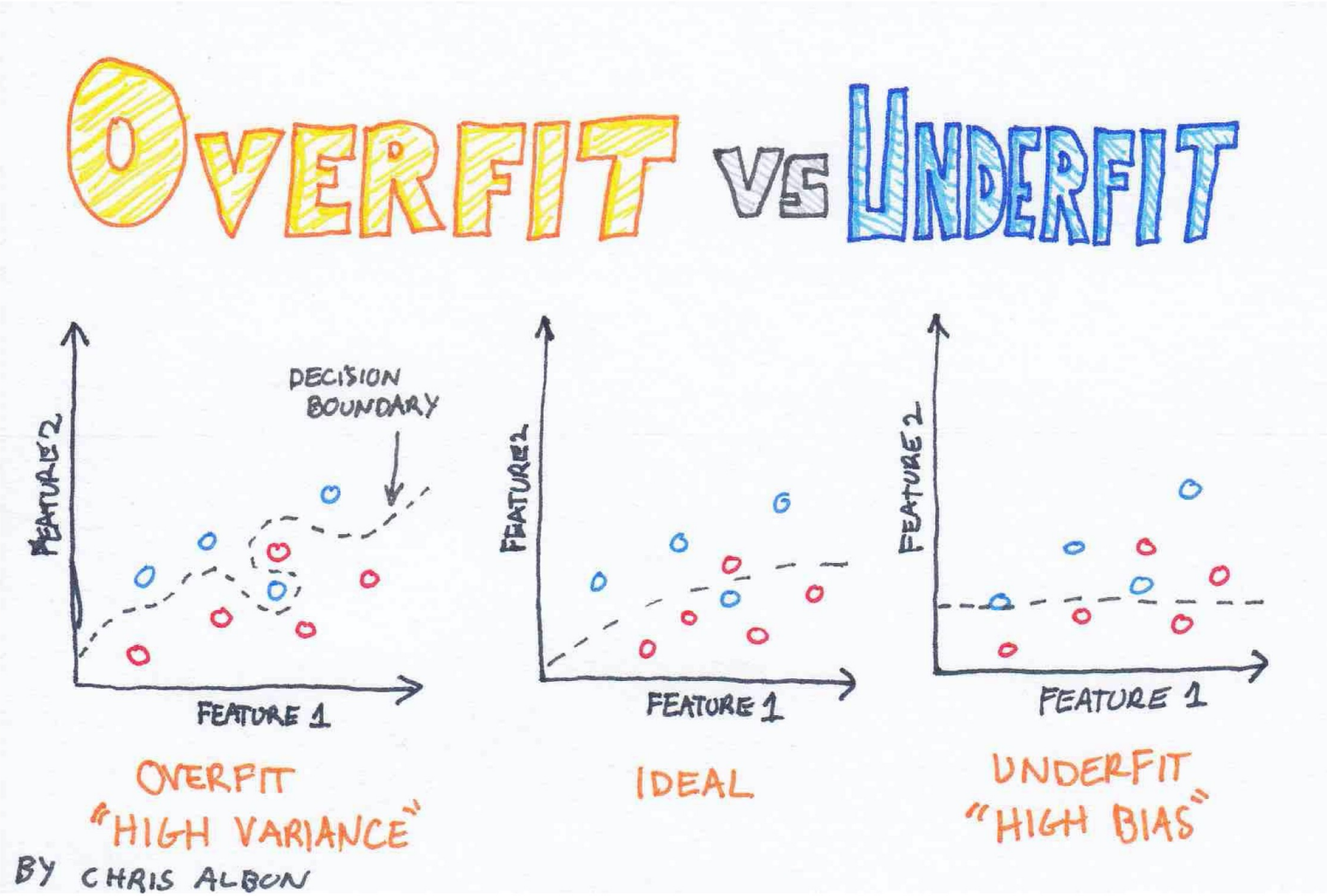


Disyuntivas en el desarrollo de modelos

Overfitting y underfitting

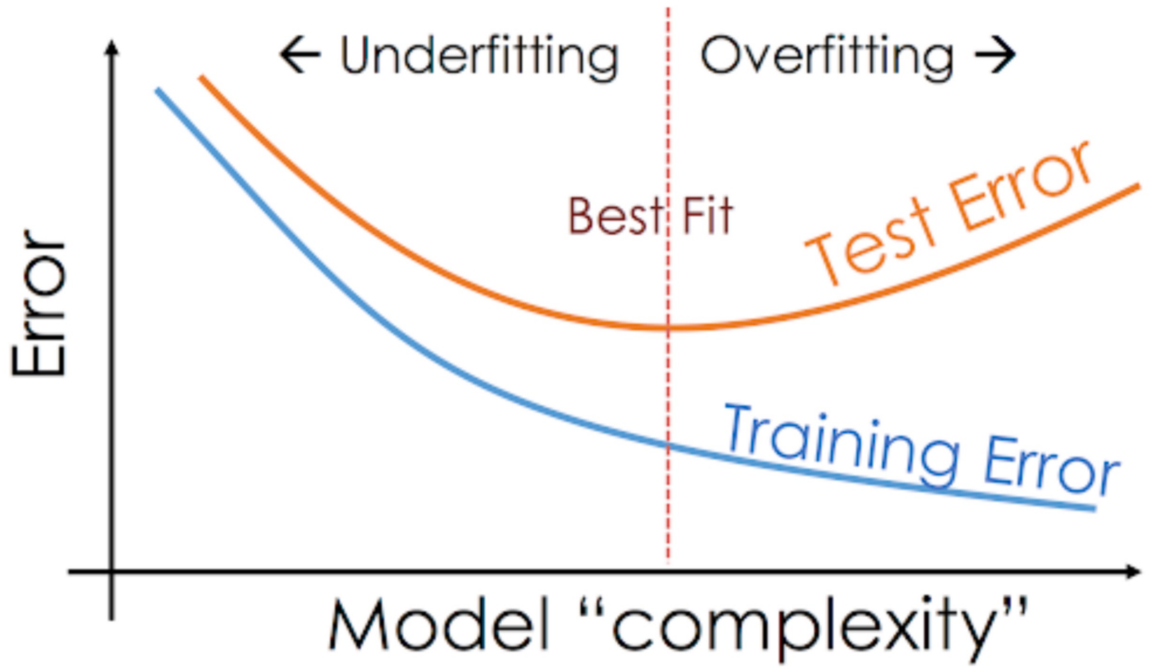
Overfitting: Cuando un modelo se sobreajusta demasiado a los datos de entrenamiento no generaliza sus resultados adecuadamente

Underfitting: Cuando el modelo se ajusta muy ligeramente a los datos no se realiza una buena predicción



Overfitting y underfitting

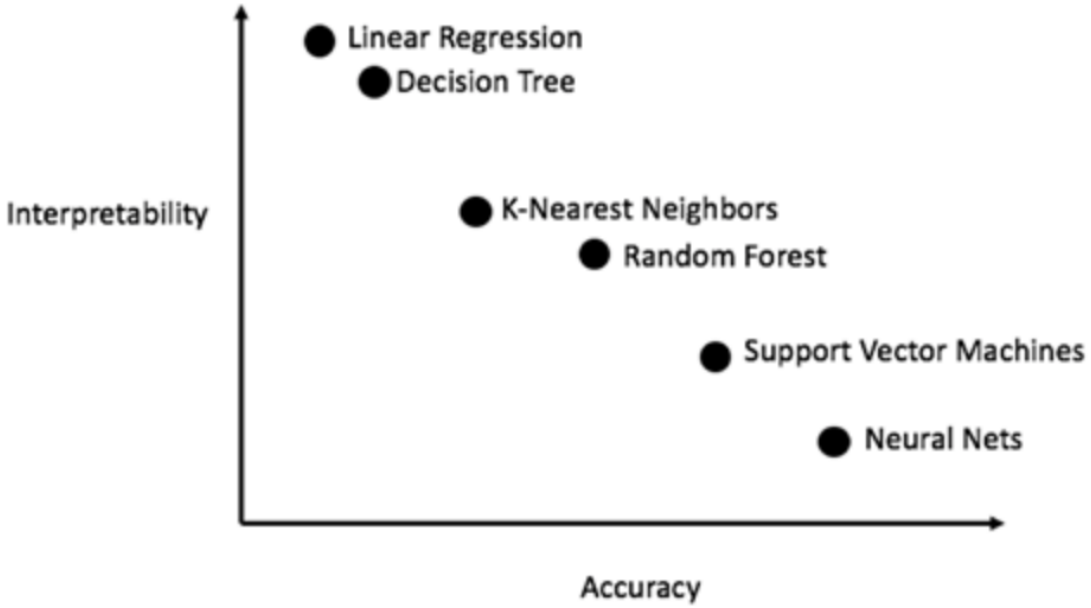
Para encontrar el equilibrio es necesario observar los resultados en train y test.

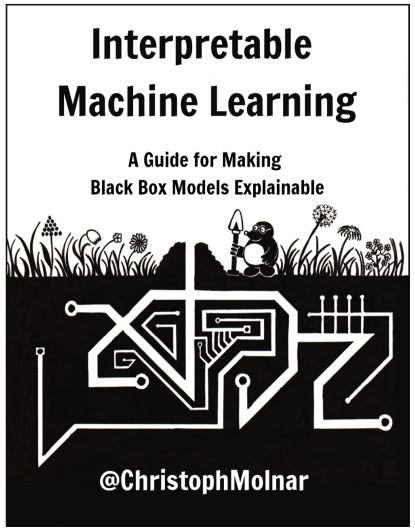


My model on training data

My model on test dataset

Complejidad vs Interpretabilidad





Evaluación de clasificación

Evaluación de modelos de clasificación: Matriz de confusión

- Es una tabla de tamaño m x m, donde m es el número de valores que toma la variable objetivo

- Un clasificador perfecto tendría todos los elementos en la diagonal

Clasificado como …

Clase Original		Spam	No Spam
	Spam	120	40
	No Spam	20	820

Evaluación de modelos de clasificación: Matriz de confusión

Clasificado como …

Clase Original	Spam	No Spam
Spam	Verdaderos Positivos	Falsos Negativos
No Spam	Falsos Positivos	Verdaderos Negativos

Evaluación de modelos de clasificación: Matriz de confusión

Clasificado como …

Clase Original

	Spam	No Spam
Spam	True Positive (TP)	False Negative (FN)
No Spam	False Positive (FP)	True Negative (TN)

Evaluación de modelos de clasificación: Accuracy

- Eficacia: Es la capacidad del modelo de **predecir correctamente** la clase o etiqueta de los registros

$$accuracy = \frac{registros\ clasificados\ correctamente}{total\ de\ registros}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

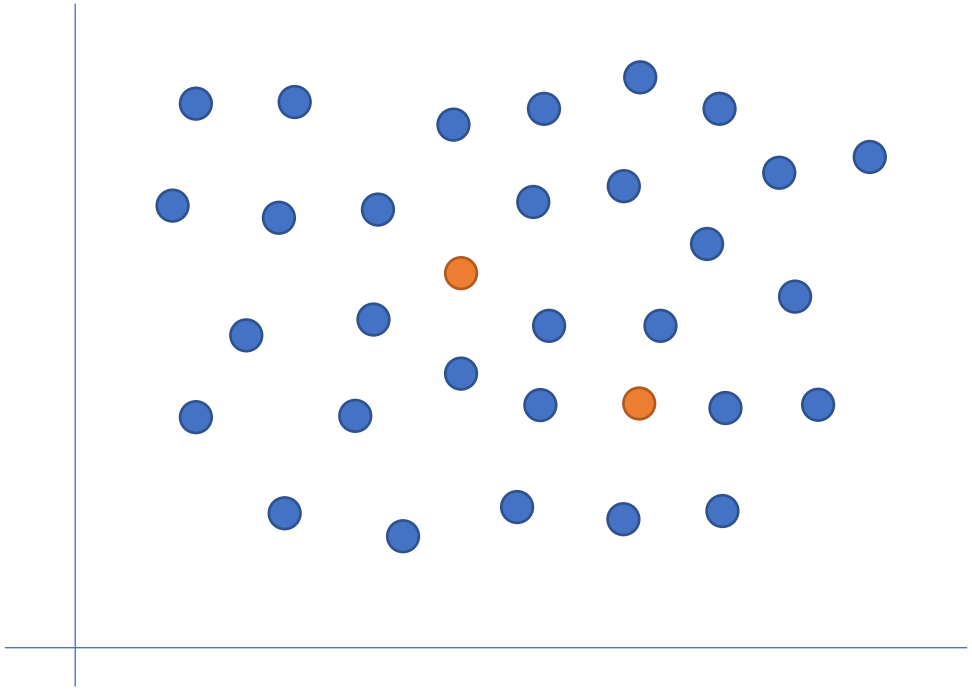
Clase Original

Clasificado como …

	Spam	No Spam
Spam	TP	FN
No Spam	FP	TN

El accuracy no siempre es suficiente para evaluar un modelo

- Si tenemos un modelo clasificador de spam, supongamos que 998 mensajes no son spam y solo 2 son spam
- Si nuestro modelo predice que ningún mensaje es spam nuestra accuracy sería 998 / 1000 = 99.8 %
- El valor es engañoso porque nuestro modelo no detecta ningún correo spam
- En general, el accuracy no es una métrica adecuada cuando tenemos clases desbalanceadas



Evaluación de modelos de clasificación: Precisión

- De los que predije como spam, ¿qué proporción efectivamente era spam?

$$precision = \frac{registros\ positivos\ clasificados\ correctamente}{total\ de\ registros\ clasificados\ como\ positivos}$$

Clasificado como …

Clase Original	No Spam	Spam
No Spam	TN	FP
Spam	FN	**TP**

$$precision = \frac{TP}{TP + FP}$$

Evaluación de modelos de clasificación: Recall

- De todos los que efectivamente son correos spam, ¿qué proporción identifica mi modelo?

$$recall = \frac{\begin{matrix} registros\ positivos \\ clasificados\ correctamente \end{matrix}}{\begin{matrix} total\ de\ registros \\ positivos \end{matrix}}$$

$$recall = \frac{TP}{TP + FN}$$

Clasificado como …

Clase Original	No Spam	Spam
No Spam	TN	FP
Spam	FN	**TP**

Evaluación de modelos de clasificación: F1 score

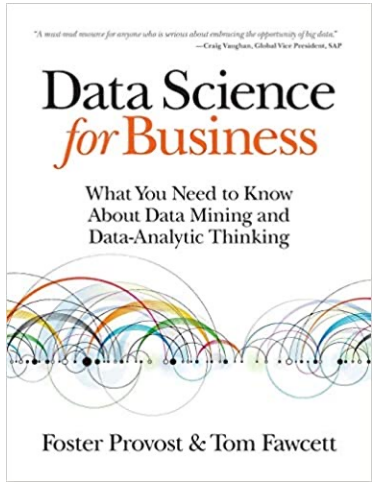
- Es la media armónica de la precisión y el recall

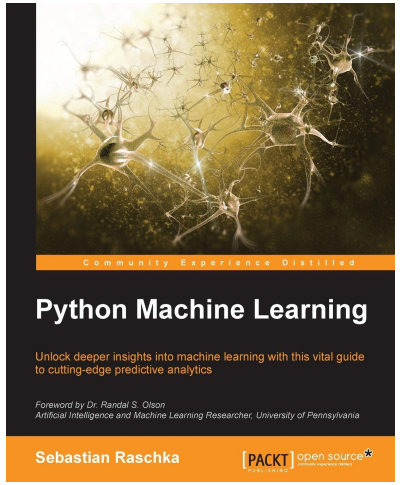
$$f1 = \frac{2 \ * precision \ * recall}{precision + recall}$$

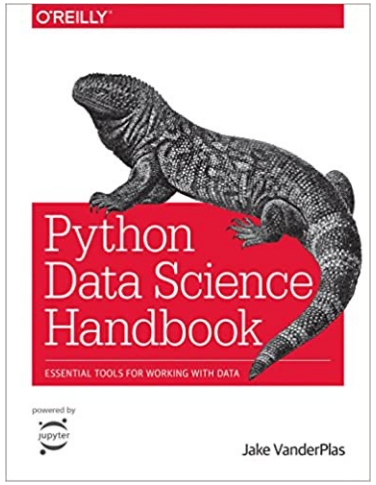
Clasificado como …

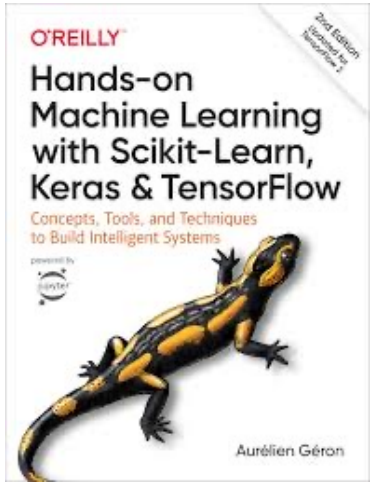
Clase Original		No Spam	Spam
	No Spam	TN	FP
	Spam	FN	**TP**

Bibliografía recomendada









Referencias

- Python Machine Learning – Sebastian Raschka (Tercera edición): https://github.com/rasbt/python-machine-learning-book-3rd-edition
- Documentación de scikit-learn: https://scikit-learn.org/stable/
- Machine learning flashcards – Chris Albon: https://machinelearningflashcards.com
- Interpretable Machine Learning – Cristoph Molnar: Interpretable Machine Learning: https://christophm.github.io/interpretable-ml-book/