



Machine Learning Advanced

- Aprendizaje Supervisado -

Introducción a Machine Learning

Docente: Manuel Montoya



Expositor: Manuel Montoya

<https://www.linkedin.com/in/manuel-montoya-gamio/>

manuel.montoya@pucp.edu.pe



Ingeniería Informática

Diplomado en Inteligencia Artificial



Micromaster in Statistics and Data Science



Data Science and Big Data Analytics



Deep Learning Nanodegree



Data Engineer Nanodegree



Data Scientist



Data Scientist

Data Engineer

REGLAS



Se requiere **puntualidad** para un mejor desarrollo del curso.



Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.



Las preguntas se realizarán **a través del chat** y en caso de que lo requieran **podrán activar el micrófono**.



Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.



Identificarse en la sala Zoom con el primer nombre y primer apellido.

Agenda

1. Logística del curso
2. Machine Learning: Tipos de aprendizaje
3. Disyuntivas en el desarrollo de modelos

1. Logística del curso

Contenido del curso

Aprendizaje supervisado

- Técnicas de preprocessamiento y balanceo de target.
- Validación cruzada. Optimización de hiperparámetros.
- Modelos ensamblados: Bagging, Boosting, Stacking

Práctica: Aprobación automática de préstamos con XGBoost

Aprendizaje no supervisado

- Clustering

Práctica: Segmentación de clientes de un supermercado

Sistemas de recomendación

- Factorización de matrices

Práctica: Recomendador de películas con movielens

Contenido del curso

Minería de textos

- Análisis exploratorio. Wordclouds
- Representación de texto
- Detección de spam automática

Práctica:

- Segmentación de películas según su contenido
- Filtro automático de SPAM

Introducción a Deep Learning

- Redes neuronales: Arquitectura de redes, feedforward, backpropagation, dropout, funciones de activación
- Redes neuronales convolucionales. Feature extraction, clasificación de imágenes
- Redes neuronales recurrentes.

Práctica: Clasificador automático de objetos (Fashion MNIST, CIFAR)

CALIFICACIÓN

Asistencia (Curso):

mínimo 80% sesiones para recibir la certificación

Proyecto Final

(60%)

+

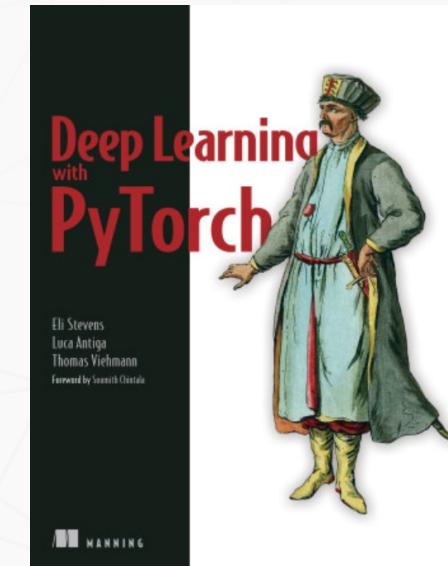
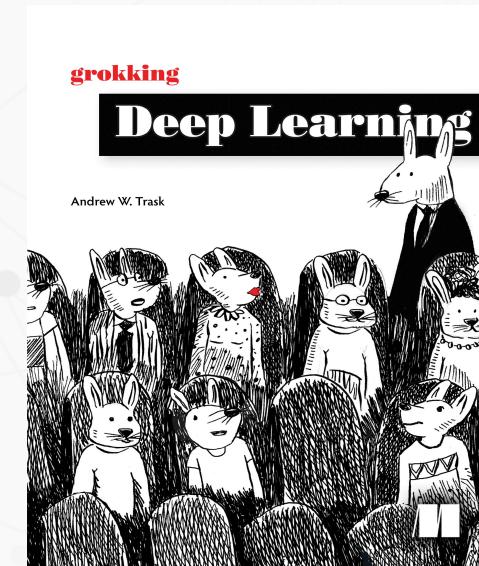
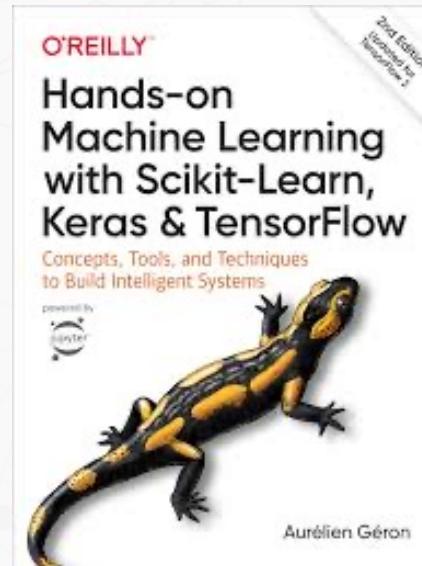
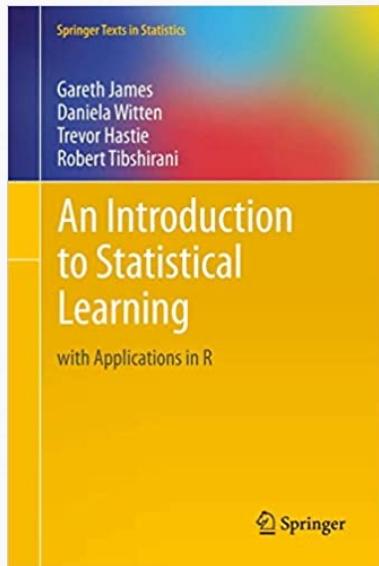
Trabajos

(40%)

Herramientas de desarrollo



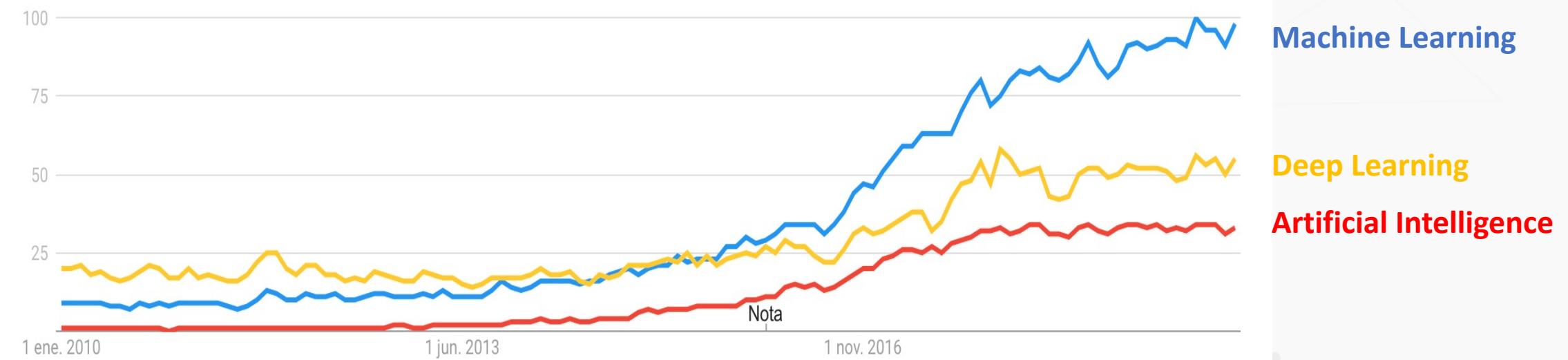
Recursos extras



2. Machine Learning

Tipos de aprendizaje

Google Trends en términos de búsqueda desde el 2010



Machine Learning

Deep Learning

Artificial Intelligence

Visión General

Inteligencia Artificial

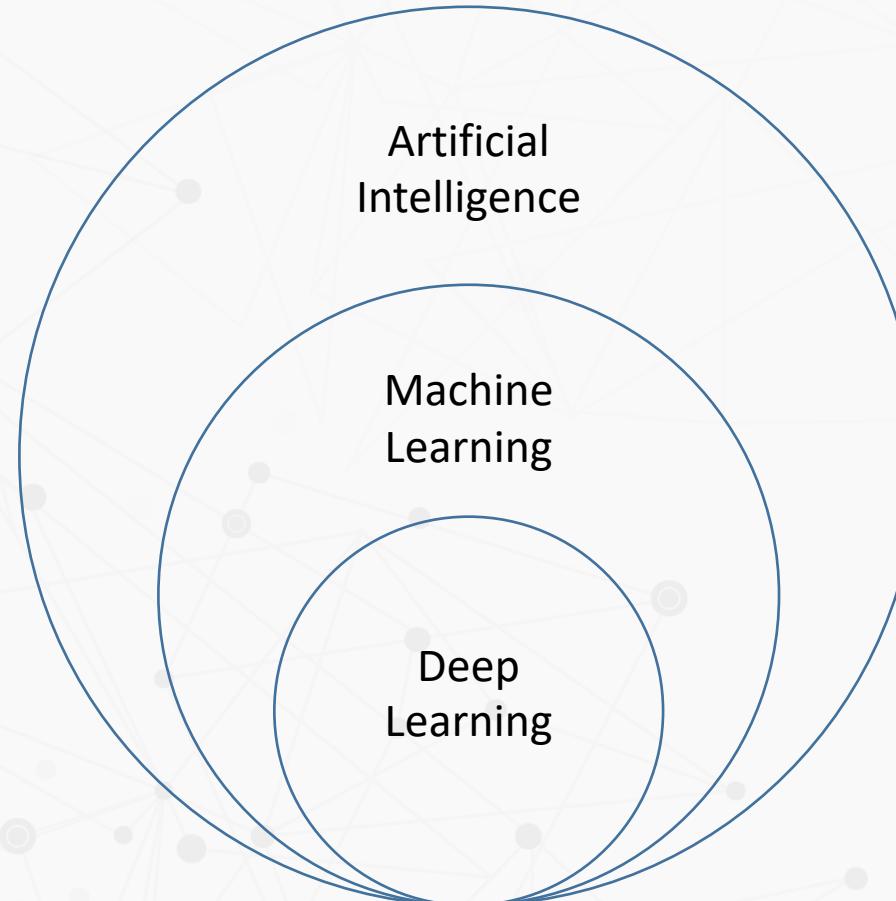
Emular el comportamiento humano

Machine Learning

Aprender a realizar tareas a partir de
experiencias previas

Deep Learning

Mejorar el aprendizaje y representación
de los datos



¿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

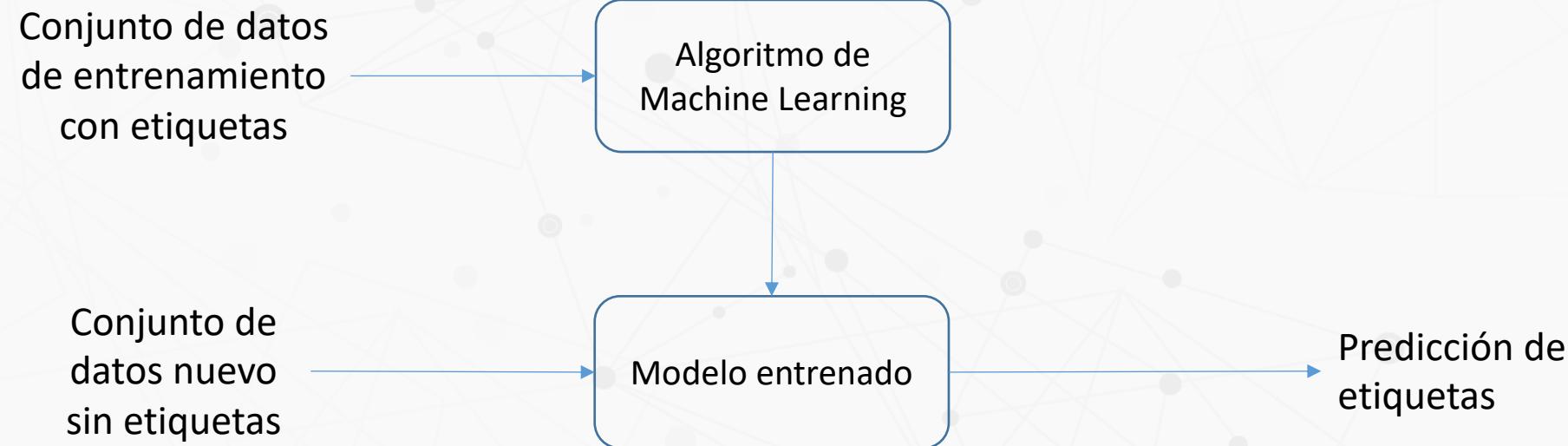
Sueldo	Antiguedad en sistema financiero	Edad	Capacidad de endeudamiento	Pagó el préstamo
				Sí
				Sí
				Sí
				No
				No

Cantidad de muestras

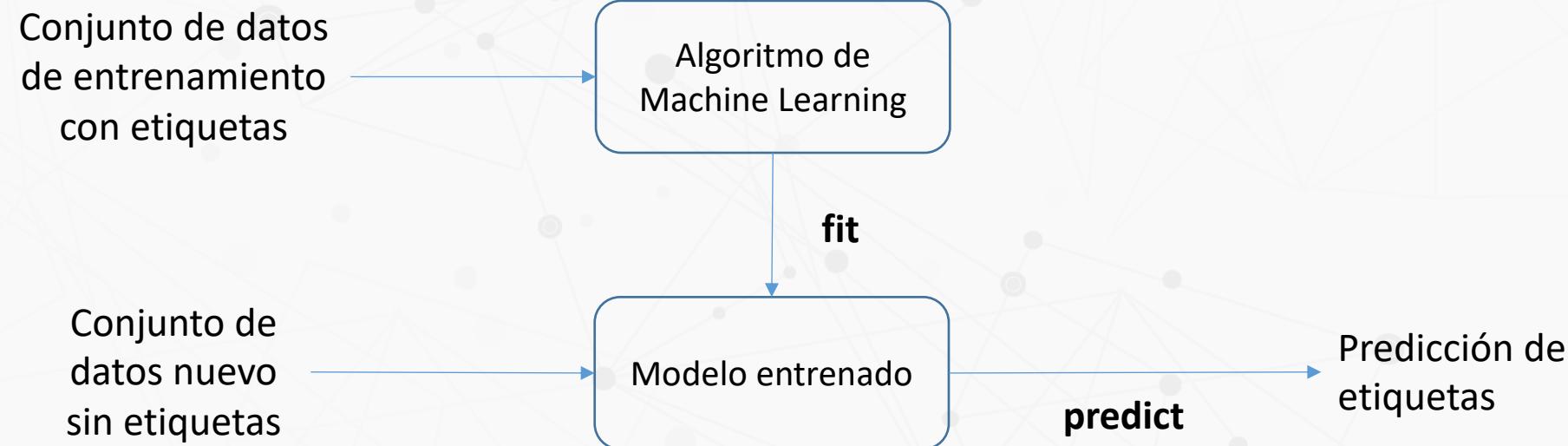
Características
(Features, atributos, dimensiones)

Etiqueta
(Label / Target)

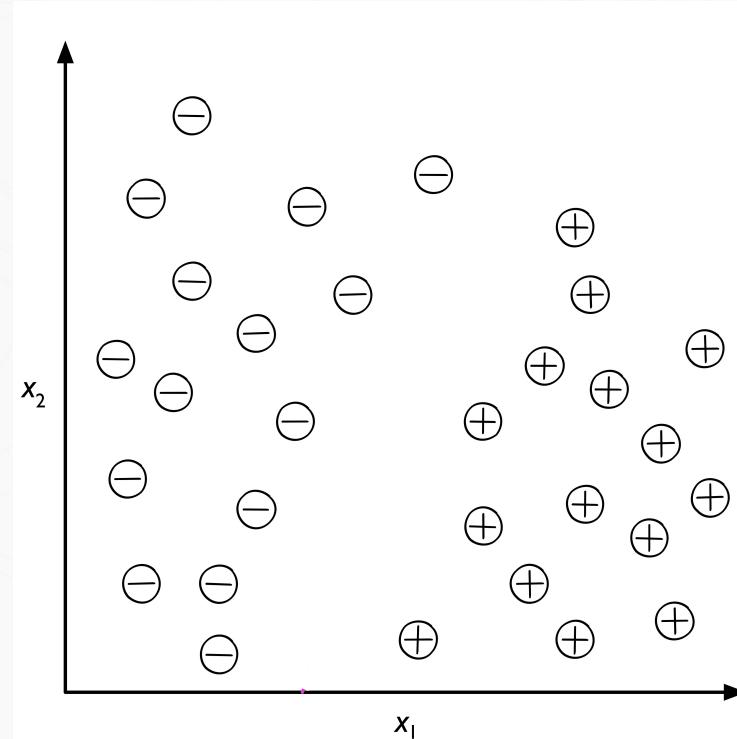
Aprendizaje supervisado: haciendo predicciones sobre el futuro



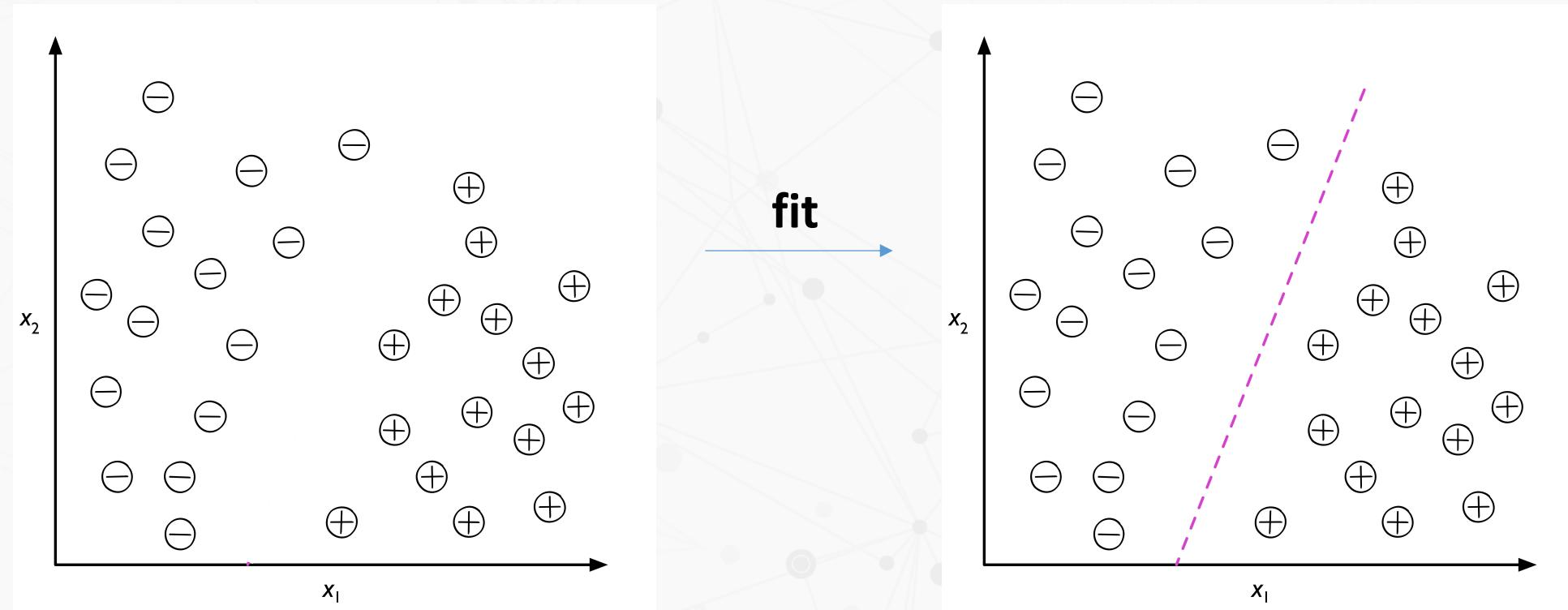
Aprendizaje supervisado: haciendo predicciones sobre el futuro



Aprendizaje supervisado: clasificación para predecir etiquetas



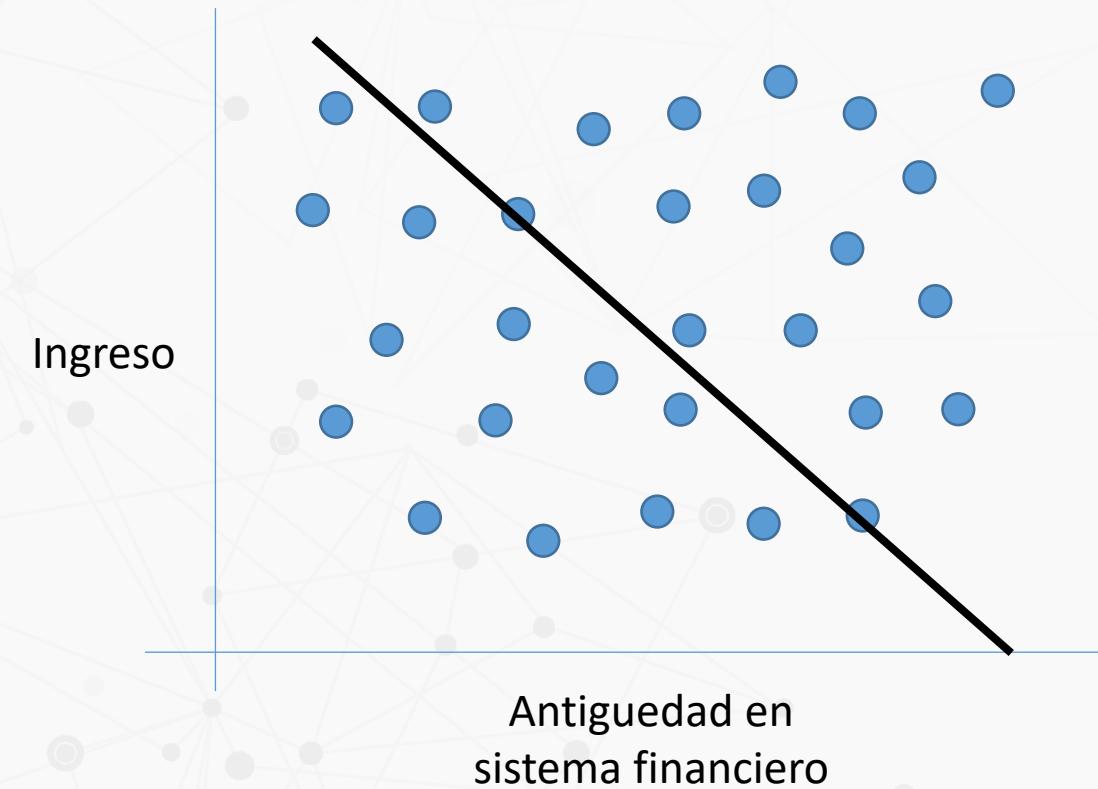
Aprendizaje supervisado: clasificación para predecir etiquetas



Comparación de clasificadores: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

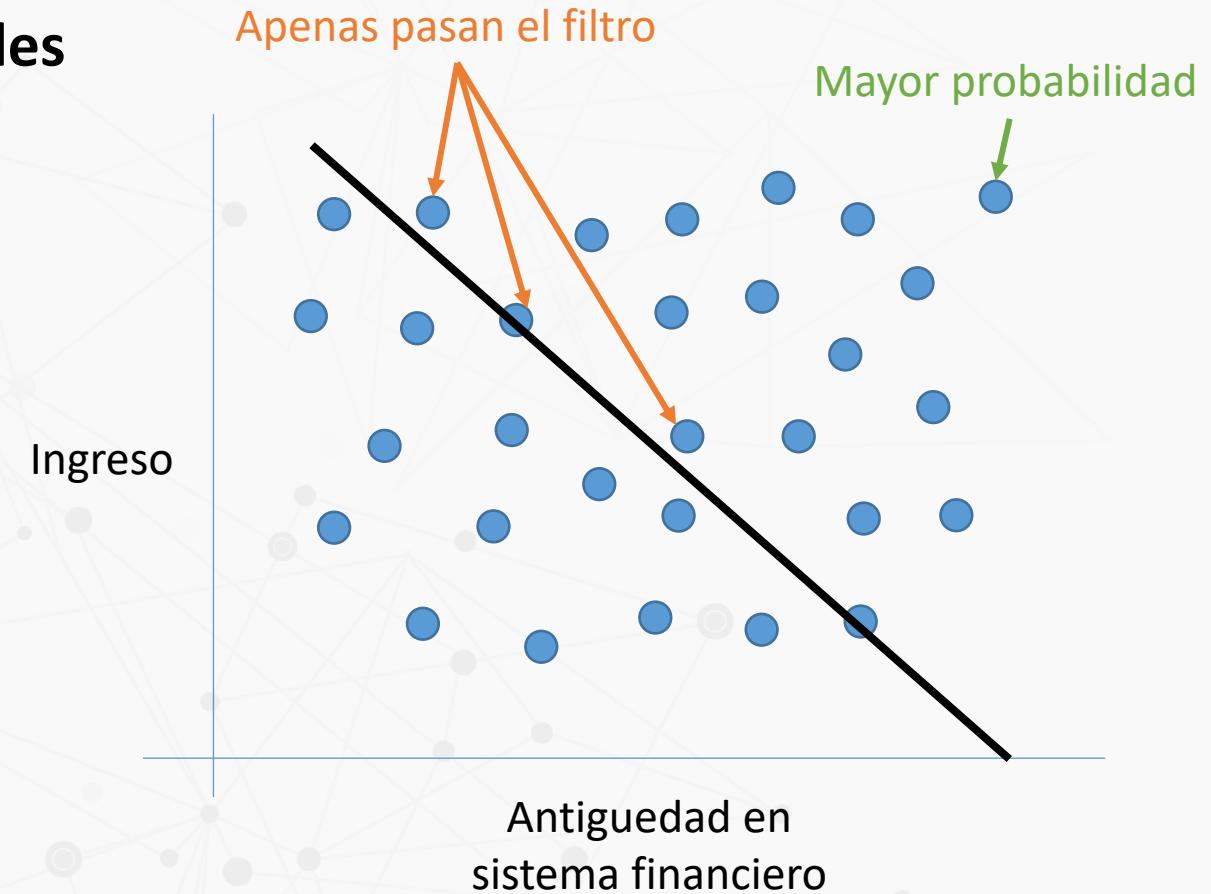
Clasificación: Predicción de probabilidades

- En los modelos de predicción reales, no se predice directamente una clase, sino se predice una probabilidad.
- ¿Qué tan probable es que un cliente me pague un crédito?
- Esto nos permite ordenar los registros en base a la probabilidad de mayor a menor y priorizar aquellos clientes que estamos seguros nos van a pagar.



Clasificación: Predicción de probabilidades

- En los modelos de predicción reales, no se predice directamente una clase, sino se predice una probabilidad.
- ¿Qué tan probable es que un cliente me pague un crédito?
- Esto nos permite ordenar los registros en base a la probabilidad de mayor a menor y priorizar aquellos clientes que estamos seguros nos van a pagar.



Clasificación: Predicción de probabilidades

- En los modelos de predicción reales, no se predice directamente una clase, sino se predice una probabilidad.
- ¿Qué tan probable es que un cliente me pague un crédito?
- Esto nos permite ordenar los registros en base a la probabilidad de mayor a menor y priorizar aquellos clientes que estamos seguros nos van a pagar.

Clientes



+ Mayor probabilidad de pago

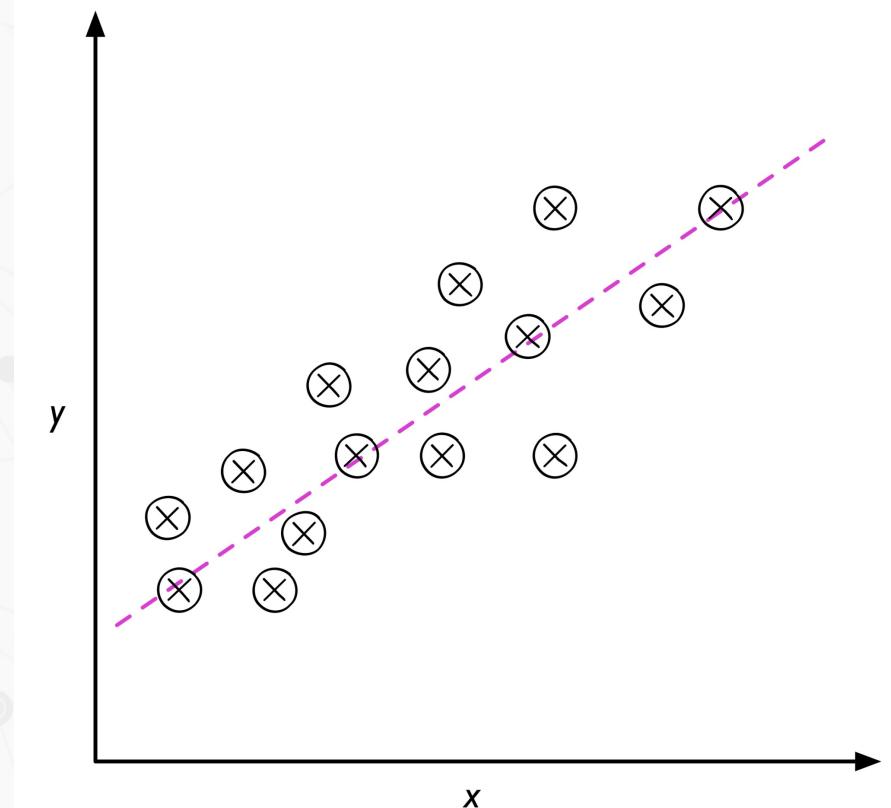
- Menor probabilidad de pago

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

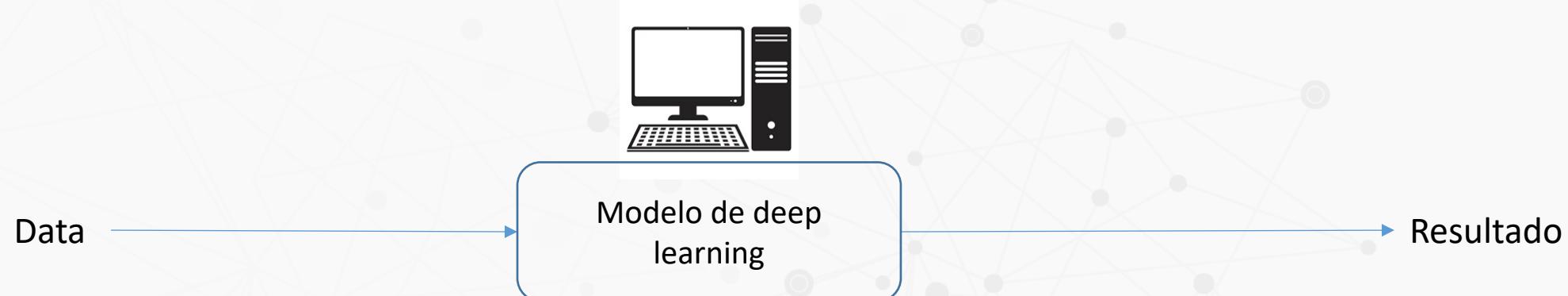
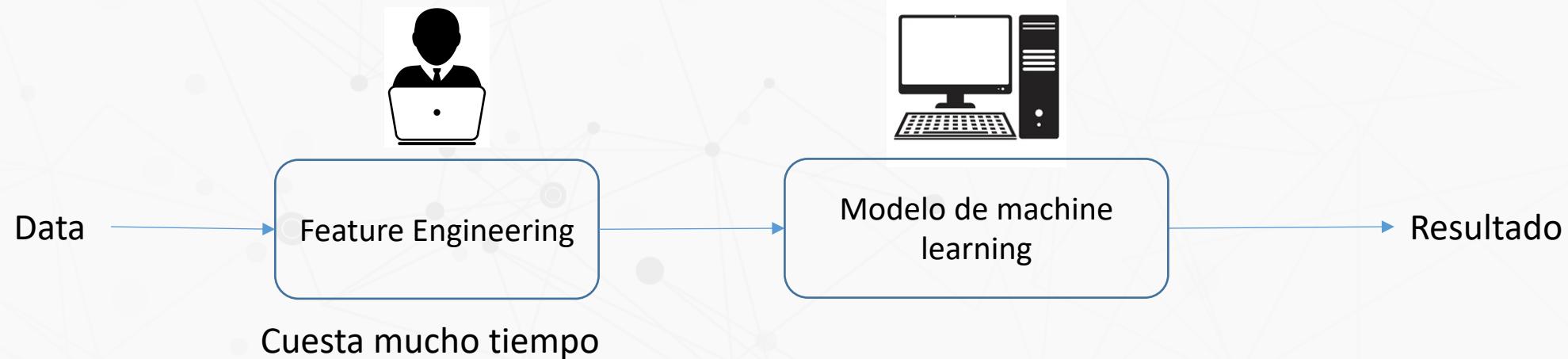
La línea elegida (el modelo) es resultado de ajustar un modelo en el que se intenta minimizar la distancia entre los valores reales y la predicción del modelo.



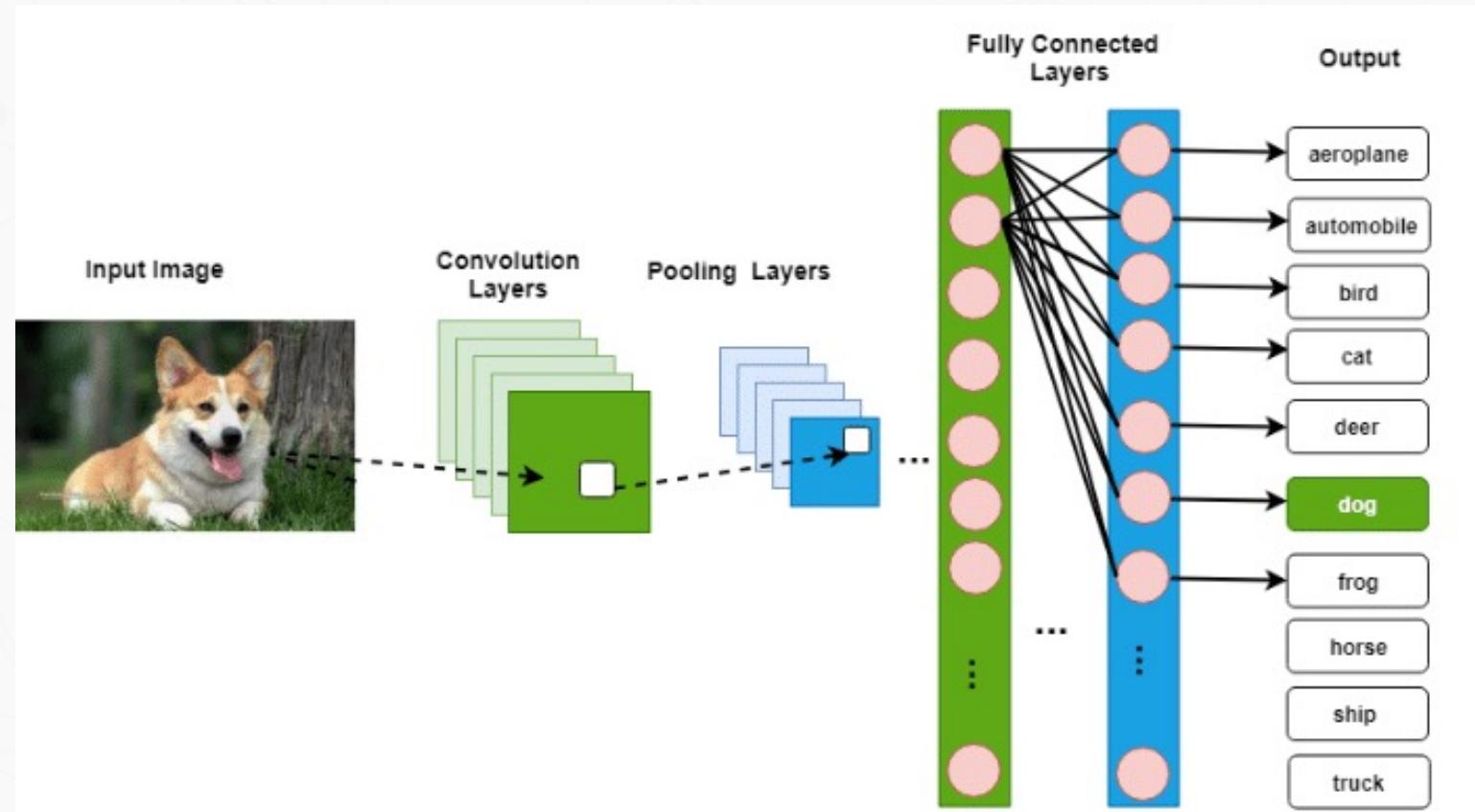
¿Qué es Deep Learning?

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)



¿Qué es Deep Learning?



¿Qué es Deep Learning?



'Godfathers of AI' honored with Turing Award, the Nobel Prize of computing
<https://www.theverge.com/2019/3/27/18280665/ai-godfathers-turing-award-2018-yoshua-bengio-geoffrey-hinton-yann-lecun>

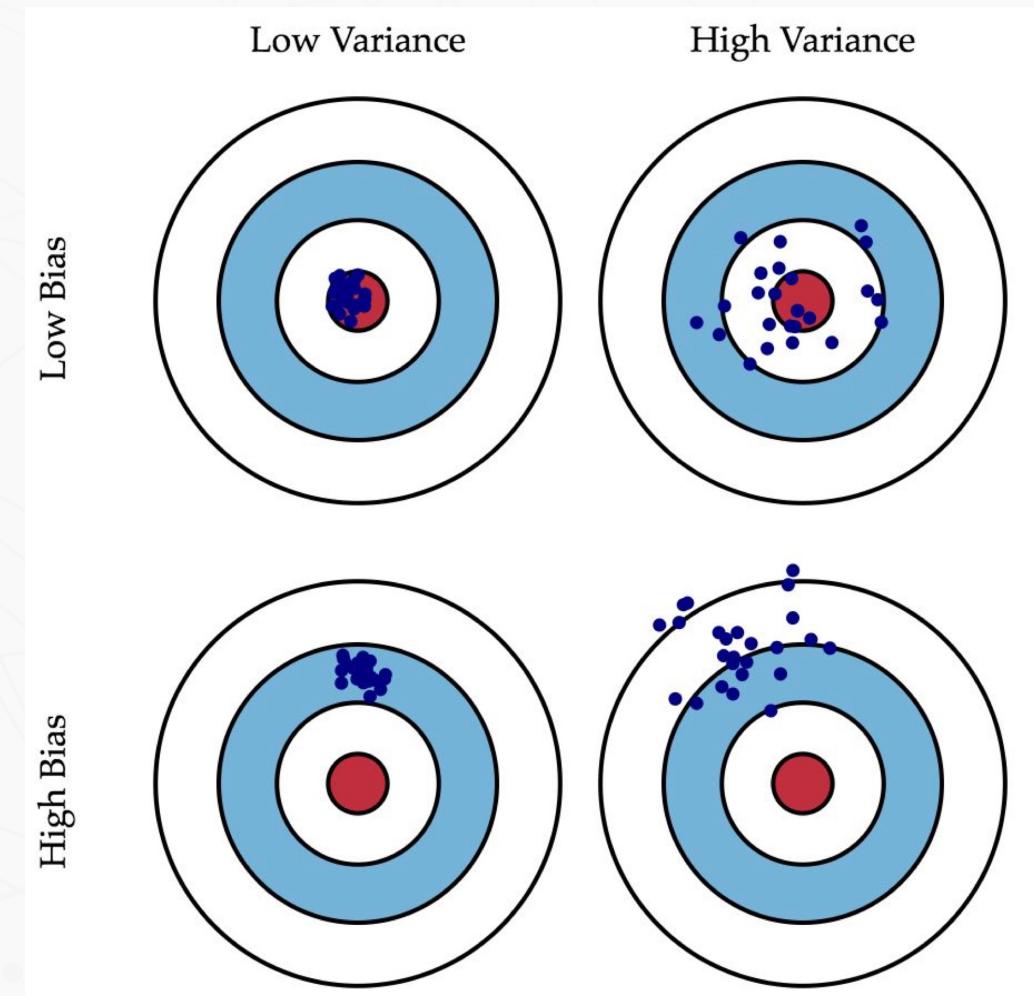
3. Disyuntivas en el desarrollo de modelos

3.1 Fuentes de error

Un modelo predictivo intenta **estimar** la realidad.
Debido a esto nunca va a ser 100% exacto.

El error obtenido proviene de tres fuentes

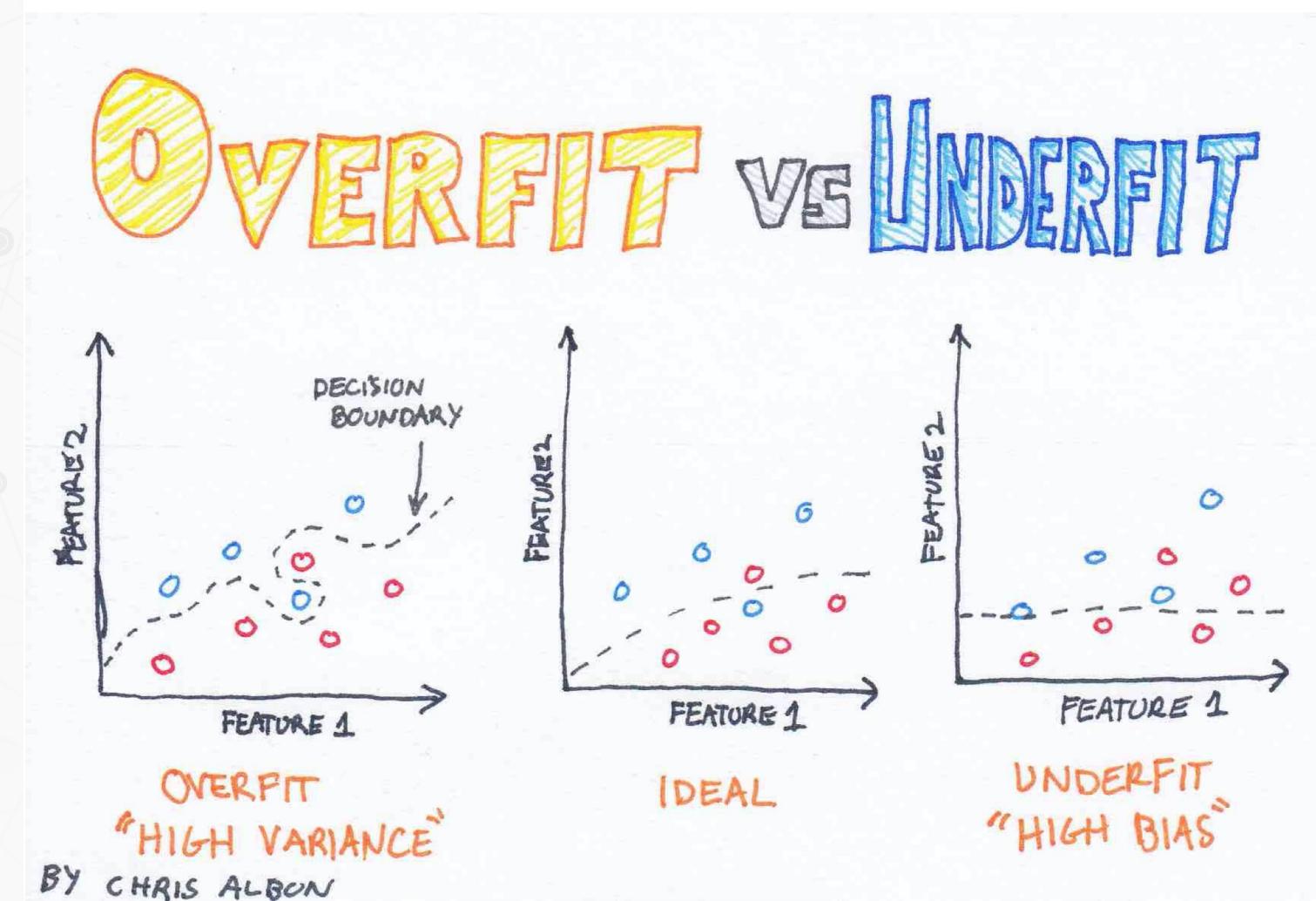
- **Bias:** Diferencia entre la predicción y el valor real
- **Varianza:** Error debido a la variabilidad de la predicción promedio del modelo
- **Error irreducible:** Ruido presente en la generación de muestras



3.2 Overfitting y underfitting

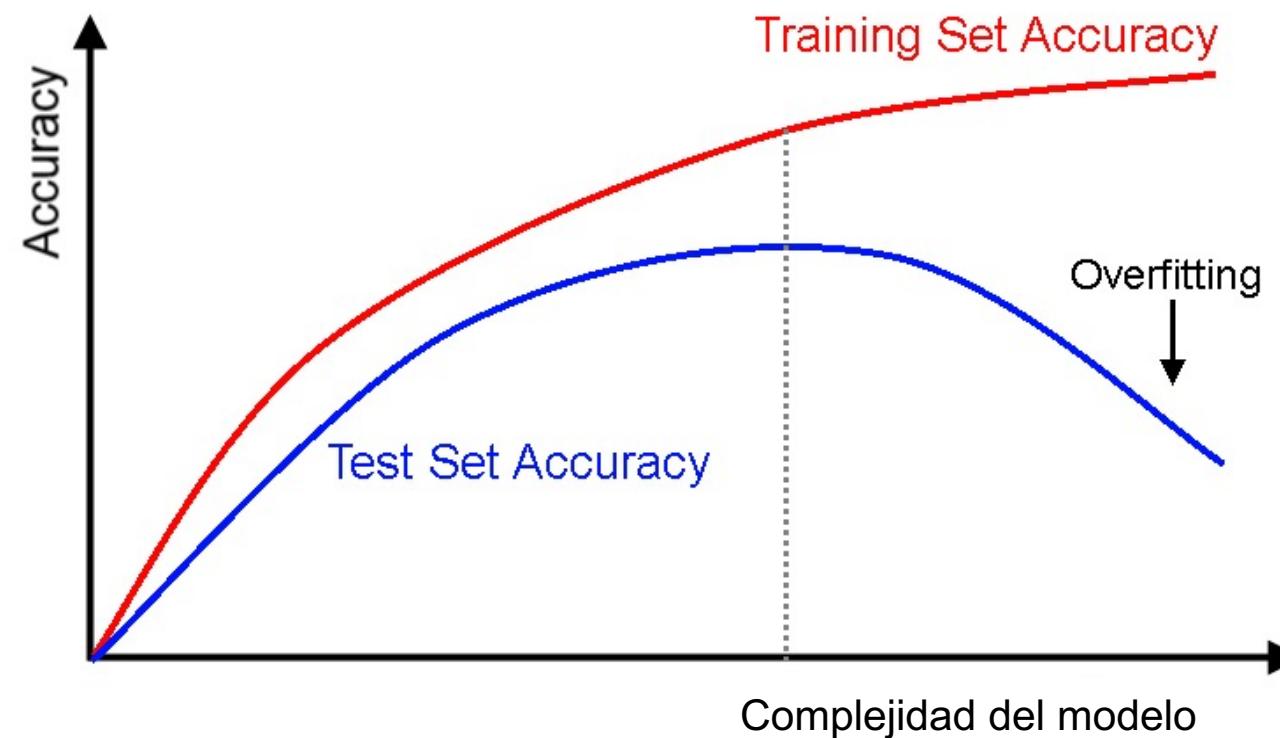
Overfitting: Cuando un modelo se sobreajusta demasiado a los datos de entrenamiento no generaliza sus resultados adecuadamente

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



3.2 Overfitting y underfitting

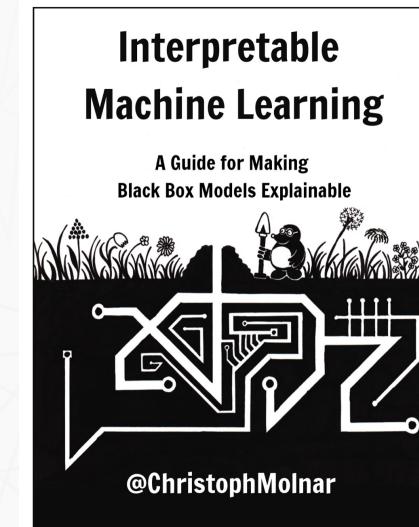
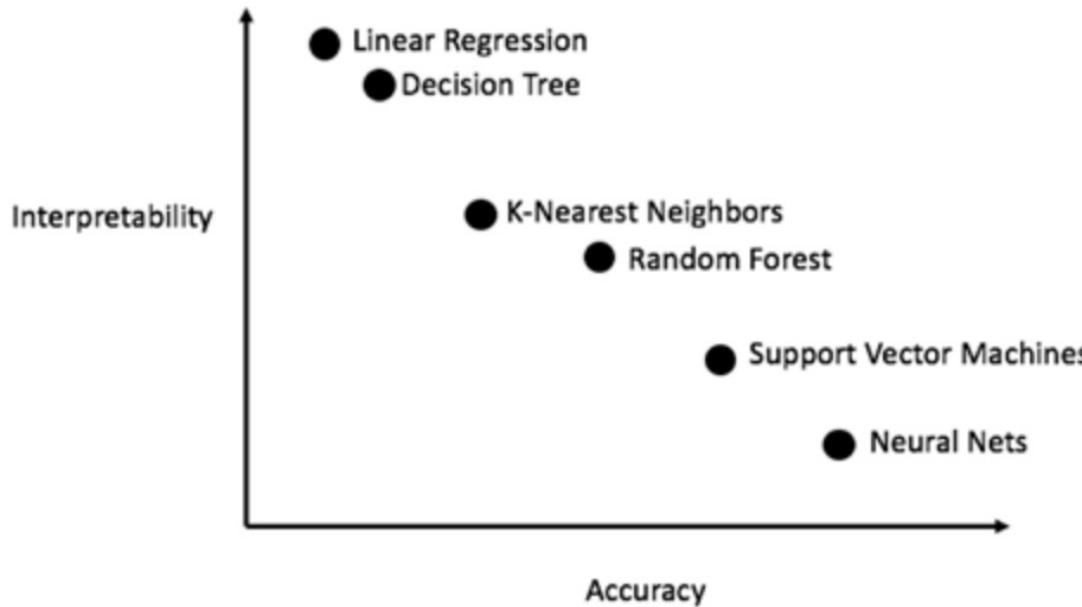
Para encontrar el equilibrio entre overfitting y underfitting es necesario observar los resultados en los conjuntos de train y test.



3.2 ¿Cómo combatir el overfitting?

- Utilizar modelos simples
- Utilizar conjuntos de prueba y validación
- Cross-validation
- Regularización
- Obtener una mayor muestra de datos
- Algoritmos ensamblados
- Detención temprana

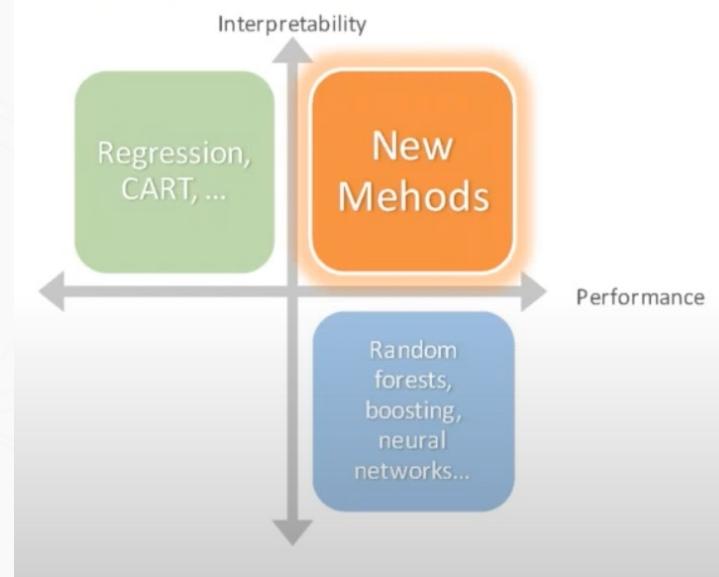
3.3 Complejidad vs Interpretabilidad



Interpretable Machine Learning: <https://christophm.github.io/interpretable-ml-book/>

3.4 Complejidad vs Interpretabilidad

En los últimos años han aparecido nuevos modelos con alta performance y alta interpretabilidad



Optimal classification trees (Dimitris Bertsimas - Jack Dunn)

https://www.mit.edu/~dbertsim/papers/Machine%20Learning%20under%20a%20Modern%20Optimization%20Lens/Optimal_classification_trees_MachineLearning.pdf

Dimitris Bertsimas (MIT) Optimal Classification Trees and Interpretable AI

<https://www.youtube.com/watch?v=WGW0mygEW44>

3.5 Otros tipos de sesgos en modelos predictivos

Los errores asociados a la predicción de un algoritmo no son el único tipo de sesgo posible

Twitter is looking into why its photo preview appears to favor white faces over Black faces

Users discovered the problem with the neural network that crops photo previews

By Kim Lyons | Sep 20, 2020, 4:20pm EDT

<https://www.theverge.com/2020/9/20/21447998/twitter-photo-preview-white-black-faces>

Apple's credit card is being investigated for discriminating against women

Customers say the card offers less credit to women than men

By James Vincent | Nov 11, 2019, 5:57am EST

<https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation>

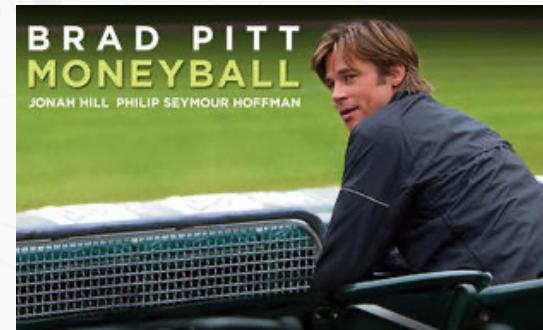
Referencias

- 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/>

Lecturas recomendadas

- Python Machine Learning – Capítulo 1: Giving computers the ability to learn from data
- Data Science for Business – Capítulo 1: Introduction: Data-Analytic Thinking
- Data Science for Business – Capítulo 2: Business Problems and Data Science Solutions

Documentales y películas



¡Gracias!