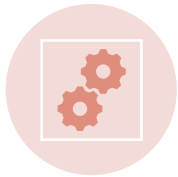




Aprendizaje de máquinas

Ciencia de Datos para físicos

Agenda



Introducción al
aprendizaje de
máquinas



Tipos de
aprendizaje



Algoritmos
populares

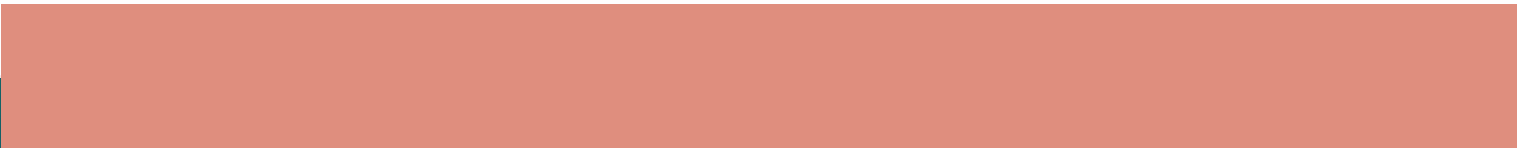


Evaluación de
modelos



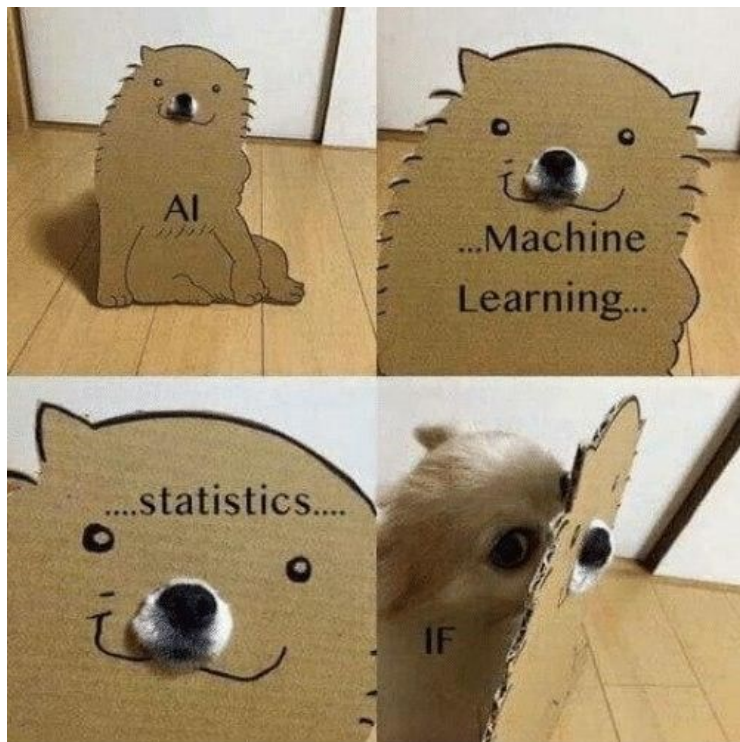
Conceptos
relevantes

Introducción al aprendizaje de máquinas

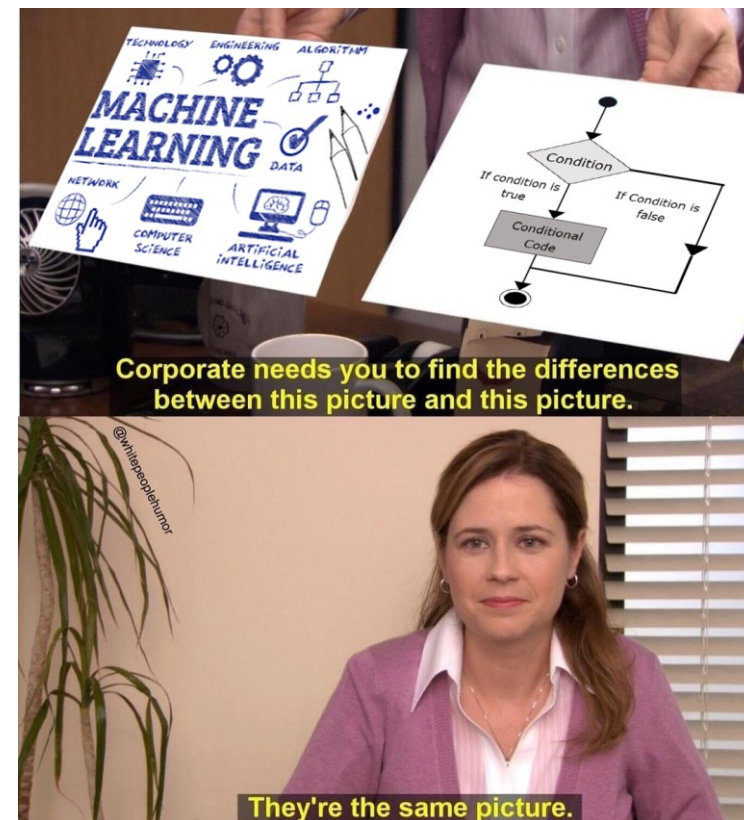
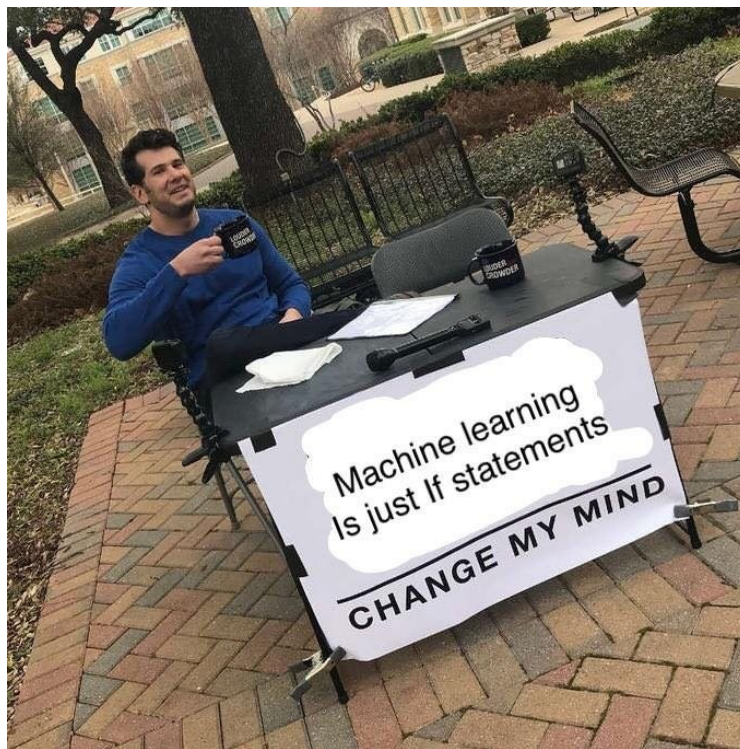


¿Qué es el aprendizaje de máquinas?

Es una rama de la inteligencia artificial que se enfoca en **desarrollar** algoritmos y modelos estadísticos que permiten que **una computadora aprenda a partir de datos**, sin ser programada explícitamente para resolver una tarea específica. El objetivo es que la máquina pueda generalizar patrones y relaciones en los datos de entrenamiento para hacer predicciones precisas sobre nuevos datos.



IF IF IF IF IF IF IF IF WE!





¿Qué problemas puede resolver?

Una amplia gama de problemas, como clasificación, regresión, agrupamiento, detección de anomalías y recomendación, entre otros. Se puede aplicar en diversas áreas, como la medicina, finanzas, marketing, seguridad, entre otras.

¿Cómo se diferencia del aprendizaje automático?

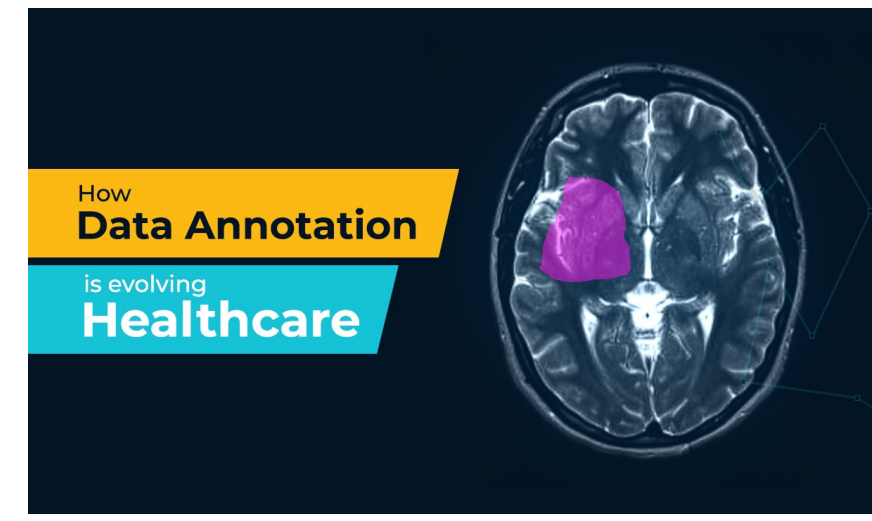
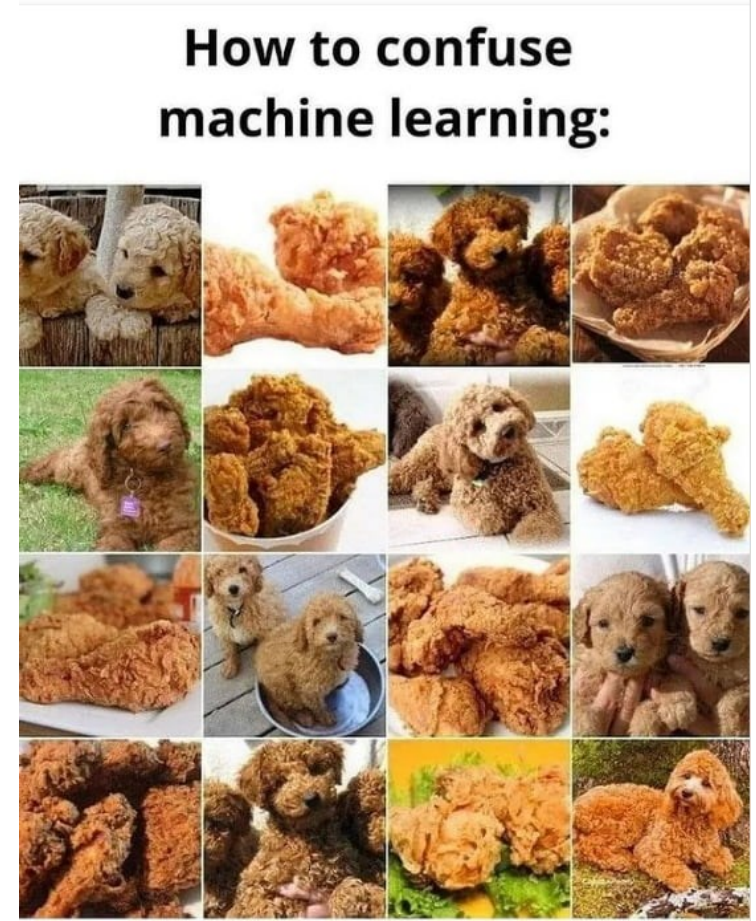
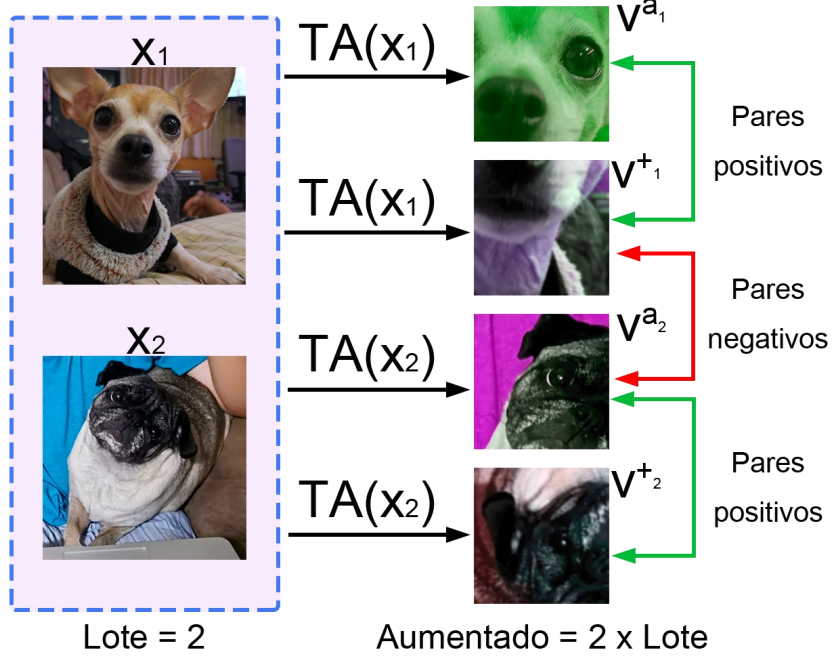
El ***aprendizaje de máquinas*** se enfoca específicamente en el ***desarrollo de algoritmos que permiten que una máquina aprenda a partir de datos***. El ***aprendizaje automático*** se refiere a cualquier tipo de modelo o ***algoritmo que permita que una máquina aprenda***, incluyendo métodos que no se basan en datos, como los algoritmos genéticos.

Tipos de aprendizaje

A solid orange horizontal bar is positioned below the title, spanning most of the width of the white text box.

Aprendizaje supervisado

En este tipo de aprendizaje, el modelo aprende a partir de ejemplos de entrada y salida previamente etiquetados, es decir, se le proporciona información sobre qué respuesta se espera para cada entrada. El objetivo es que el modelo pueda aprender una función que relacione las entradas con las salidas.



Aprendizaje supervisado

El aprendizaje supervisado se divide en dos categorías: la **clasificación** y la **regresión**. En la clasificación, el objetivo es predecir una etiqueta categórica, como identificar si una imagen es de un perro o un gato. En la regresión, el objetivo es predecir un valor numérico, como predecir el precio de una casa basado en características como la ubicación, el tamaño y el número de habitaciones.

Aprendizaje supervisado

Es uno de los enfoques más usados, por su “comodidad”.

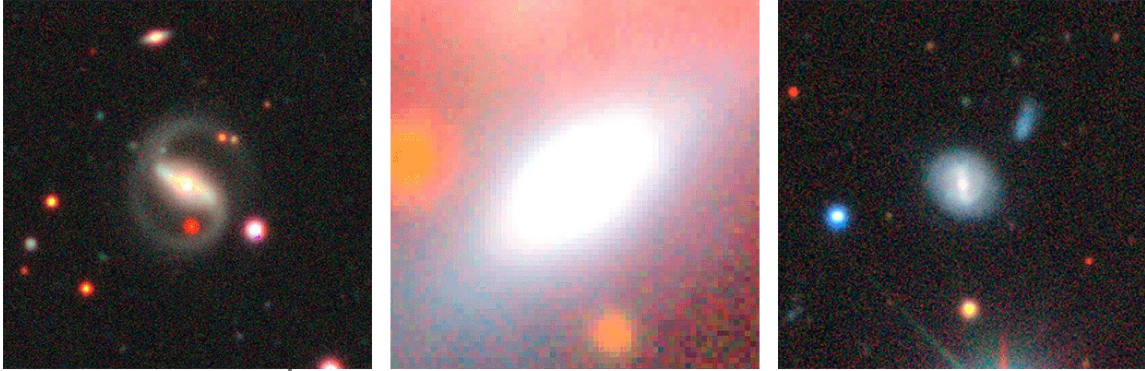
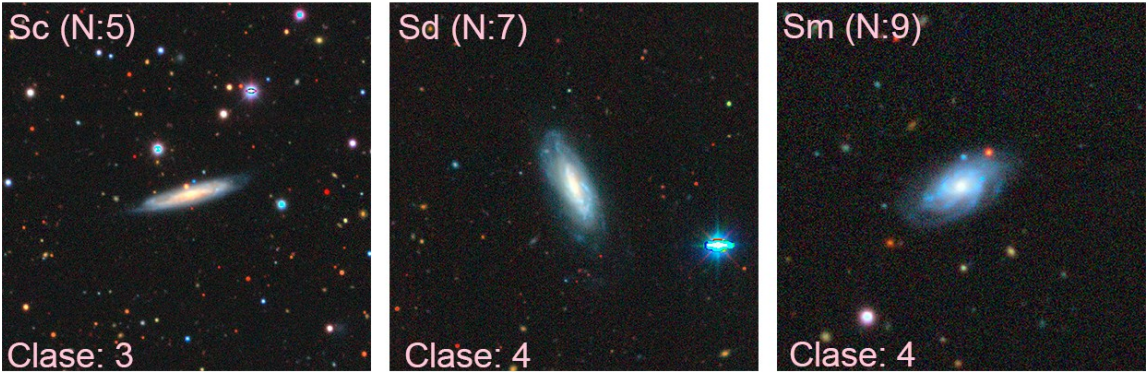
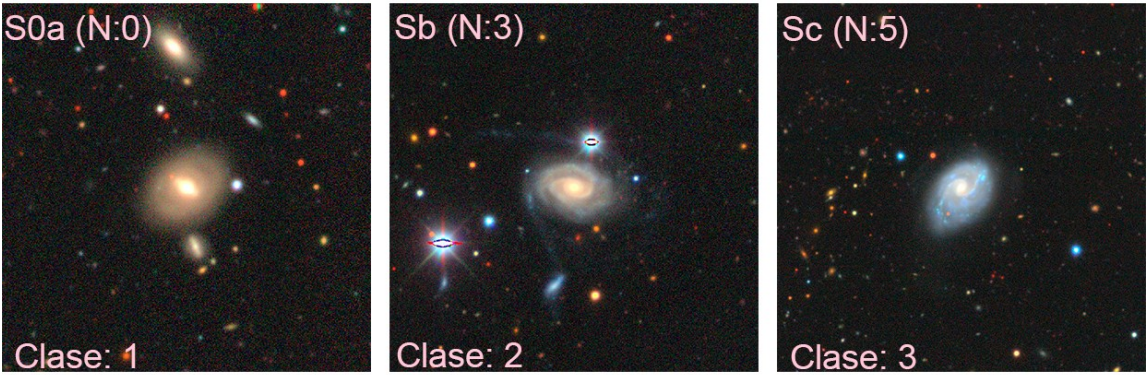
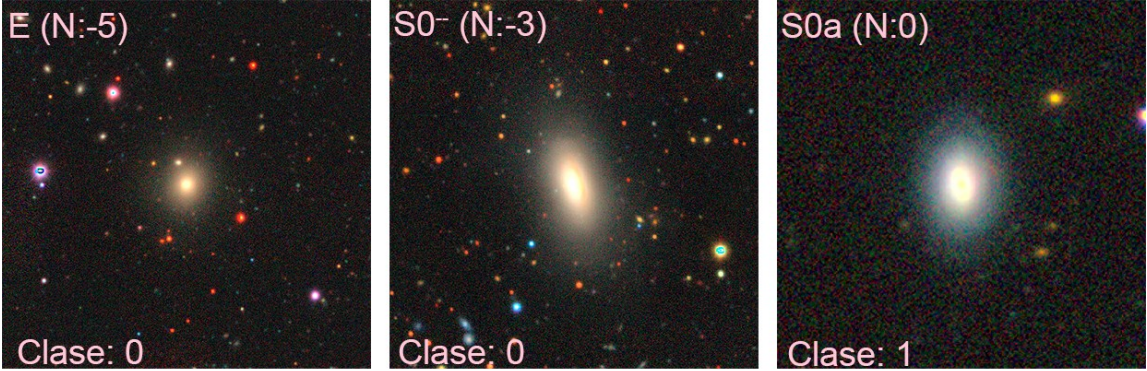
Los resultados son en la mayoría de las veces buenos.

Este enfoque es uno de los más estudiados y conocidos.

El gran inconveniente se encuentra en las etiquetas, sin suficientes datos etiquetados es muy probable que el rendimiento del modelo sea malo.

Es muy costoso etiquetar, debido al volumen y el conocimiento técnico que se requiere.

En la vida real, la gran mayoría del tiempo se tendrán datasets desbalanceados, y el desbalance no es muy bueno, ya que puede sesgar al modelo.



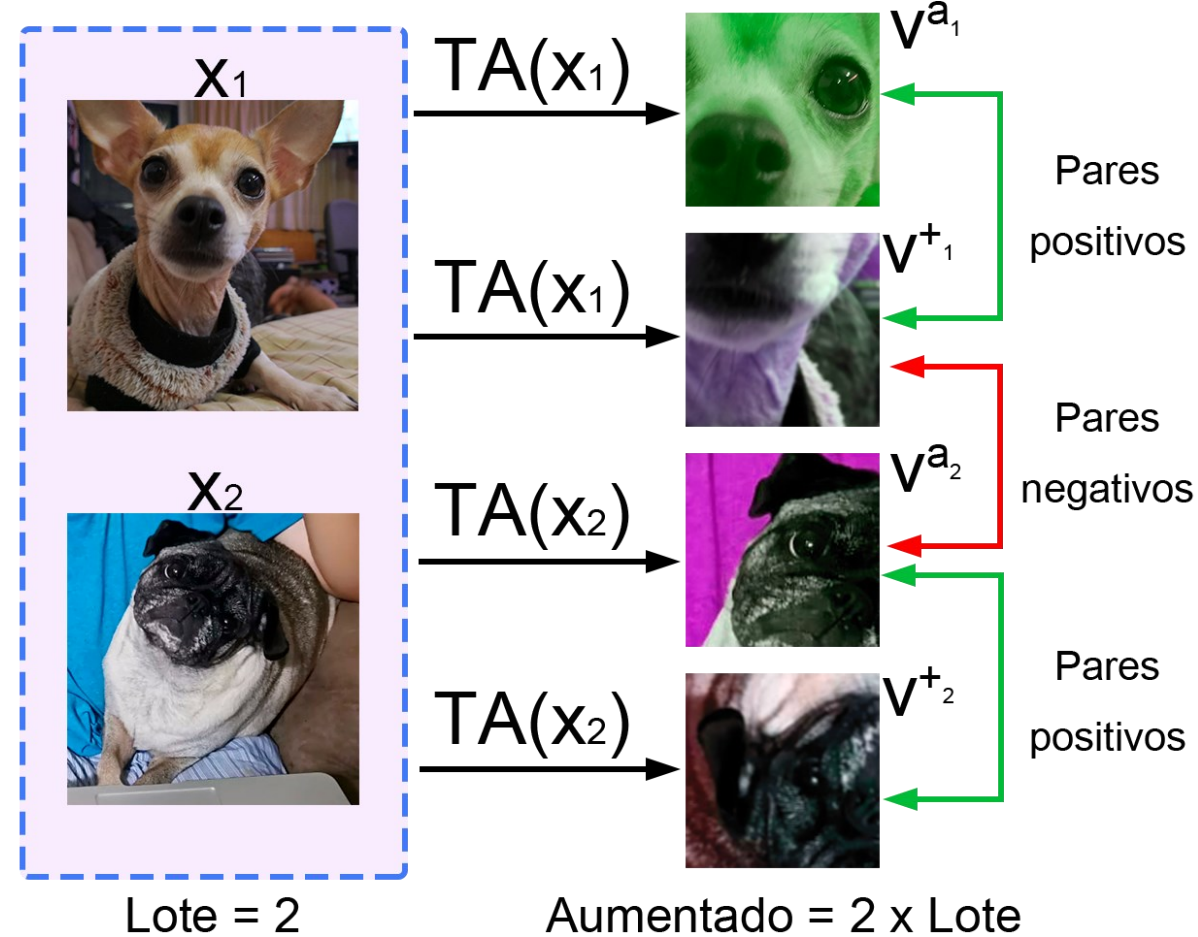
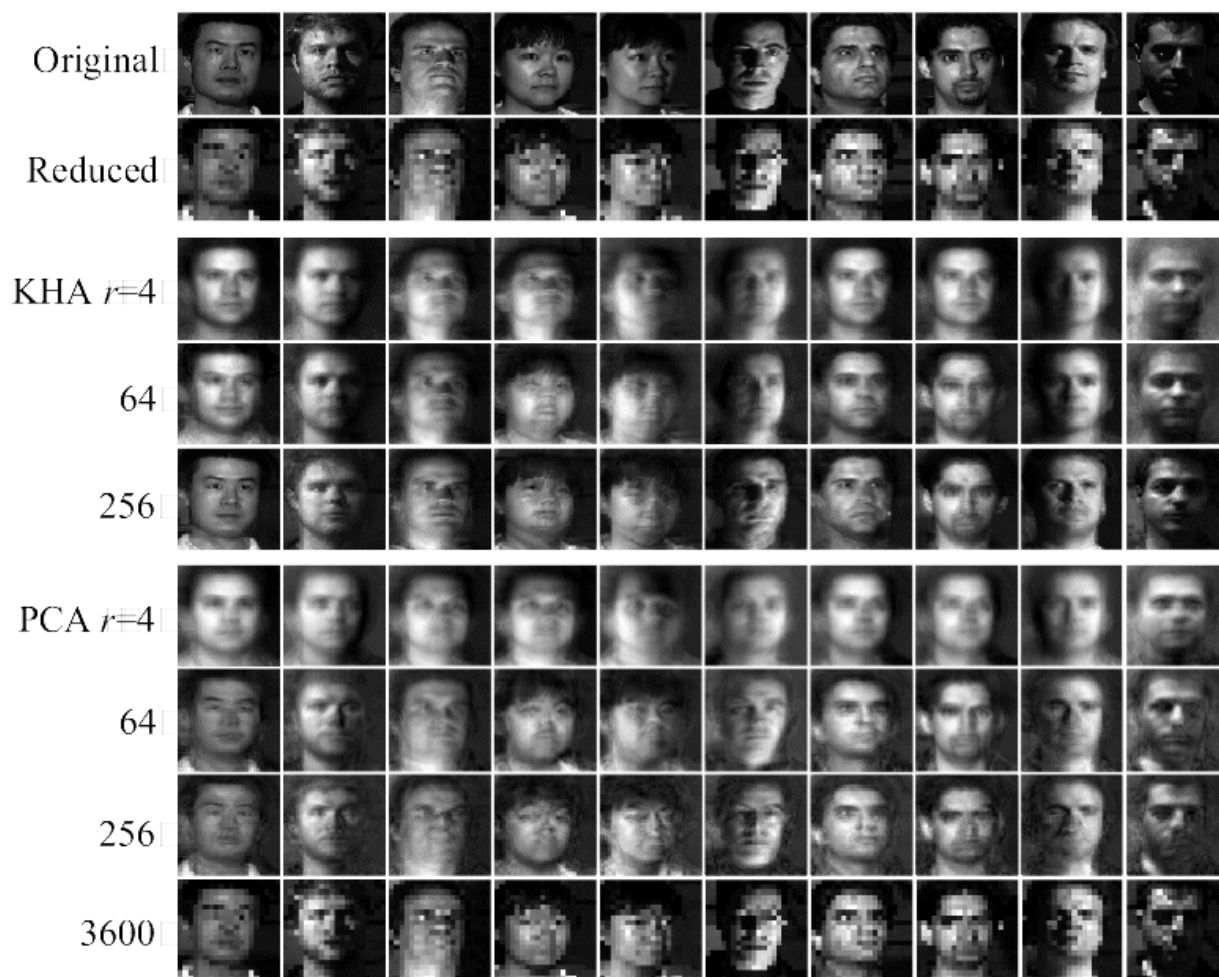
Aprendizaje no supervisado

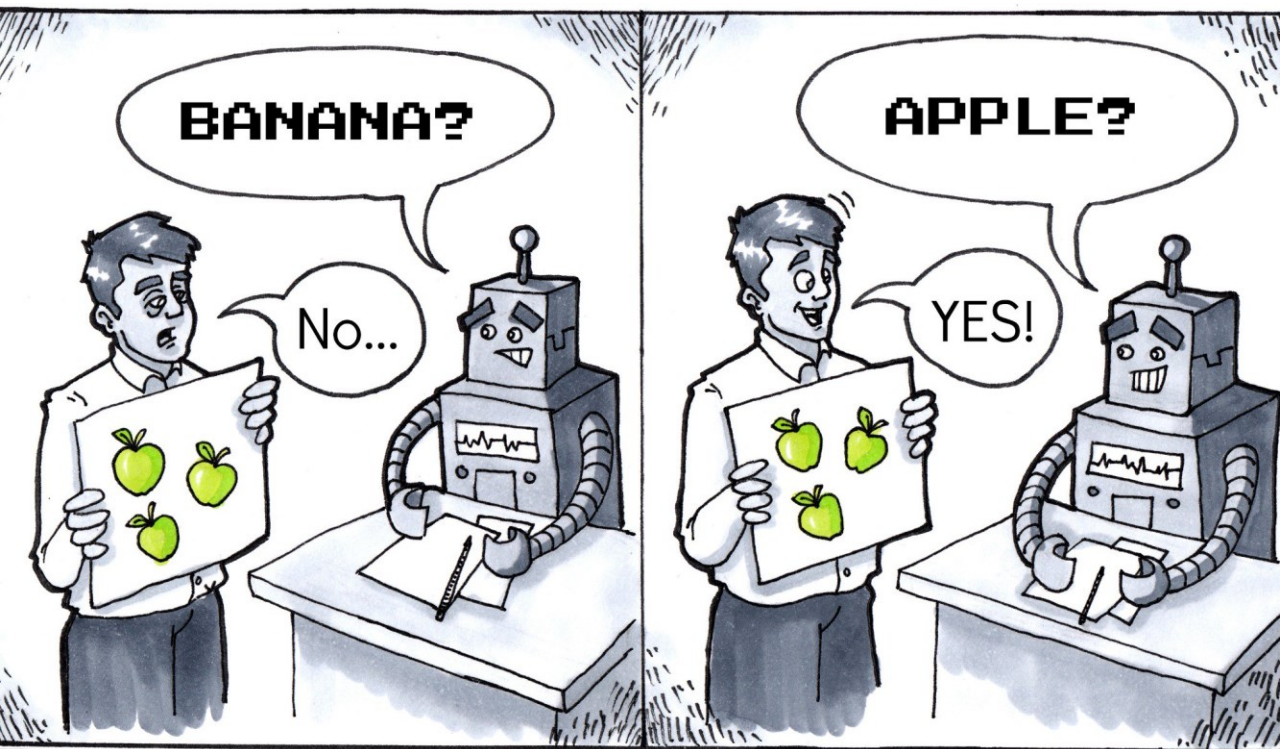
En este tipo de aprendizaje, el modelo no recibe información explícita sobre qué salida se espera para cada entrada, sino que debe descubrir patrones y estructuras por sí mismo.

El objetivo es identificar patrones interesantes en los datos y agruparlos en categorías o clústeres. Esto puede ayudar a revelar relaciones ocultas y simplificar la comprensión y análisis de grandes conjuntos de datos.

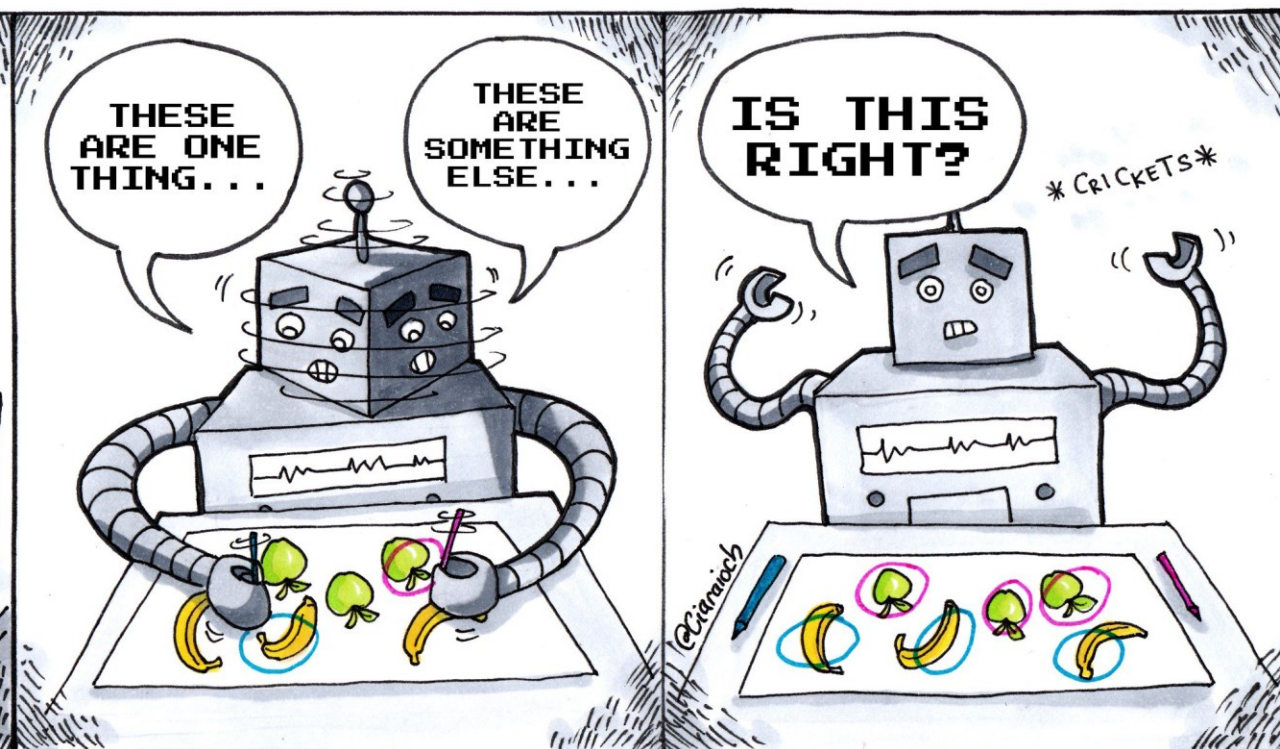
Aprendizaje no supervisado

Se puede utilizar en una variedad de aplicaciones, como la detección de anomalías, la reducción de la dimensionalidad y la exploración de datos. Algunos ejemplos de algoritmos de aprendizaje no supervisado son el agrupamiento k-means, análisis de componentes principales (PCA) y el análisis de conglomerados jerárquicos (HCA).





Supervised Learning



Unsupervised Learning

Algoritmos populares

A solid orange horizontal bar is positioned below the title, spanning most of the width of the white text box.

Regresión lineal

establecer una relación lineal entre una variable de entrada predictor (X) y una variable de salida o respuesta (Y). El objetivo de la regresión lineal es encontrar la línea recta que mejor se ajuste a los datos para poder hacer predicciones precisas sobre nuevos datos.

$$y = mx + b = \theta^T X$$

Regresión lineal

Se buscan los valores óptimos de los parámetros m y b que minimizan la diferencia entre los valores predichos por el modelo y los valores reales observados en los datos de entrenamiento. Esta diferencia se mide a través de una función de costo.

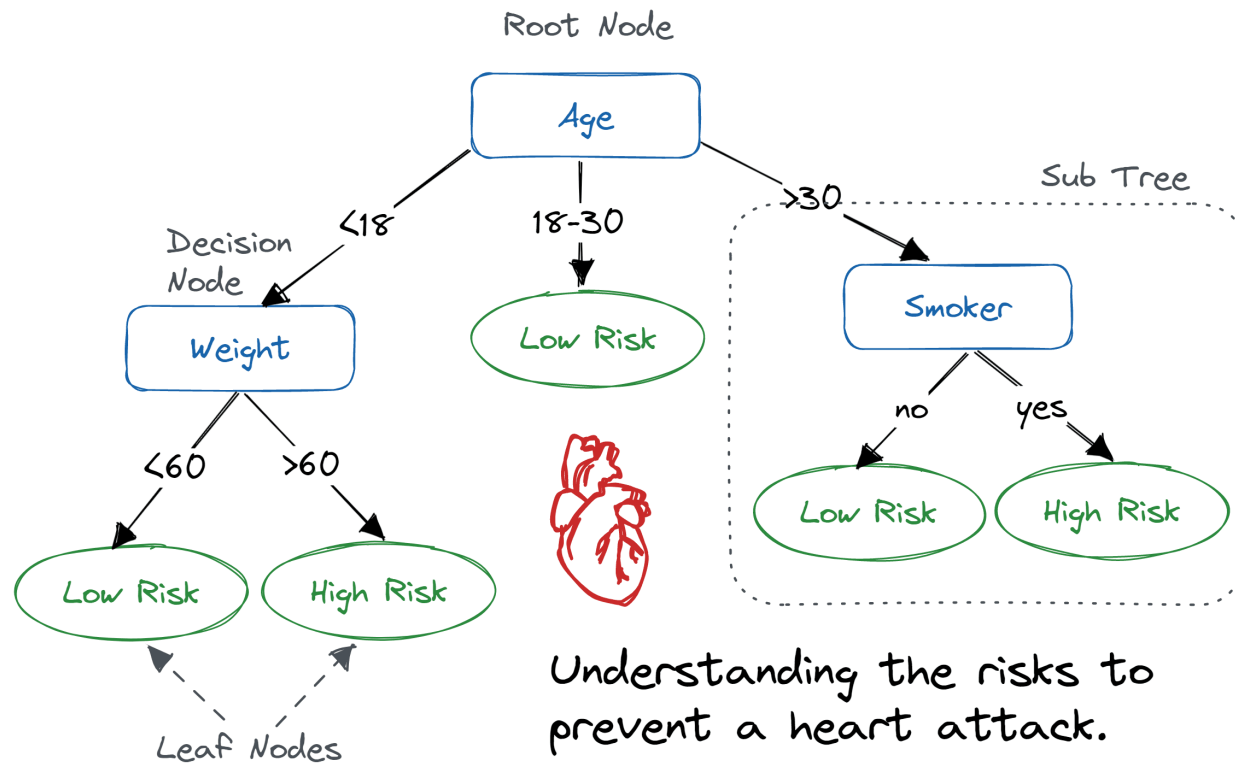


Árboles de decisión

En los árboles de decisión, se divide el conjunto de datos en subconjuntos más pequeños y homogéneos en términos de la variable de salida. Esto se logra a través de una serie de preguntas o decisiones basadas en las características de los datos.

Pueden usarse para problemas de regresión o clasificación

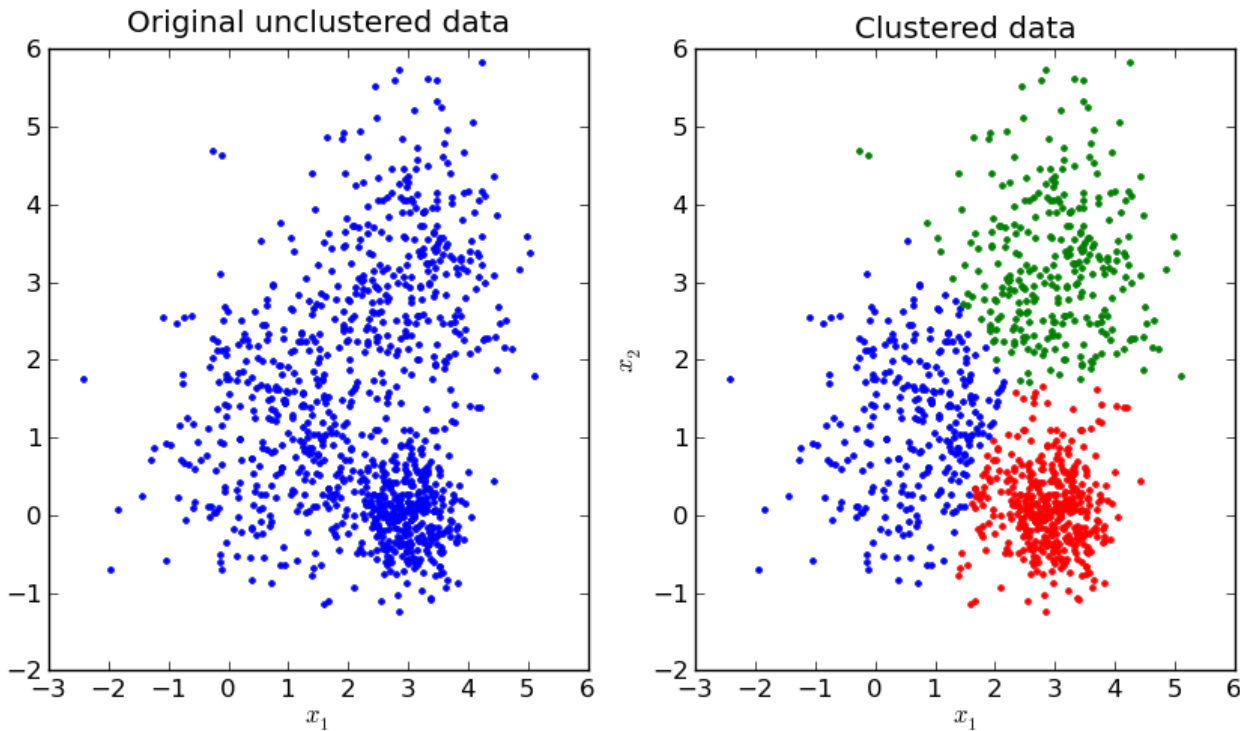
Árboles de decisión



El objetivo es maximizar la pureza o homogeneidad de las clases en cada hoja, mientras que en un árbol de decisión para regresión, el objetivo es minimizar la varianza o el error en cada hoja.

Es muy fácil interpretarlos.

K-means



Algoritmo de aprendizaje autosupervisado para agrupar datos en K grupos distintos. El objetivo del algoritmo es encontrar los centroides de los grupos, que son los puntos centrales alrededor de los cuales se agrupan los datos.

Se basa en la minimización de la suma de las distancias al cuadrado entre cada punto de datos y el centroide de su grupo asignado.

Evaluación de modelos

A solid orange horizontal bar is positioned below the title, spanning most of the width of the white text box.

Evaluación de modelos

Es un proceso **crucial** para determinar la eficacia y precisión de los modelos.

Es importante tener en cuenta que **ninguna técnica de evaluación es perfecta**, y la elección de la técnica de evaluación dependerá del tipo de problema y del conjunto de datos en cuestión.

Evaluación de modelos de clasificación

- Accuracy -> cuidado con esta, puede ser engañosa
- Precision
- Recall
- F1
- Curvas ROC y su AUC
- Confusion Matrix

Evaluación de modelos de regresión

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Otras similares a las dos anteriores

Conceptos Relevantes

A horizontal bar chart with a single salmon-colored bar. The bar is positioned below the title and extends across a significant portion of the slide's width. The background of the slide features a teal-to-brown gradient with wavy, organic patterns.



Overfitting

Sobreajuste es un problema común en el aprendizaje de máquinas, en el que el modelo se ajusta demasiado bien a los datos de entrenamiento y pierde su capacidad de generalización en datos nuevos o desconocidos. En otras palabras, el modelo "memoriza" los datos de entrenamiento en lugar de aprender patrones generales que se puedan aplicar a nuevos datos.



Overfitting

Cuando un modelo tiene un overfitting, suele tener una precisión alta en el conjunto de entrenamiento, pero una precisión baja en el conjunto de prueba o en datos nuevos.



Underfitting

Subajuste es un problema común en el aprendizaje de máquinas, en el que el modelo es demasiado simple para capturar los patrones en los datos de entrenamiento y no puede generalizar correctamente a nuevos datos. En otras palabras, el modelo "no aprende lo suficiente" de los datos de entrenamiento.



Underfitting

Cuando un modelo tiene un underfitting, suele tener una precisión baja tanto en el conjunto de entrenamiento como en el conjunto de prueba o en datos nuevos.

La maldición de la dimensionalidad

Se refiere a un problema común en el aprendizaje de máquinas y en la estadística en general, en el que la cantidad de características o variables en un conjunto de datos aumenta significativamente, lo que dificulta la identificación de patrones significativos en los datos.

Más datos no siempre garantizan que el modelo mejore su eficiencia

El teorema de No Free Lunch

Establece que, en promedio, cualquier algoritmo de aprendizaje de máquinas funcionará igual de bien para cualquier problema en particular. En otras palabras, no hay un algoritmo de aprendizaje de máquinas que sea mejor que cualquier otro para todos los tipos de problemas.

No existe un modelo que le gane a todos los otros modelos para cualquier problema de ML

