Aprendizaje automático

Clase 1

Martin Pustilnik, Iris Sattolo,
Maximiliano Beckel

UNAHUR

Agenda

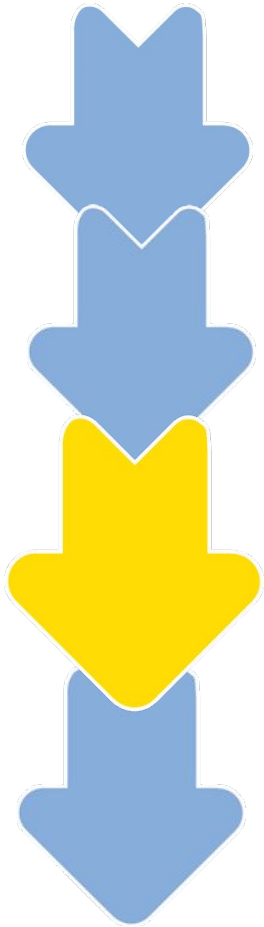
00 Contexto

01 Introducción

02 Aprendizaje automático

UNAHUR

00-Modelado y descubrimiento del conocimiento en Python



01 **Python**

02 **Procesamiento y exploración de datos**

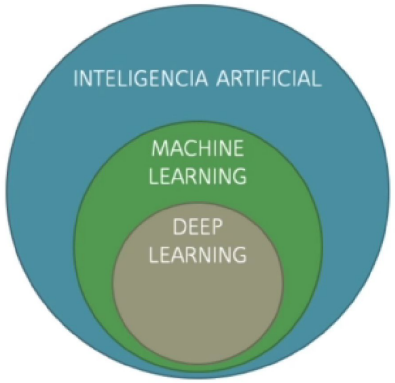
03 **Aprendizaje automático**

04 **Modelado y aprendizaje no supervisado**

UNAHUR

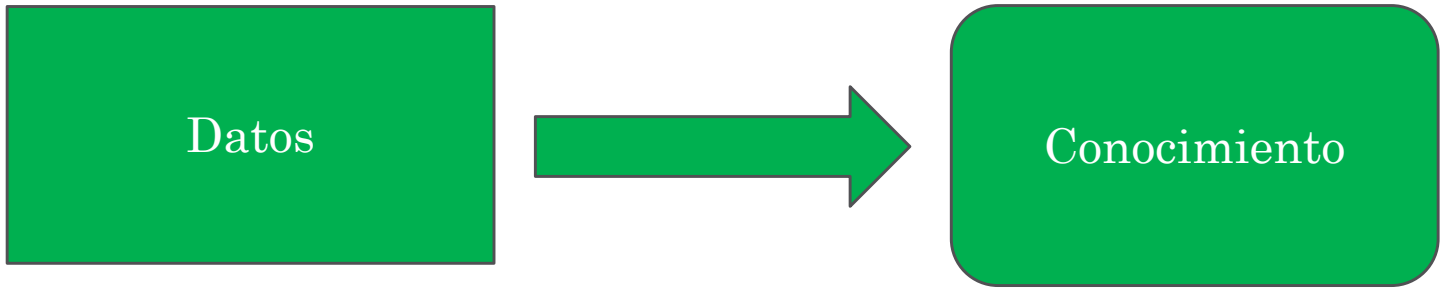
01-Introducción

Una de las cuestiones principales a la hora de abordar el tema de ciencia de datos y aprendizaje automático, es el inmenso número de términos que aparecen relacionados con la disciplina, con el área. Hemos hablado muchas veces de **minería de datos**, otras veces, hablamos de **ciencia de datos**, podemos hablar de **análisis de datos,** inteligencia empresarial, hoy se habla muchísimo de **Big Data** y evidentemente todos está relacionado con el **aprendizaje automático** o incluso hay gente que confunde todos estos términos con lo que puede ser el **aprendizaje profundo** o incluso el área más genérica de lo que es la **inteligencia artificial**.



01-Introducción

Más allá de sus diferencias, todas estas áreas o disciplinas tienen algo en común: su principal objetivo es transformar datos en **conocimiento**. Y lo importante de este paso es que el conocimiento nos va a permitir tomar mejores decisiones!

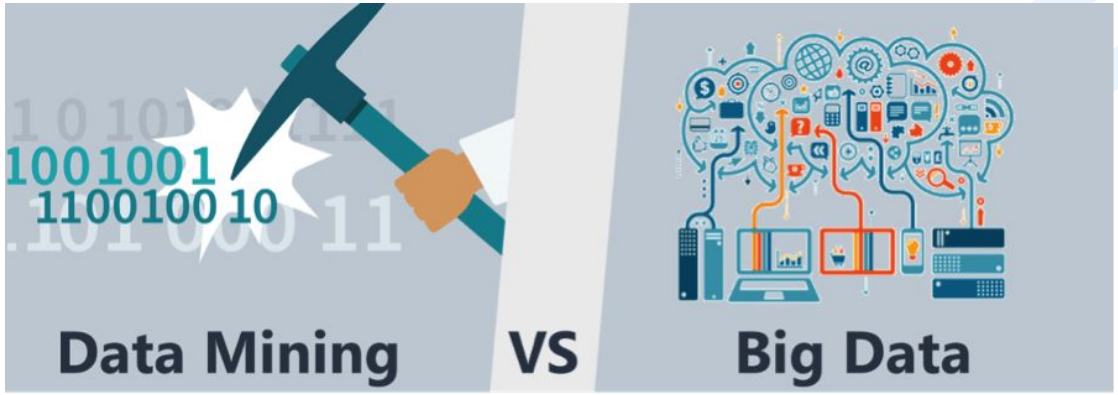


01-Introducción

De qué hablamos cuando hablamos de…

Big Data: está relacionado con la posibilidad actual de poder almacenar y procesar grandes volúmenes de datos. Por lo que cuando hablamos de Big Data hacemos referencia a aquella tecnología que nos permite capturar, gestionar y procesar de forma eficiente montañas de datos.

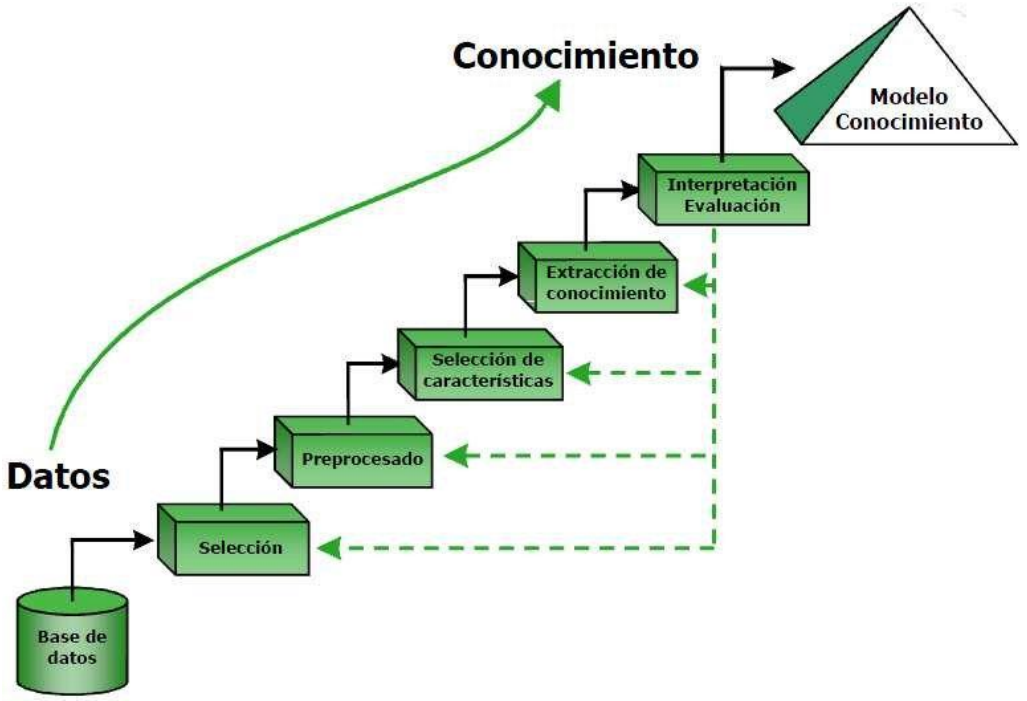
Data Mining: en ese océano de datos es imprescindible poder diferenciar "la paja del trigo". ¿Qué es lo importante y qué no? La minería de datos es el proceso mediante el cual vamos a poder identificar patrones en nuestros datos que nos van a permitir estructurarlo de una manera que sea más comprensible para los futuros análisis que hagamos



01-Introducción

De qué hablamos cuando hablamos de…

Data Mining: en ese océano de datos es imprescindible poder diferenciar “la paja del trigo”. ¿Qué es lo importante y qué no? La minería de datos es el proceso mediante el cual vamos a poder identificar patrones en nuestros datos que nos van a permitir estructurarlo de una manera que sea más comprensible para los futuros análisis que hagamos



01-Introducción

De qué hablamos cuando hablamos de…

Data Science: es un campo interdisciplinario en el que se combinan conocimientos como estadística, ciencias de la computación, data mining, machine learning, además de conocimientos específicos sobre los problemas que quieren abordarse: economía, ciencias, marketing, finanzas, medicina, etc.

01-Introducción

De qué hablamos cuando hablamos de… ¿Y el Aprendizaje Automático?

Samuel (1959)
Campo de estudio que le da a las computadoras la habilidad de aprender sin ser programadas de manera explícita.

Mitchell (1998)
Un programa de computadora se dice que aprende de una **experiencia** E con respecto a una clase de **tareas** T y una medida de **performance** P, si su performance en las tareas T, medidas por P, mejoran con la experiencia E.

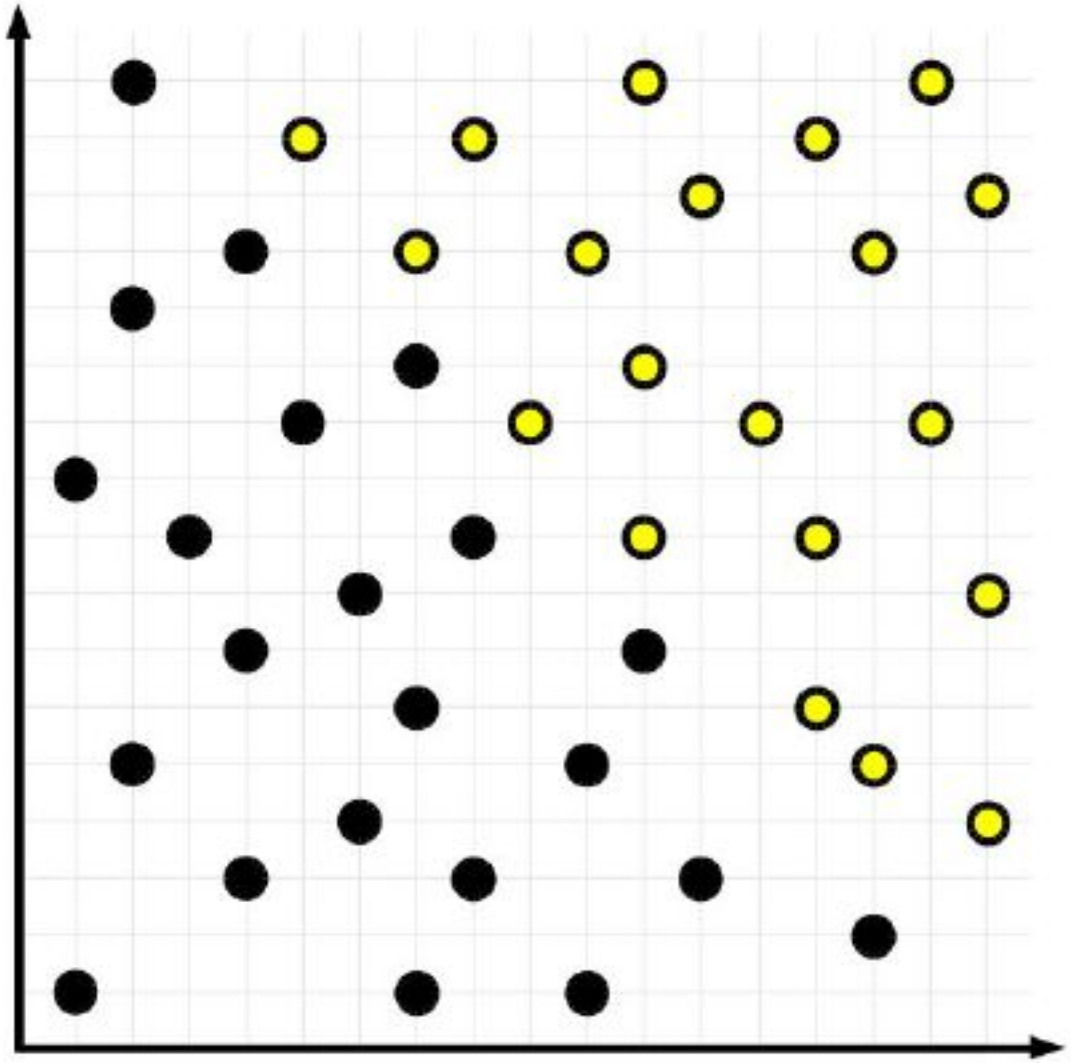
02-Aprendizaje Automático

Aprendizaje Automático

Un programa aprende una tarea si su performance mejora con la experiencia, y el aprendizaje es automático porque no le enseñamos a realizar esa tarea de manera explícita.

Tenemos que definir…

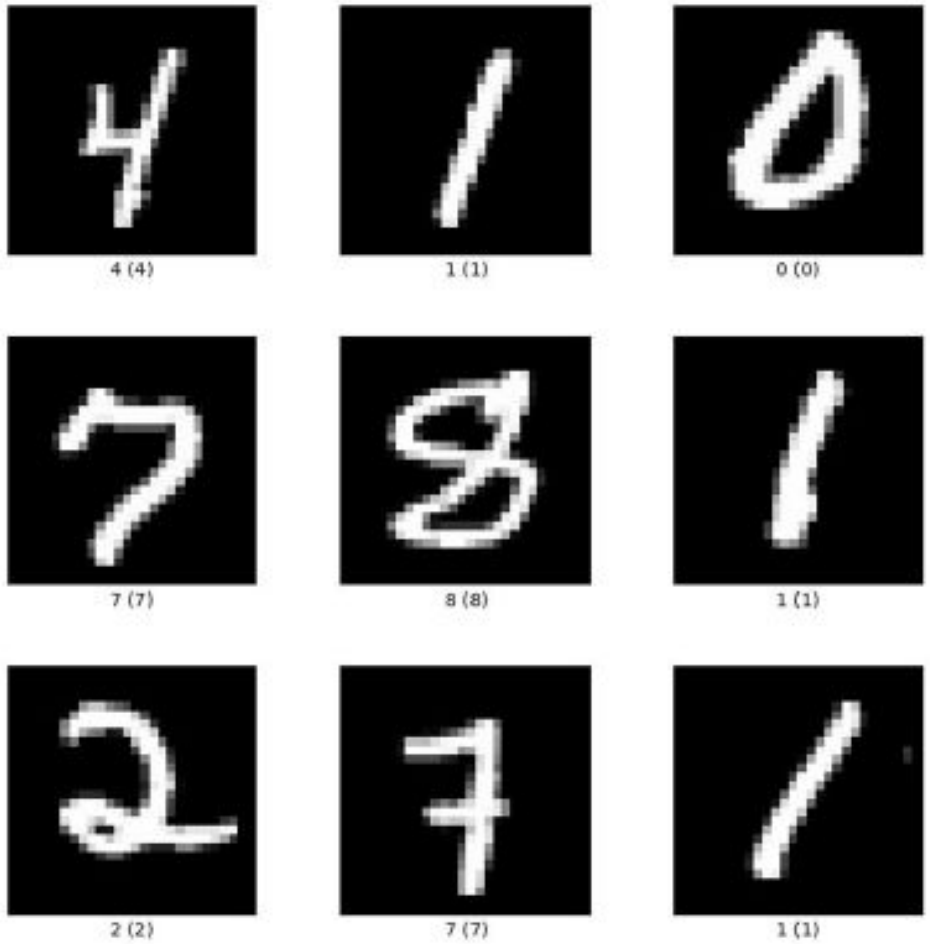
Tarea: objetivo
Experiencia: datos, ejemplos
Medida de performance: distancia formal a mi objetivo



Aprendizaje Automático Ejemplos

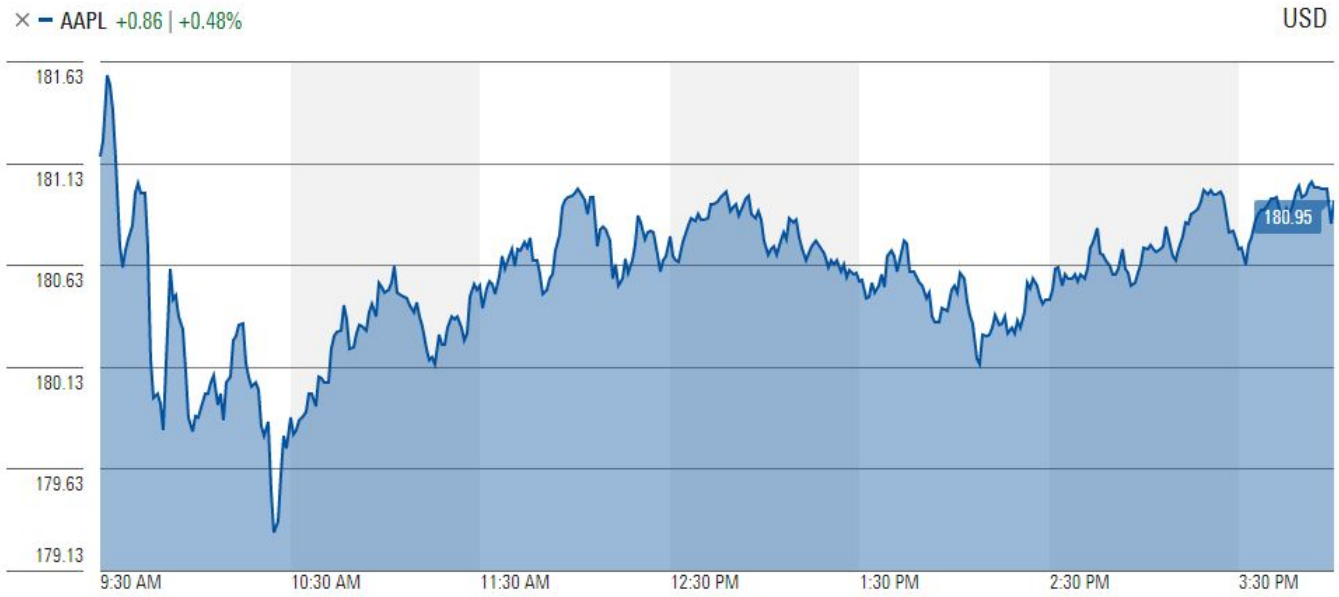
- **Tarea T**: predecir el color de un punto.
- **Experiencia de entrenamiento E**: Base de datos de puntos con sus respectivos colores.
- **Medida de desempeño P**: ¿qué porcentaje de puntos pude predecir correctamente su color?

Aprendizaje Automático Ejemplos



- **Tarea T**: digitalizar números escritos a mano.
- **Experiencia de entrenamiento E**: Imágenes de números a mano con sus etiquetas.
- **Medida de desempeño P**: % de digitalizaciones correctas.

Aprendizaje Automático Ejemplos



- **Tarea T**: Predecir el valor de una acción.
- **Experiencia de entrenamiento E**: Valor de la acción en el pasado.
- **Medida de desempeño P**: % predicciones correctas

Aprendizaje Automático Ejemplo

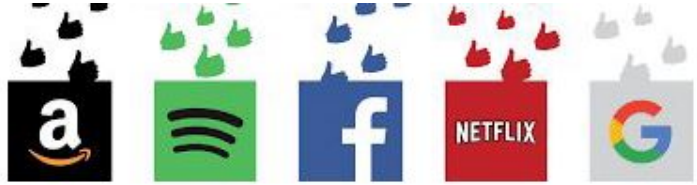
- conducción autónoma

  - **Tarea T**: conducción en carreteras públicas de cuatro carriles utilizando sensores de visión.
  - **Experiencia de entrenamiento E**: una secuencia de imágenes y comandos de dirección registrados mientras se observa a un conductor humano.
  - **Medida de desempeño P**: distancia promedio recorrida antes de un error (según lo juzgado por un supervisor humano)

UNAHUR

02-Aprendizaje automático

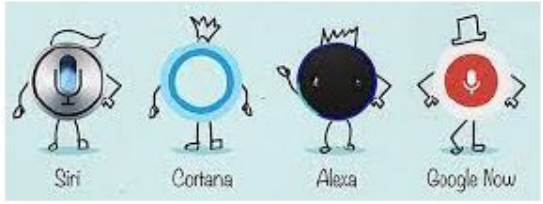
Sistemas de recomendación



Traductores

Predicción de tiempo de viaje, camino óptimo

Reconocimiento del habla, asistentes virtuales



Publicidad online, chatbots, text2image, detectar spam, reconocimiento de rostros/patentes, diagnósticos clínicos, vehículos autónomos, jugar al Go, …

02-Aprendizaje automático

MODELO: representación de una parte de la realidad que nos interesa entender con un determinado objetivo.

Modelar la realidad tiene sus limitaciones…

Sesgo Inductivo: conjunto de suposiciones que uno asume a la hora de construir un modelo a partir de mis datos.

- ➔ Tipo de modelo elegido
- ➔ Sesgo en nuestros datos

02-Aprendizaje automático

Tipos de Modelos según el tipo de Experiencia:

Aprendizaje supervisado:

Los datos están anotados con la respuesta correcta que quiero predecir.

➔ Clasificación: lo que quiero predecir es un clase (variable categórica)

➔ Regresión: lo que quiero predecir es un valor numérico.

Aprendizaje no supervisado:

Los datos de entrenamiento no están anotados.

➔ Encontrar patrones en nuestros datos.

➔ Agrupar nuestros datos en grupos homogéneos (clustering).

Aprendizaje por refuerzos:

Los datos surgen por interacción con el entorno, y el aprendizaje surge gradualmente en base a una recompensa.

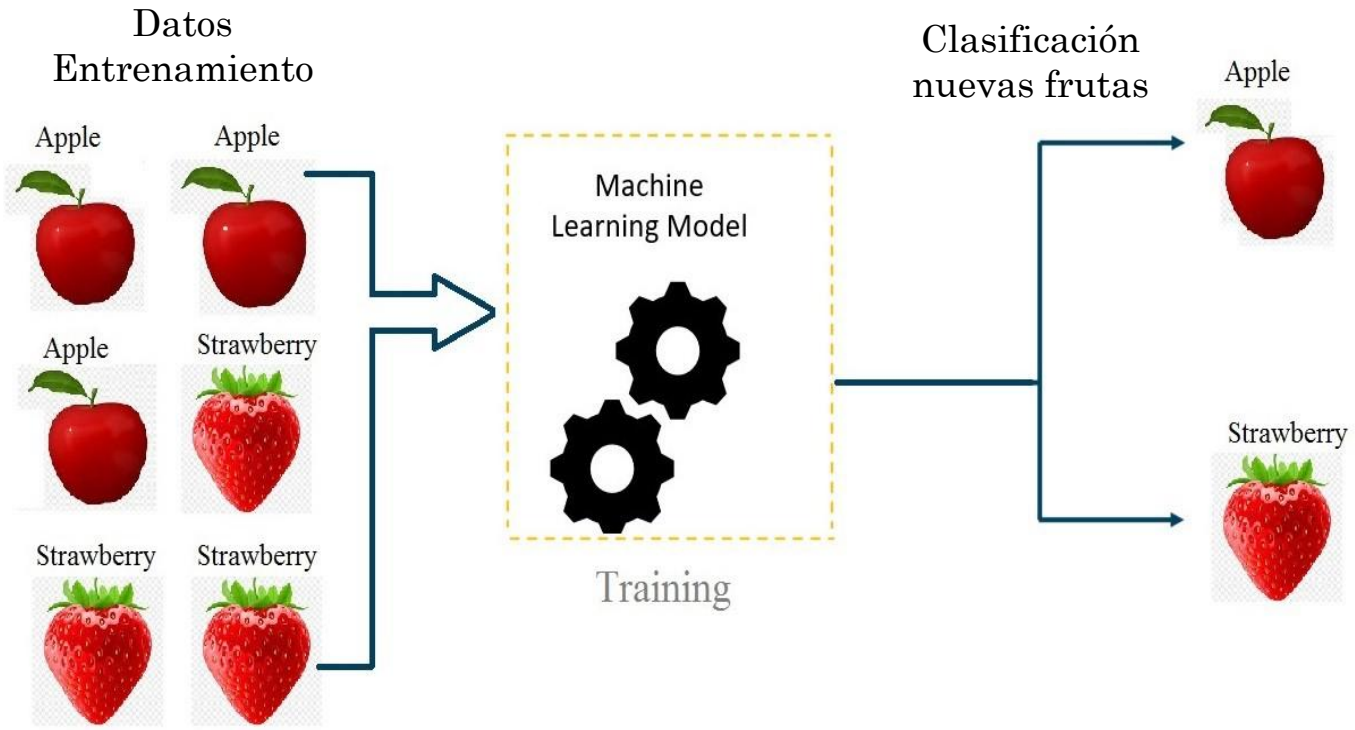
Aprendizaje supervisado

El aprendizaje supervisado permite que los algoritmos 'aprendan' de datos históricos/de entrenamiento y los apliquen a entradas desconocidas para obtener la salida correcta.

Para funcionar, el aprendizaje supervisado utiliza árboles de decisión, bosques aleatorios y Gradient Boosting Machine.

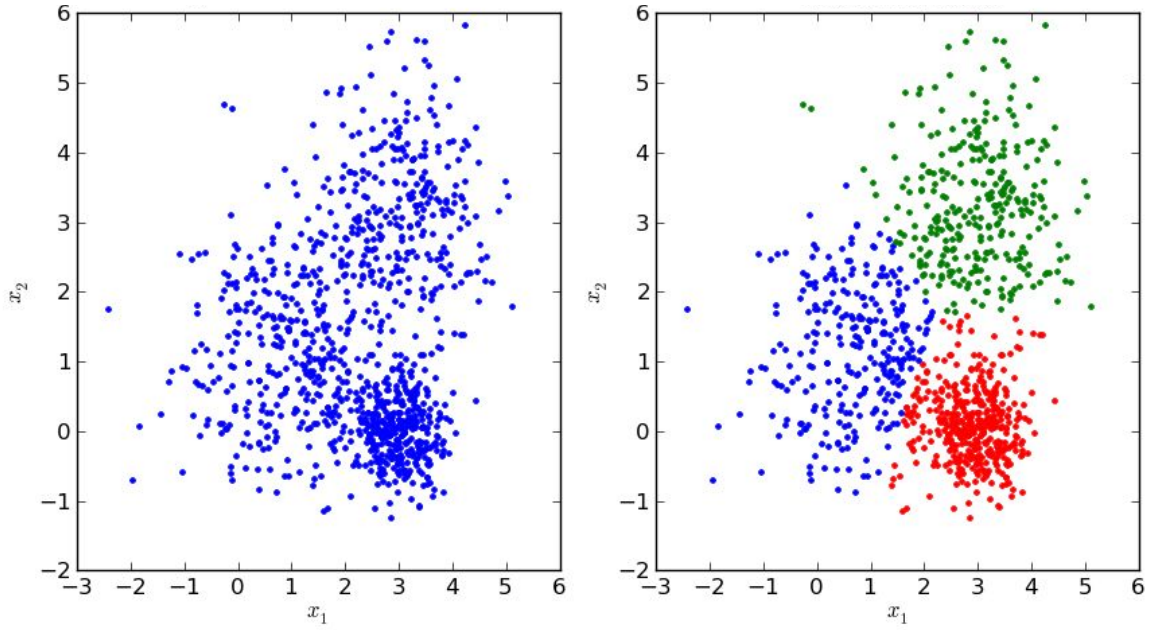
Existen dos tipos principales de aprendizaje supervisado;

clasificación y regresión



Aprendizaje no supervisado

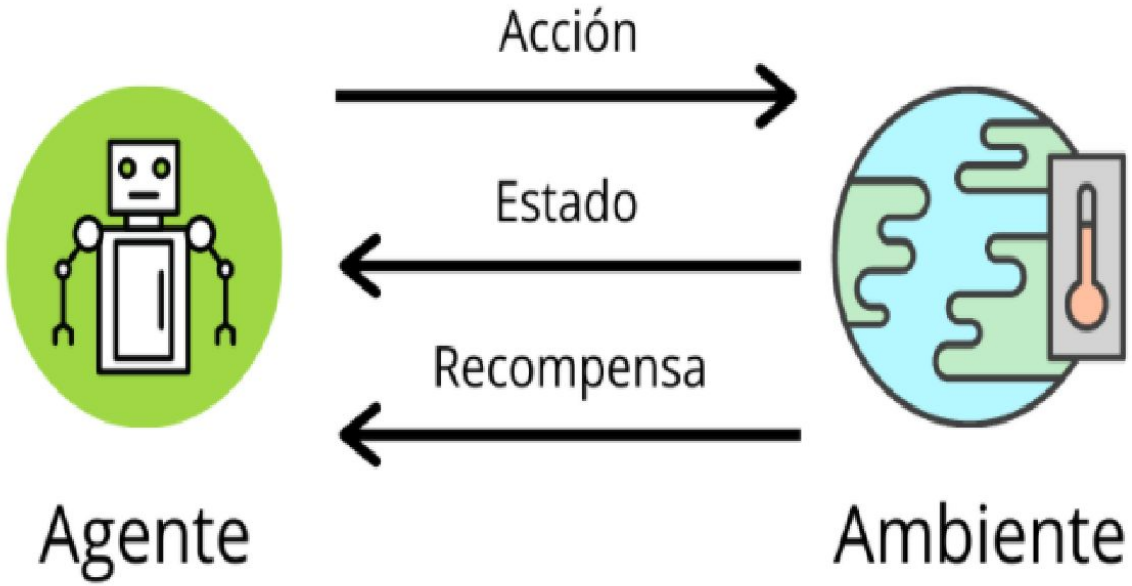
El aprendizaje no supervisado tiene datos sin etiquetar que el algoritmo tiene que intentar entender por sí mismo. El objetivo es simplemente dejar que la máquina aprenda sin ayuda o indicaciones de los científicos de datos, también deberá aprender a ajustar los resultados y agrupaciones cuando haya resultados más adecuados, permitiendo que la máquina comprenda los datos y los procese como mejor le parezca. Se utiliza para explorar datos desconocidos. Puede revelar patrones que podrían haberse pasado por alto o examinar grandes conjuntos de datos que serían demasiado para que los abordara una sola persona



Aprendizaje por refuerzo

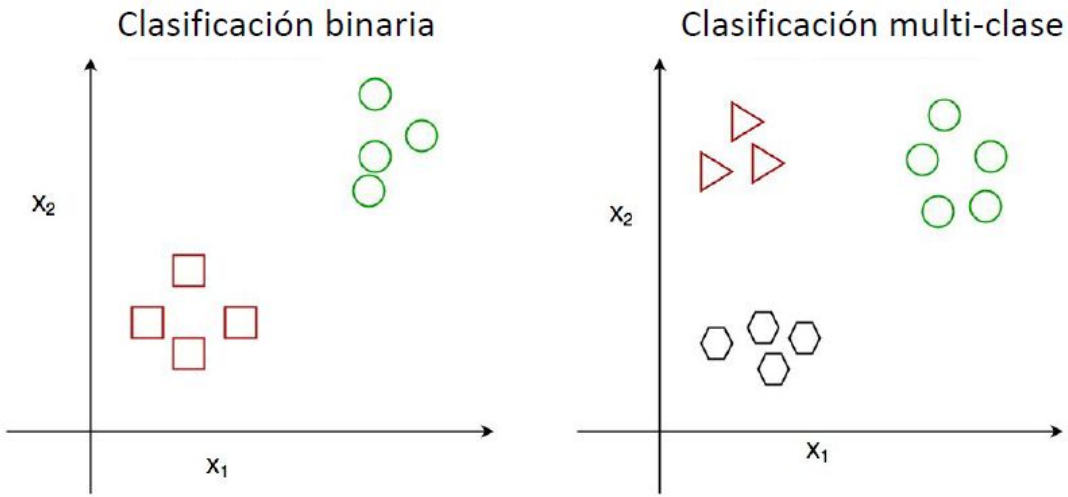
Es capaz de funcionar sin grandes cantidades de datos de entrenamiento. Tan “sólo” necesita una serie de indicaciones para ir aprendiendo a través de prueba y error. Aquí se utilizan recompensas para reforzar el comportamiento deseado.

Es muy útil cuando se conoce cuál puede ser un paso adecuado para lograr un resultado deseado, pero se desconoce el camino completo para lograrlo, lo cual requiere mucha iteración.

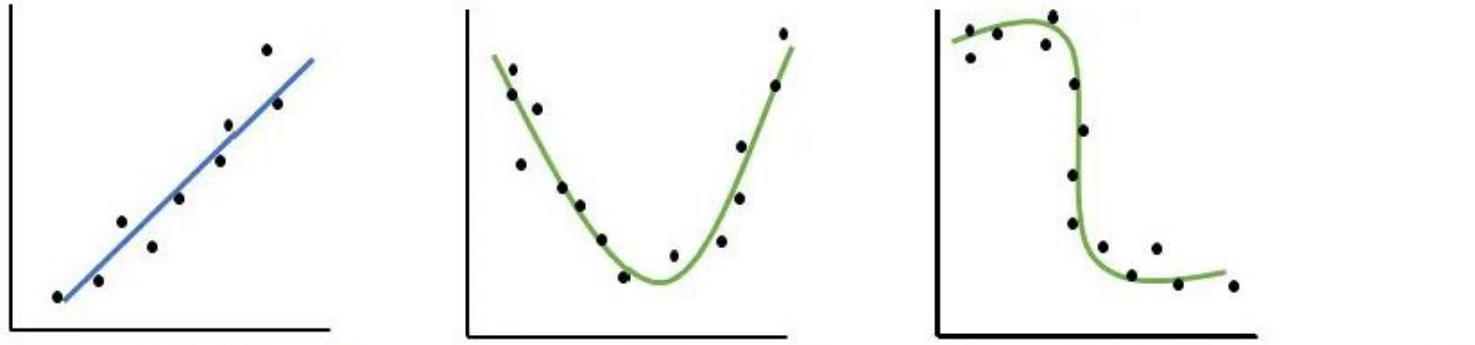


Aprendizaje supervisado

Clasificación:



Regresión:



Regresión lineal

Regresión no lineal (polinomio, etc)

Aprendizaje supervisado: clasificación

El objetivo es predecir las etiquetas de clase categóricas (discreta, valores no ordenados, pertenencia a grupo) de las nuevas instancias, basándonos en observaciones pasadas.

Clasificación Binaria: Es un tipo de clasificación en el que tan solo se pueden asignar dos clases diferentes (0 o 1). El ejemplo típico es la detección de email spam, en la que cada email es: spam → en cuyo caso será etiquetado con un 1 ; o no lo es → etiquetado con un 0.

Clasificación Multi-clase: Se pueden asignar múltiples categorías a las observaciones. Como el reconocimiento de caracteres de escritura manual de números (en el que las clases van de 0 a 9).

Aprendizaje supervisado: regresión

Al usar regresión, el resultado es un número. Es decir, el resultado de la técnica de machine learning que estemos usando será un valor numérico, dentro de un conjunto infinito de posibles resultados.

Ejemplos:

- Estimar cuánto tiempo va a tardar un vehículo en llegar a su destino
- Estimar cuántos productos se van a vender