ntic master
revolucionamos la comunicación

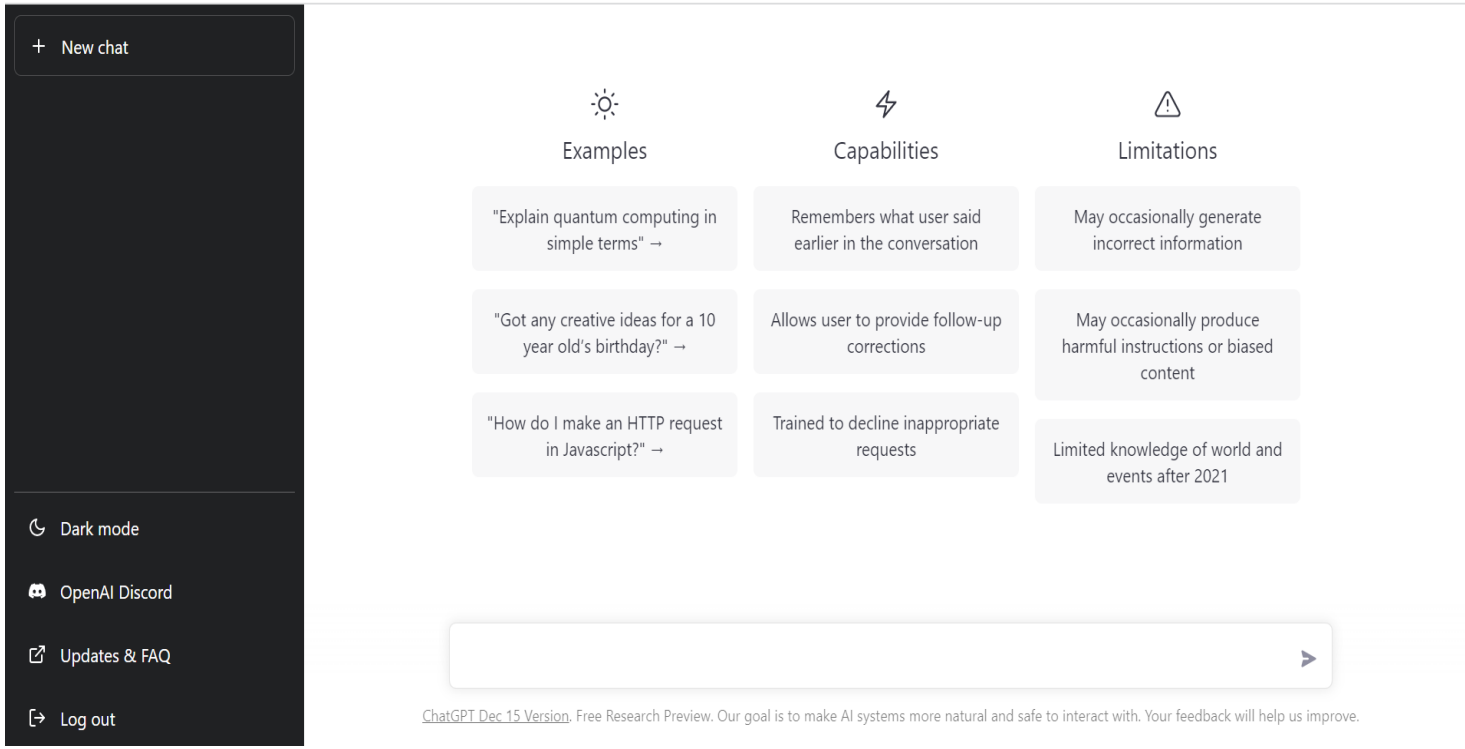
UNIVERSIDAD COMPLUTENSE DE MADRID

Machine Learning
Introducción

Inmaculada Gutiérrez García-Pardo

ChatGPT: LLM al alcance de todos

https://chat.openai.com ps://chat.openai.com



Algunas herramientas asociadas a GPT:

https://github.com/JamesHWade/gpttools

Se trata de una extensión de GPT a Rstudio para convertir código en funciones, documentar paquetes y comentar código. ¡Es capaz de crear código a partir de un comentario en el que se le explique lo que hace!

https://www.chatbcg.com/

Se trata de una extensión de GPT para generar diapositivas y presentaciones

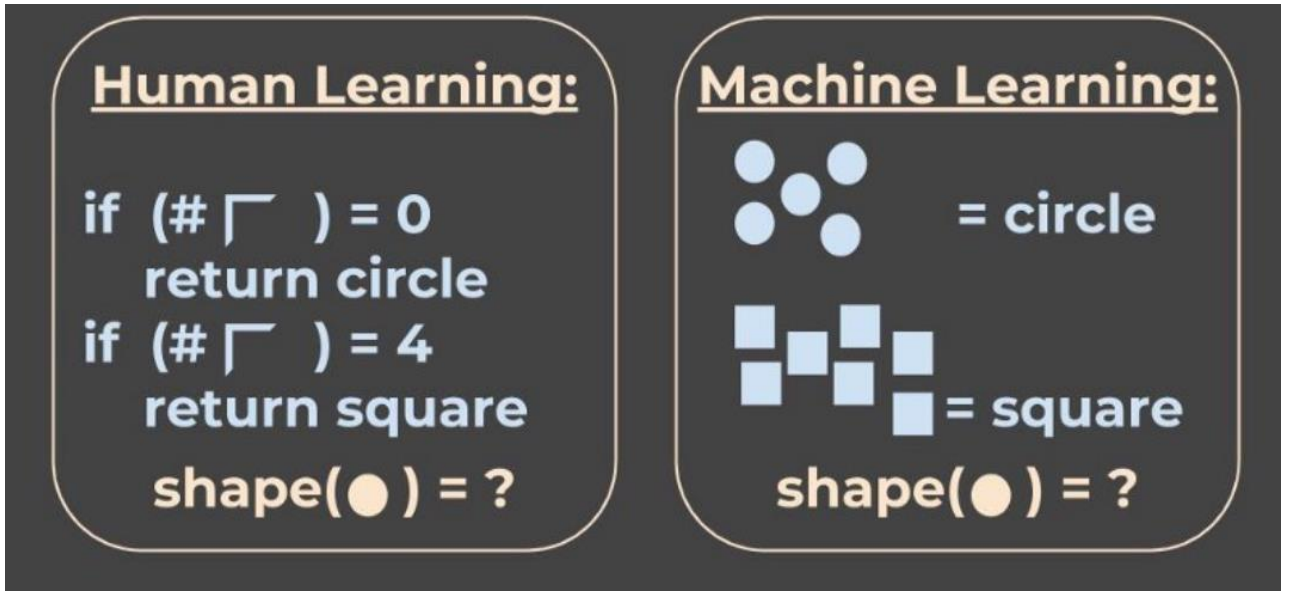
¿Qué hay detrás de esta “magia”?

MACHINE LEARNING

En este módulo se presentarán los principios del machine learning y se explicará la aplicación de algunas técnicas: redes neuronales, árboles, svm, y distintas metodologídas de ensamblado

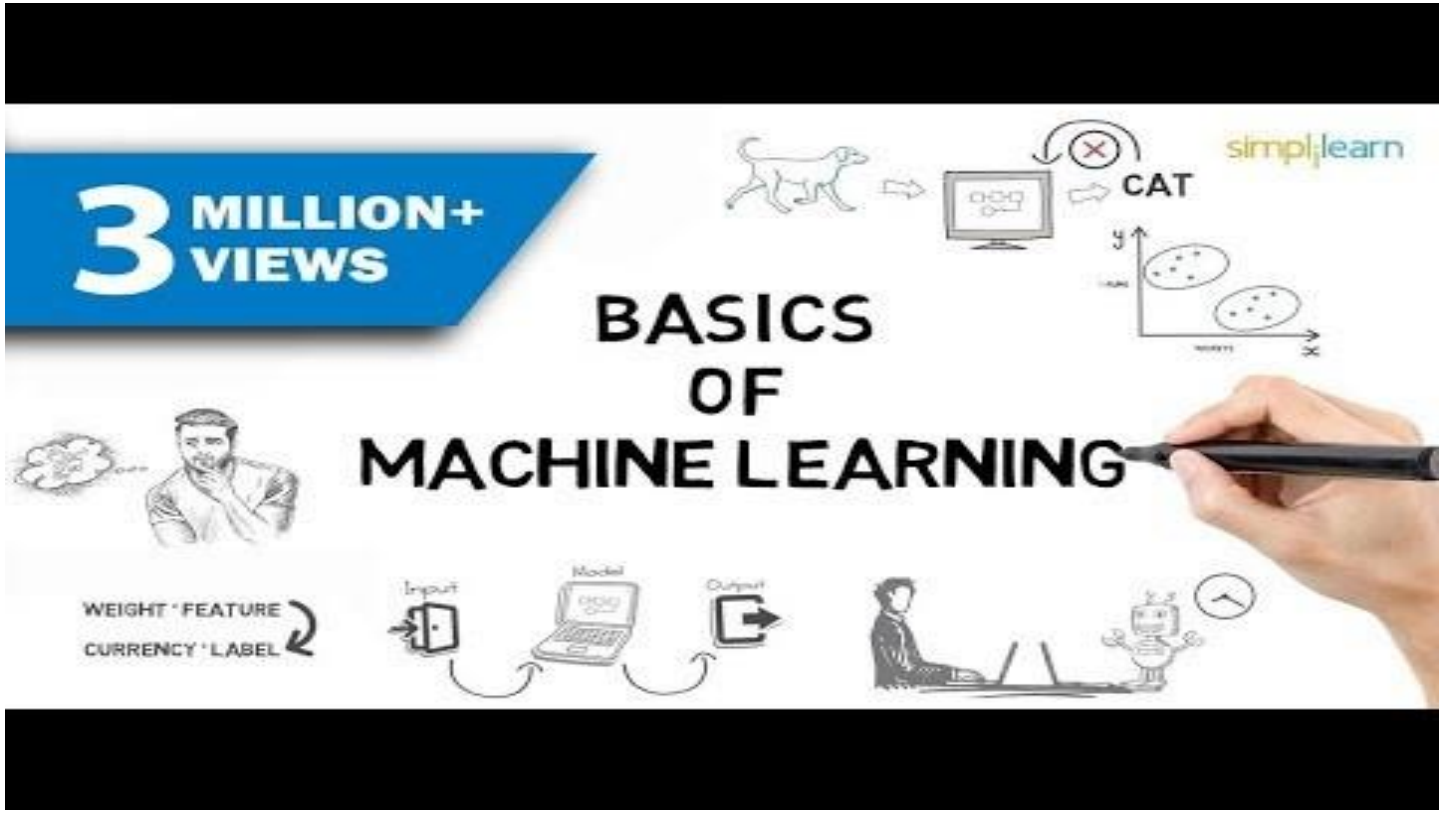
Introducción

El machine learning aprendizaje automático es un campo de la inteligencia artificial que se centra en el **desarrollo de algoritmos** y **modelos** estadísticos que permiten a las máquinas aprender y mejorar su rendimiento en una tarea específica **sin ser programadas explícitamente** para ello. Se basa en patrones e inferencias y no en programación directa



Machine Learning puede entenderse como el proceso de **aprender a través de ejemplos**

Introducción



https://www.youtube.com/watch?v=ukzFI9rgwfU

Introducción

- **¿Qué es el Machine Learning?** El Machine Learning o aprendizaje automático es una rama de la Inteligencia Artificial que permite que las máquinas “aprendan” sin haber sido expresamente programadas para ello.
- **¿Qué significa que la máquina “aprenda”?** Que su desempeño mejora con la experiencia y mediante el uso de datos, pese a que esta habilidad no estaba en su caracterización original.
- **¿Qué se consigue con técnicas de Machine Learning?** Dotar a las máquinas de la capacidad de identificar patrones en datos masivos y elaborar predicciones
- **¿De qué son capaces las técnicas de Machine Learning?** De convertir una muestra de datos en un programa informático capaz de extraer inferencias de nuevos conjuntos de datos para los que no han sido entrenadas previamente.
- **En la práctica, ¿en qué situaciones se aplica Machine Learning?** Recomendaciones de Netflix y Spotify, asistentes inteligentes como Siri o Alexa, respuestas de Gmail, detección de fraude, robótica, diagnóstico médico, mejora en motores de búsqueda, etc.

MACHINE LEARNING *VS* MINERÍA DE DATOS

- Con frecuencia se usan estos conceptos **indistintamente.**
- Sin embargo, hay una diferencia importante relativa al **objetivo** de cada disciplina (aunque es una apreciación algo personal):

  La **minería de datos** descubre patrones anteriormente **desconocidos**, el **Machine Learning** se usa **para reproducir patrones conocidos** y hacer **predicciones** basadas en los patrones.

Se podría decir que la **minería de datos** tiene una función de **exploración** y **análisis** de grandes conjuntos de datos para descubrir patrones y tendencias mientras que el **Machine Learning** se centra en la **predicción** o **creación** de **algoritmos** y **modelos** que permiten a las computadoras **aprender** de los datos para hacer predicciones o tomar decisiones en base a nuevos datos.

MACHINE LEARNING *VS* MINERÍA DE DATOS

- **Definición:**
  - **Machine Learning (ML):** Es un campo de la inteligencia artificial que se centra en el desarrollo de **algoritmos y modelos** que permiten a las máquinas **aprender patrones** y realizar tareas específicas **sin programación explícita**.
  - **Minería de Datos (Data Mining):** Es el proceso de **descubrir patrones significativos**, información y conocimientos útiles a partir de grandes conjuntos de datos. Involucra la aplicación de técnicas estadísticas y matemáticas para explorar y analizar datos.
- **Objetivo:**
  - **ML:** Se enfoca en la construcción de **sistemas** que pueden **aprender** de los datos para realizar tareas específicas sin ser programados explícitamente.
  - **Minería de Datos:** Busca **descubrir patrones** y conocimientos previamente desconocidos en grandes conjuntos de datos.
- **Enfoque:**
  - **ML:** Se centra en la **creación de modelos predictivos y descriptivos** utilizando algoritmos que pueden mejorar su rendimiento con el tiempo a medida que se les proporciona más datos.
  - **Minería de Datos:** Utiliza técnicas como la clasificación, la regresión, el agrupamiento y la asociación para analizar datos históricos y descubrir patrones y tendencias.
- **Aplicación:**
  - **ML:** Se aplica en una amplia variedad de campos, como reconocimiento de voz, visión por computadora, recomendaciones personalizadas, juegos, entre otros.
  - **Minería de Datos:** Se utiliza comúnmente en la exploración de grandes conjuntos de datos empresariales para descubrir información relevante para la toma de decisiones.
- **Énfasis en el Conocimiento:**
  - **ML:** Se centra en la **predicción** y el **aprendizaje automatizado** a partir de datos.
  - **Minería de Datos:** Se centra en la identificación de **patrones** y **conocimientos** útiles a partir de **datos existentes**.

Grupos de algoritmos de Machine Learning

❑DOS GRUPOS PRINCIPALES

- **Aprendizaje Supervisado:** algoritmos que enseñan a la máquina mediante el ejemplo. El programador proporciona al algoritmo un conjunto de datos conocidos que incluye los datos de **entrada** y **salida;** el algoritmo debe encontrar un método específico para llegar a esas salidas a partir de las entradas dadas. El objetivo es crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente. Las utilidades más comunes de las técnicas de aprendizaje supervisado son de clasificación y regresión.
- **Aprendizaje No Supervisado**: el algoritmo estudia los datos para identificar patrones en los mismos, **sin instrucciones ni conocimiento a priori**. La máquina determina las correlaciones y relaciones analizando los datos disponibles. El aprendizaje no supervisado usa datos históricos que no están etiquetados, con el fin de explorarlos para encontrar alguna estructura o forma de organizarlos. Muy frecuente en campañas de marketing altamente segmentadas que buscan agrupar clientes con características o comportamientos similares.

Grupos de algoritmos de Machine Learning

❑PERO TAMBIÉN SE PUEDE DISTINGUIR

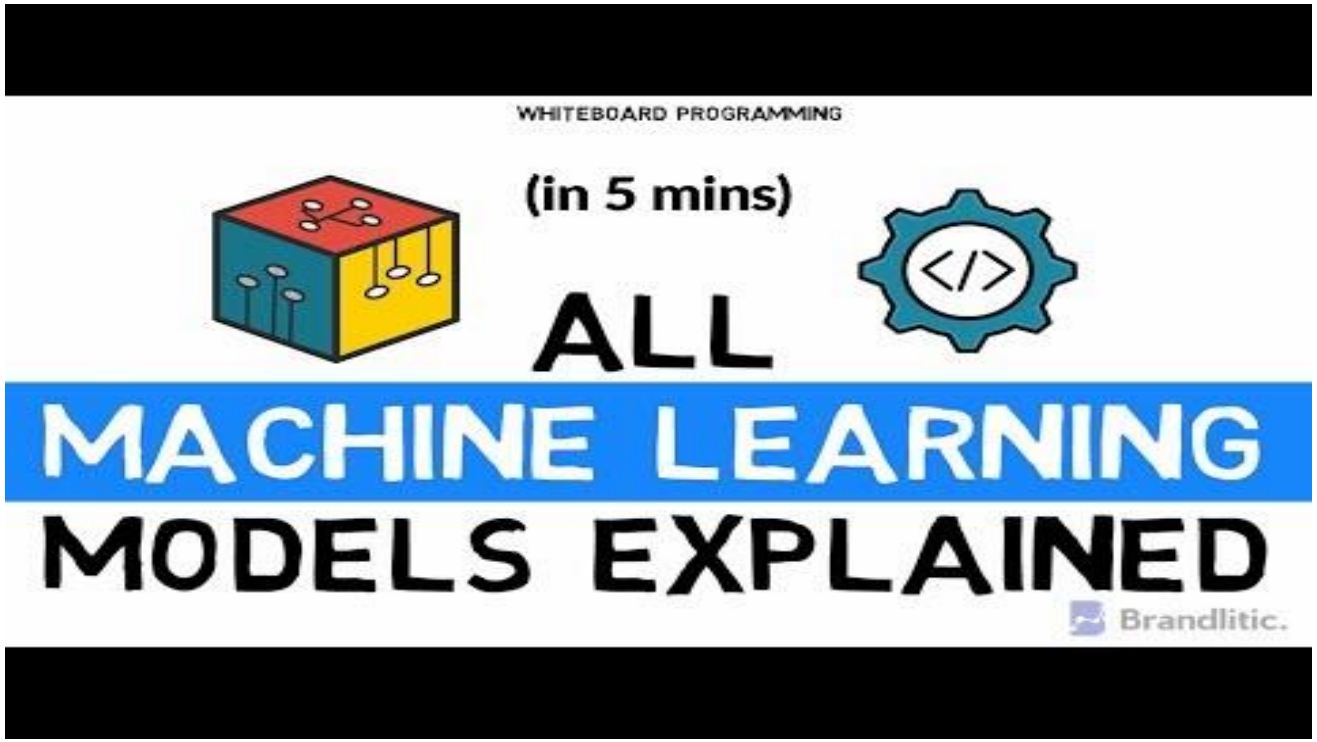
- **Aprendizaje Semisupervisado:** estos algoritmos combinan las características de los grupos anteriores. Se entrena un modelo matemático a partir de un conjunto de datos que contiene tanto datos etiquetados como no etiquetados. Utilizan la información de los datos no etiquetados para mejorar el rendimiento del modelo en la tarea de clasificación o predicción.
- **Aprendizaje por Refuerzo:** algoritmos de aprendizaje reglamentados, a los que se proporciona un conjunto de acciones, parámetros y valores finales. Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo. Este sistema enseña a la máquina a través del proceso de ensayo y error. Aprende de experiencias pasadas y comienza a adaptar su enfoque en respuesta a la situación para lograr el mejor resultado posible. modelo de aprendizaje en el que el sistema aprende a través de la interacción con su entorno. El sistema recibe retroalimentación en forma de recompensas o castigos en función de sus acciones, y su objetivo es maximizar la recompensa a largo plazo. El aprendizaje por refuerzo se utiliza comúnmente en robótica, juegos y otros escenarios donde el sistema debe aprender a tomar decisiones óptimas en un entorno cambiante

Técnicas de clasificación/regresión

- **Reglas de asociación:** detectan relaciones entre variables. Algunos de los algoritmos son el Eclat y el de Patrón Frecuente.
- **Algoritmos genéticos:** procesos de búsqueda heurística que imitan la evolución biológica como estrategia para resolver problemas de optimización de búsqueda global, explorando todo el espacio de soluciones del problema permitiendo salir de posibles óptimos locales e ir en busca de óptimos globales.
- **Árboles de decisión:** son algoritmos con una estructura de árbol similar a un diagrama de flujo que utilizan un método de bifurcación para ilustrar cada resultado posible de una decisión. Cada nodo dentro del árbol representa una prueba en una variable específica, y cada rama es el resultado de esa prueba.
- **Redes neuronales y aprendizaje profundo (*deep learning*): comprenden muchos** elementos de procesamiento interconectados, que trabajan al unísono para resolver problemas específicos. Aprenden con el ejemplo y la experiencia, y son muy útiles para modelar relaciones no lineales en datos de alta dimensión, o donde la relación entre las variables de entrada es difícil de entender.
- **Máquinas de vector soporte:** métodos utilizados para clasificación y regresión, usando un conjunto de ejemplos de entrenamiento clasificado en dos categorías para construir un modelo que prediga si un nuevo ejemplo pertenece a una u otra de dichas categorías.
- **Algoritmos de agrupamiento:** permiten la clasificación de observaciones en subgrupos, de modo que las observaciones en cada grupo se asemejen entre sí según ciertos criterios.
- **Redes bayesianas:** modelos probabilísticos que representan una serie de variables de azar y sus independencias condicionales a través de un grafo acíclico dirigido. Se usan para modelar, por ejemplo, las relaciones probabilísticas entre enfermedades y síntomas.

Modelos de Machine Learning

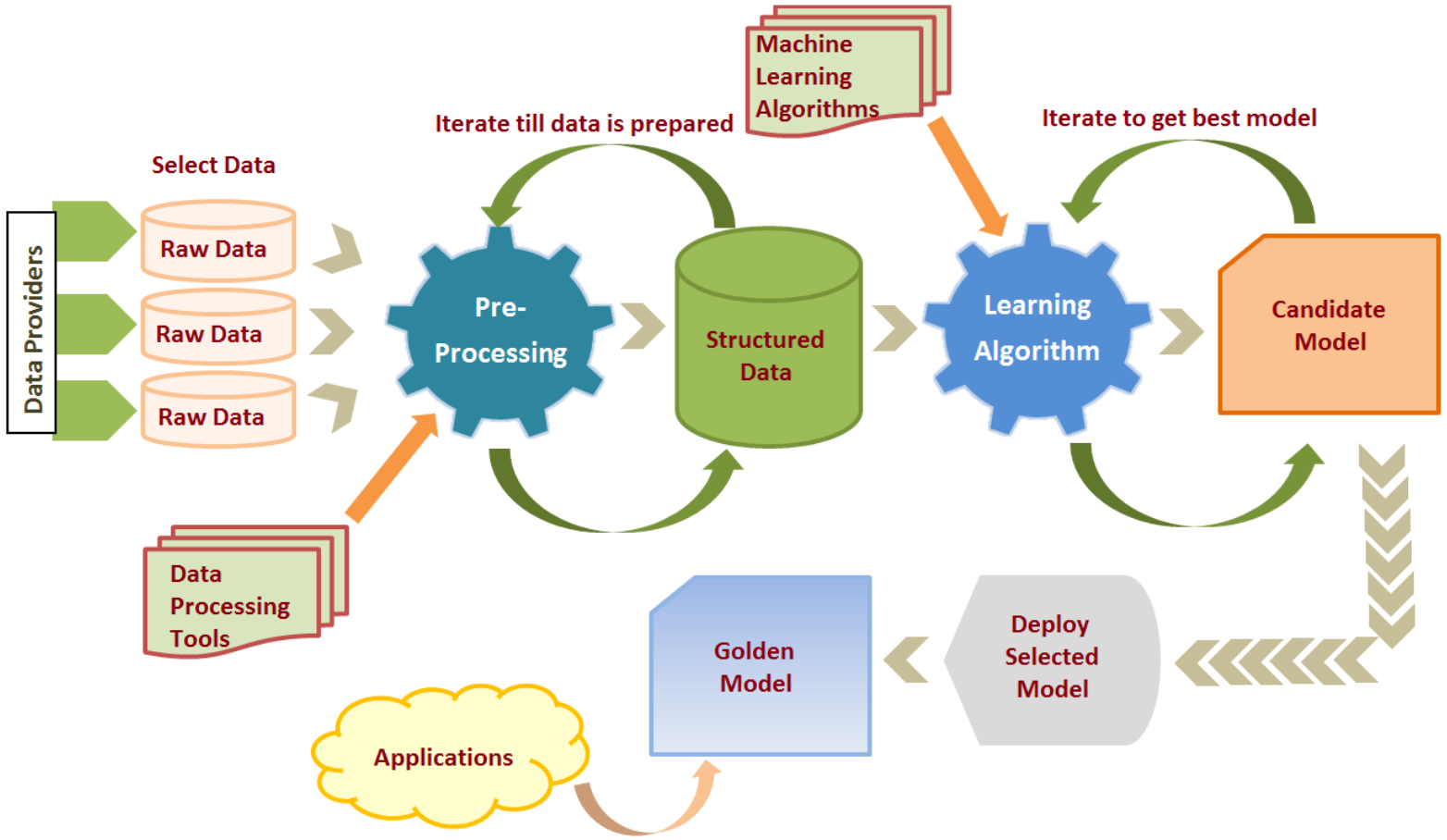
https://www.youtube.com/watch?v=yN7ypxC7838&t=207s



Fases

- **Definición del objetivo,** planteamiento de un problema que requiera una solución a medio-largo plazo.
- **Recopilación y preparación de datos,** procesamiento y limpieza de los mismos.
- **Elección del modelo:** ¿Clasificación binaria? ¿Clasificación multiclase? ¿Regresión?
- **Entrenamiento del modelo,** suministrando al algoritmo la información que necesita para el aprendizaje inicial.
- **Evaluación del modelo,** diferenciando entre los datos de prueba y los de entrenamiento.
- **Análisis de errores,** permite modelar y cambiar los aspectos no relevantes para mejorar el rendimiento.

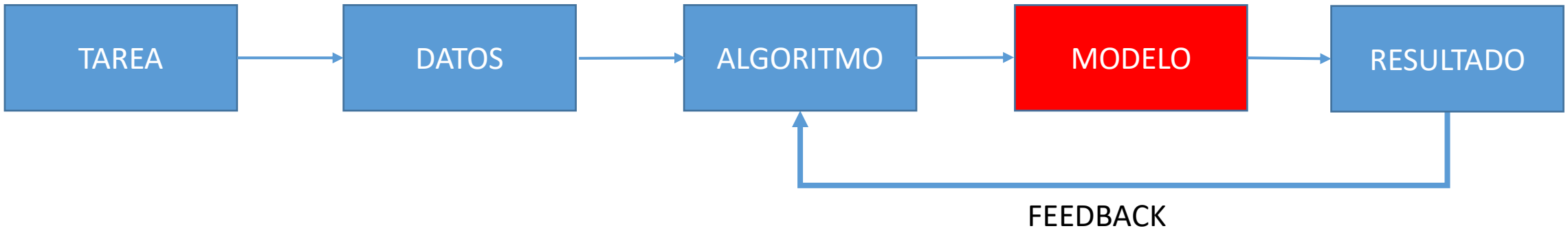
Machine Learning: es un proceso iterativo



Programación tradicional VS Machine Learning

PROGRAMACIÓN TRADICIONAL

Machine Learning



CLASIFICACIÓN *VS* REGRESIÓN

- *Clasificación* y *regresión* son nociones relacionas con técnicas de Machine Learning supervisado. La diferencia es que un sistema de **clasificación** predice una **categoría**, mientras que una **regresión** predice un **número**.
- Ejemplo de clasificación: predicción de **correos spam**. Los correos históricos se categorizan como “spam” o como “legítimos”, y a partir de ellos se trata de clasificar un correo nuevo.
- Ejemplo de regresión: predecir el **número de reservas** que se harán en 2023 en un hotel, conocido el histórico de los últimos 5 años.

Elementos básicos de Machine Learning

- **Dataset** o **conjunto de datos**: histórico de datos usado para entrenar al sistema seleccionado para detectar patrones. Un conjunto de datos se compone de **instancias** (*filas* u *observaciones*), que a su vez constan de **factores** *(columnas* o *variables),* características o propiedades.
- **Instancia**: cada uno de los datos disponibles para el análisis es una instancia, compuesta a su vez de las características que la definen.
- **Característica**: son los atributos que describen cada una de las instancias del conjunto de datos.
- **Objetivo**: atributo que se quiere predecir.
- **Confianza**: probabilidad de acierto que calcula el sistema para cada predicción hecha.
- **Aprendizaje** o **entrenamiento** **(learning, training):** proceso en el que se detectan los patrones de un conjunto de datos. Tras la identificación de los patrones, se pueden hacer predicciones incorporando nuevos datos al sistema entrenado.

IMPORTANTE: división de datos

Los métodos de Machine Learning **aprenden** de los datos con los que los **entrenamos**. A partir de ellos, intentan ***inferir el patrón*** que en que se organizan, para a su vez utilizar este patrón para predecir el resultado de nuevos casos.

Para evaluar el funcionamiento de un sistema de Machine Learning es IMPRESCINDIBLE calibrarlo, probando su ejecución en un conjunto de datos diferente al utilizado para entrenarlo.

En todo proceso de Machine Learning, los datos **disponibles** se dividen en dos partes: datos de **entrenamiento** y datos de **test** (que a su vez pueden constar de parte de validación y parte de test).

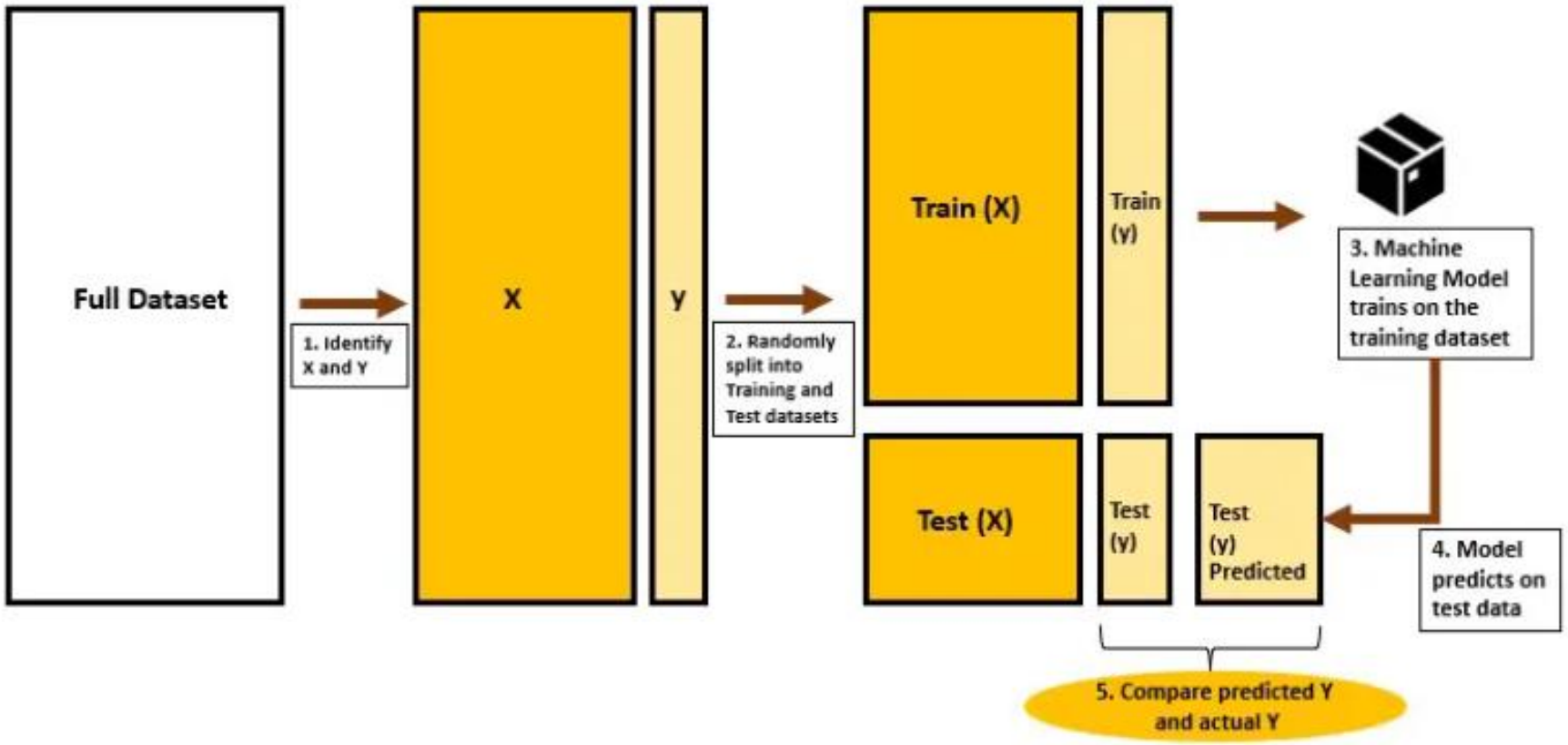
División de datos

- **Datos de entrenamiento** (*training data*)*:* son los datos que usados para **entrenar** un modelo. La **calidad del modelo** está directamente relacionada con la **calidad** de los **datos** de **entrenamiento**. IMPRESCINDIBLE llevar a cabo un proceso de **depuración** y **limpieza** de datos antes de iniciar el ajuste del modelo.

- **Datos de prueba** o **validación** (*testing data*)*:* son los datos reservados para comprobar si el modelo que generado a partir de los datos de entrenamiento "funciona"; esto es, si las respuestas predichas por el modelo para un caso totalmente nuevo son acertadas o no.

- Importante que el *conjunto de* ***datos de prueba*** tenga un ***volumen suficiente*** como para generar resultados estadísticamente significativos, y a la vez, que sea ***representativo*** del conjunto de datos global.

División de datos



Entrenamiento / Validación / Test

- **Datos de entrenamiento:** son los datos que entrenan los modelos, usados para ajustar los parámetros
- **Datos de validación:** se usan para seleccionar el mejor de los modelos entrenados. Proporcionan una evaluación imparcial del ajuste de un modelo en el conjunto de datos de entrenamiento mientras ajusta lo parámetros del modelo. Estos datos se suelen utilizar cuando se aplica regularización mediante *early stopping.*
- **Datos de test**: se usan para calcular el error real cometido con el modelo seleccionado

IMPORTANTE: no mezclar datos!!! Cada instancia sólo puede pertenecer a uno de estos subconjuntos. Imprescindible que haya suficiente cantidad total de datos para poder hacer la división.

Algunas divisiones comunes: (80, 10, 10); (70, 20, 10); (70, 15, 15); (60, 20, 20)

Entrenamiento / Test

- **Datos de entrenamiento:** son los datos que entrenan los modelos, usados para ajustar los parámetros
- **Datos de test**: se usan para calcular el error real cometido con el modelo seleccionado

IMPORTANTE: no mezclar datos!!! Cada instancia sólo puede pertenecer a uno de estos subconjuntos. Imprescindible que haya suficiente cantidad total de datos para poder hacer la división.

Algunas divisiones comunes: (80, 20); (67, 33); (50, 50)

Problemas que surgen

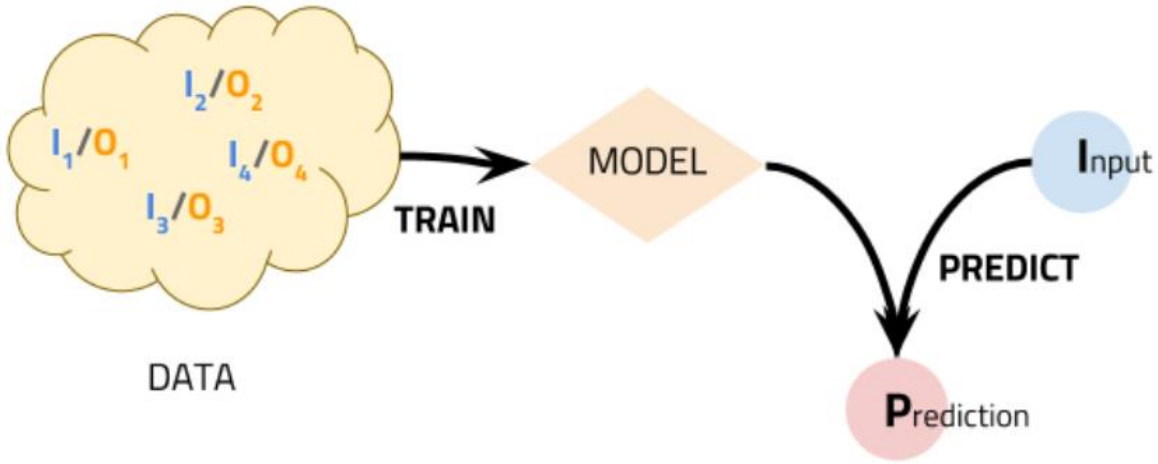
- **Sobreajuste u overfitting:** ocurre cuando un modelo está “sobre-entrenado”. Ocurre en modelos complejos que se ajustan tan exactamente al conjunto de datos de entrenamiento a partir del cual se han creado, que no se generalizan bien a los datos de test, perdiendo gran parte de su capacidad predictiva. Esto se debe a que los datos siempre presentan cierto grado de error, imprecisión y variabilidad, e intentar ajustarse demasiado a ellos, complica el modelo inútilmente y le resta utilidad y capacidad de generalización.
- Se produce cuando un sistema de Machine Learning se **entrena demasiado** o con **datos anómalos**, provocando que el algoritmo *aprenda* patrones que no son generales y características específicas de los datos utilizados. Los modelos **más complejos** tienden a **sobreajustar**. Un mismo modelo puede tender a sobreajustarse a menor cantidad de datos.
- Un método habitual para evaluar el sobreajuste consiste en la comparación del error obtenido en los datos de *train* y el error cometido en los datos de *test.* Lo ideal es que ambos tipos de error se parezcan lo más posible.
- **¿Cómo evitar el sobreajuste?:** incorporando más datos, provocando que el algoritmo generalice mejor; modificando los parámetros de los algoritmos y proponiendo sistemas sencillos, e incorporando parámetros de regularización.

Problemas que surgen

- **Subajuste o underfitting:** es justamente lo contrario al sobreajuste. Sucede cuando el conjunto de datos de entrenamiento es insuficiente o poco representativo, o tiene ruido en alguna de sus dimensiones. Estas características pueden derivar en el entrenamiento de un modelo excesivamente simple y con poco valor predictor.

Para generar un buen modelo de Machine Learning, es importante encontrar el **equilibrio** entre sobreajuste y subajuste: *Bias vs Varianza*.

Escenario típico en ML



Input	Output	Prediction
I1	O1	Ô1
I2	O2	Ô2
I3	O3	Ô3
I4	O4	Ô4
I5		**Ô5**
I6		**Ô6**

- Datos conocidos siempre: **VARIABLES DE ENTRADA**, características que describen la variable de interés a predecir: la SALIDA.
- Datos conocidos **solo** para **algunos casos**: **VARIABLE DE SALIDA**, la variable de interés para la cual se desean predicciones. Su valor es conocido para algunos casos utilizados para entrenar los modelos que permitirán predicciones.
- **PREDICCIÓN**: una vez que el modelo ha sido entrenado utilizando variables explicativas (ENTRADA) para aprender cómo se comporta la variable de interés (SALIDA) en cada caso, el modelo puede utilizarse para hacer predicciones para casos en los que la variable de SALIDA es desconocida.

Ejemplo típico en ML: *breast cancer Wisconsin dataset*

- Conjunto de datos popular sobre la predicción del cáncer de mama.
- Las características se calculan a partir de una **imagen digitalizada** de una aspiración con aguja fina (FNA) de una masa mamaria. **Describen** las **características** de los núcleos celulares presentes en la imagen.
- Características de entrada: 30 atributos numéricos predictivos y la clase.
- 569 instancias.
- **Propósito**: predecir si el tumor es benigno (1) o maligno (0).
- **Hipótesis de predicción**: en relación con su descripción, *¿es el tumor benigno?*

Ejemplo típico en ML: *breast cancer Wisconsin dataset*

Datos disponibles, la naturaleza del tumor se **conoce.** Estas filas se usan para **entrenar** los modelos.

Nuevos datos, no se conoce la naturaleza del tumor. Estas filas se usan para **predecir** nuevos valores de la variable objetivo.

target	radius	texture	permiter	area	smoothness	compactness
0	17,99	10,38	122,8	1001	0,1184	0,2776
0	20,57	17,77	132,9	1326	0,08474	0,07864
0	19,69	21,25	130	1203	0,1096	0,1599
0	19,81	22,15	130	1260	0,09831	0,1027
1	13,54	14,36	87,46	566,3	0,09779	0,08129
1	13,08	15,71	85,63	520	0,1075	0,127
1	9,504	12,44	60,34	273,9	0,1024	0,06492
0	15,34	14,26	102,5	704,4	0,1073	0,2135
0	21,16	23,04	137,2	1404	0,09428	0,1022
1	8,196	16,84	51,71	201,9	0,086	0,05943
	16,02	23,24	102,7	797,8	0,08206	0,06669
	15,78	17,89	103,6	781	0,0971	0,1292
	19,17	24,8	132,4	1123	0,0974	0,2458
	15,85	23,95	103,7	782,7	0,08401	0,1002
	13,73	22,61	93,6	578,3	0,1131	0,2293


ntic
master
revolucionamos la comunicación
UNIVERSIDAD
COMPLUTENSE
DE MADRID