

Fundamentos de la ID y el ML

- Introducción a los tipos de
machine learning

- Todos los modelos de ID, independiente de su complejidad (desde generativa hasta una regresión lineal) de machine learning siempre requiere una base de datos de aprendizaje o entrenamiento.

El proceso general incluye:

- 1) Utilizar una gran cantidad de datos
- 2) Alinear algoritmos

2) Obtener un algoritmo entrenado
3) Utilizar nuevos datos llamados
de inferencia o predicción
que son ya pasados por el
algoritmo entrenado para
una predicción.

En la base de datos, cada fila
se denomina instancia y cada
columna características. La
única columna que se considera
características es aquella que
se quiere predecir

Base de datos de entrenamiento o aprendizaje



Datos nuevos de inferencia o predicción



Característica							Label (conocida)
	Edad	Sexo	Peso	Test 1	Test 2	Test 3	
	35	V	74	No	Sí	No	No
	19	M	56	Sí	Sí	No	Sí

	52	M	70	No	No	Sí	Sí

Entrenamiento de un modelo de IA con la base de datos disponible.



60

V

89

Sí

No

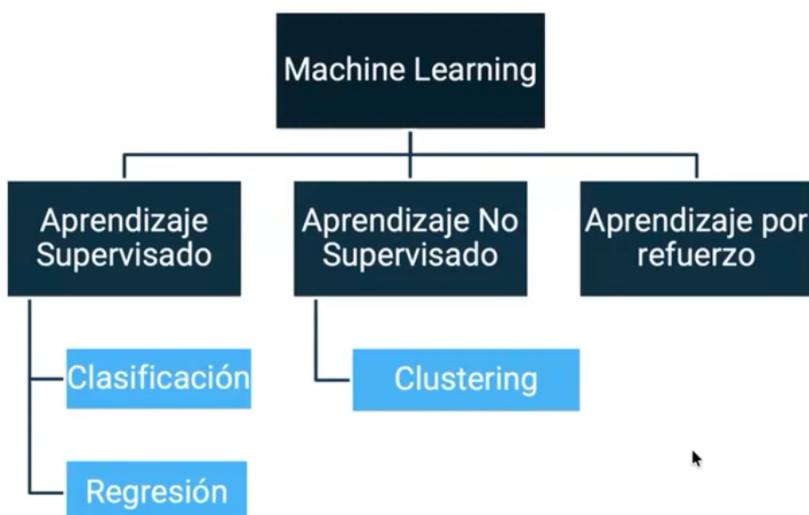
No



Predicción

- Tipología de Machine Learning

Tipos de IA



El machine learning se definió como el proceso en el que a partir de unos datos se

aprende para obtener una predicción. Dado que esto es cierto para toda la, el ML se considera la versión matemáticamente o computacionalmente más sencilla y ha sido la más utilizada y rentable en la industria durante los últimos 25 años.

Dentro del machine learning se distinguen tres tipos de aprendizaje:

1. Aprendizaje supervisado.

- Es aquel en el que el label (lo que se quiere predecir) es conocido en la base de datos de entrenamiento.

- Se divide en dos tipos

- * Clasificación: la variable objetivo es discreta (s/n) (0/1/-1)
- * Regresión: la variable objetivo es continua (esperanza de vida, un número continuo)

2. Aprendizaje o supervisado

- Es aquel en el que se descubre completamente el label. Por ejemplo si se sabe que pacientes enfermarán - Estos modelos se usan para segmentación o clasificación y buscar grupos o patrones matemáticos similares que una guía práctica

3. Spredicaje por refuerzo

- Es una combinación de los dos anteriores, utilizando instancias etiquetadas y no etiquetadas.
- Ciclo de un proyecto de ML

El flujo de trabajo en la muestra se describe con las siguientes etapas:

- 1) Fuentes de datos: conocimiento propio de la empresa (registro de pacientes)
- 2) Análisis del dato y Preprocesamiento:
 - EDD (Exploratory Data Analysis):

Analisis del dato para ver que
información se puede sacar
antes de entregar.

- Data Wrangling / Preprocesamiento :

Proceso para limpiar los datos y
tener el modelo lo más impo.
posible .

3) Modelado:

- Ingeniería de Características (Feature Engineering) : Crear de nuevas características o columnas a partir de información existente para ajustar "chicha" al modelo (calcular IMC, a través de peso y altura).
- Entrenamiento y Ajuste de Modelos (Model Training) : Proceso iterativo

para entregar el modelo y luego ajustar los hiperparámetros

4) Despliegue y Producción: Hacer que el modelo entrenado esté disponible para su uso práctico (P): notificar a un médico en una operación) Es un proceso de mejorar continua donde las predicciones generan feedback que se usa para entrenar y mejorar el modelo. Todo esto se gestiona mediante el procedimiento de CI/CD para el desarrollo del código.

Imagine el proceso de ML como hornear un pastel. La base de datos cruda es la bolsa de ingredientes que usted compra. El Análisis de Datos y Preprocesamiento (EDA y

Wrangling) es revisar los ingredientes, asegurarse de que el azúcar no tiene grumos y el huevo no está roto, dejando solo los ingredientes limpios y listos. La **Ingeniería de Características** es el arte de decidir si añade ralladura de naranja para un sabor extra. Finalmente, el **Modelo Entrenado** es la receta ajustada (con los parámetros correctos) que usted puede usar una y otra vez para hornear un pastel consistente.

Enunciado Claro del Ejercicio: Modelo de Clasificación Supervisado de Vinos

El objetivo de este ejercicio es desarrollar un **modelo de clasificación supervisado** utilizando el mismo conjunto de datos de vino (*dataset*) que fue empleado en el *notebook* de trabajo en clase,

"**2_unsupervised_classification_model.ipynb**" [Enunciado del usuario].

Instrucciones específicas:

1. Dataset y Modelo: Se debe emplear el *dataset* de los vinos visto en la sesión de modelos no supervisados para entrenar un modelo de clasificación supervisado [Enunciado del usuario]. Se recomienda utilizar como referencia el *notebook* "**1_supervised_classification_model.ipynb**" [Enunciado del usuario], el cual, según el instructor, está orientado a entrenar un modelo **XG Boost**.

2. Proceso de Desarrollo: El desarrollo del modelo debe seguir el flujo de trabajo estándar de Machine Learning Supervisado, tomando como guía los pasos realizados en la clase sobre el modelo de clasificación supervisado [Enunciado del usuario]:

- **Análisis Exploratorio de Datos (EDA):** Examinar y analizar las características del conjunto de datos [Enunciado del usuario].
 - **Ingeniería de Características:** Realizar el preprocessamiento y la ingeniería de características necesarias, como el tratamiento de variables categóricas o la normalización de variables numéricas, para que el modelo XGBoost funcione correctamente [Enunciado del usuario, 46, 47].

- **Construcción y Entrenamiento del Modelo:** Construir y entrenar el modelo de clasificación (idealmente XGBoost).

- **Evaluación de Resultados:** Evaluar el rendimiento del modelo utilizando métricas de clasificación [Enunciado del usuario].

3. Expectativa de Resultados: Dado que el *dataset* de vinos está muy tratado y es "muy cuidadito", el objetivo es obtener un modelo con métricas muy buenas, idealmente una **matriz de confusión perfecta**.

Entrega:

Para la entrega, se debe adjuntar directamente el/los archivos requeridos o comprimirlos en un archivo ".zip" [Enunciado del usuario].

En resumen analógico:

Si el *notebook 2* era la demostración de cómo segmentar un mapa de clientes sin saber quiénes eran (aprendizaje no supervisado), la tarea (Pregunta 1) consiste en tomar esos mismos datos de clientes y, como ahora sabemos a qué categoría pertenecen (el "target" o etiqueta), utilizar la "receta" de clasificación supervisada del *notebook 1* para construir el modelo más preciso posible que identifique

correctamente a cada grupo.

