

# INTRODUCCIÓN A MACHINE LEARNING

## ¿QUÉ ES MACHINE LEARNING (APRENDIZAJE AUTOMÁTICO)?

El aprendizaje automático o Machine Learning (ML) impulsa algunas de las tecnologías más importantes que utilizamos, desde aplicaciones de traducción hasta vehículos autónomos. Este documento explica los conceptos básicos del aprendizaje automático.

El ML ofrece una nueva forma de resolver problemas, responder a preguntas complejas y crear nuevos contenidos. El ML puede predecir el tiempo, estimar los tiempos de viaje, recomendar canciones, autocompletar frases, resumir artículos y generar imágenes nunca vistas.

En términos básicos, el ML es el proceso de entrenar una pieza de software, llamada **modelo**, para **hacer predicciones útiles o generar contenidos a partir de datos**.

Por ejemplo, supongamos que queremos crear una aplicación para predecir las precipitaciones. Podríamos utilizar un enfoque tradicional o un enfoque de ML. Con un enfoque tradicional, crearíamos una representación basada en la física de la atmósfera y la superficie de la Tierra, calculando cantidades ingentes de ecuaciones de dinámica de fluidos. Esto es increíblemente difícil.

Con un enfoque basado en ML, proporcionaríamos a un modelo ML enormes cantidades de datos meteorológicos hasta que el modelo ML acabara aprendiendo la relación matemática entre los patrones meteorológicos que producen diferentes cantidades de lluvia. A continuación, le daríamos al modelo los datos meteorológicos actuales y éste predeciría la cantidad de lluvia.

### Tipos de sistemas de ML

Los sistemas de ML se clasifican en una o varias de las siguientes categorías en función de cómo aprenden a hacer predicciones o a generar contenidos:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje por refuerzo
- IA generativa

### Aprendizaje Supervisado

Los modelos de aprendizaje supervisado pueden hacer predicciones tras ver muchos datos con las respuestas correctas y descubrir después las conexiones entre los elementos de los datos que producen las respuestas correctas. Es como si un estudiante aprendiera material nuevo estudiando exámenes antiguos que contienen preguntas y respuestas. Una vez que el estudiante se ha entrenado con suficientes exámenes antiguos, está bien preparado para presentarse a un nuevo examen. Estos sistemas ML están "supervisados" en el sentido de que un humano proporciona al sistema ML datos con los resultados correctos conocidos.

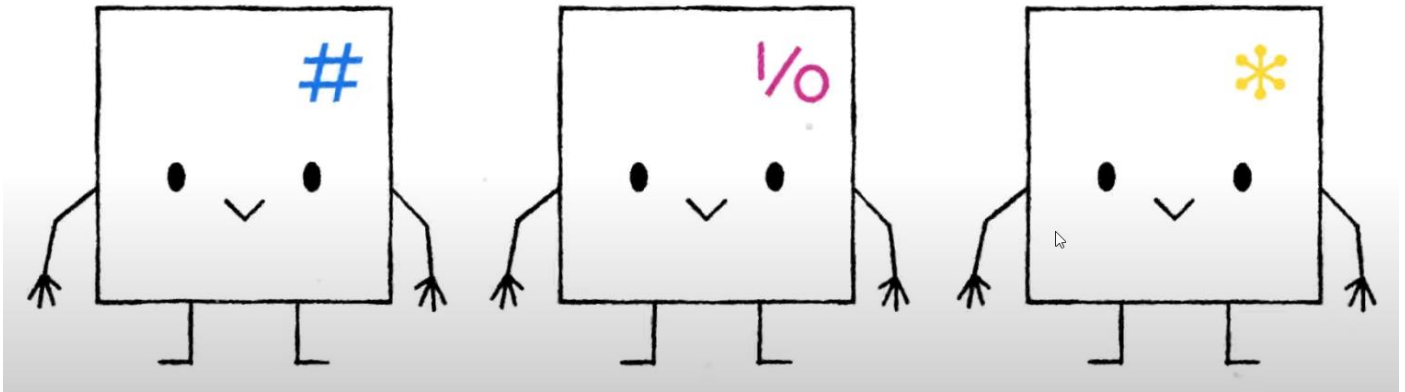
Dos de los casos de uso más comunes del aprendizaje supervisado son la **regresión** y la **clasificación**.

# INTRODUCCIÓN A MACHINE LEARNING

Regression

Binary  
Classification

Multiclass  
Classification



## Regresión

Un modelo de regresión **predice un valor numérico**. Por ejemplo, un modelo meteorológico que predice la cantidad de lluvia, en pulgadas o milímetros, es un modelo de regresión.

Consulta la tabla siguiente para ver más ejemplos de modelos de regresión:

Escenario	Posibles datos de entrada	Predicción numérica
<b>Precio futuro de la vivienda</b>	Metros cuadrados, código postal, número de dormitorios y cuartos de baño, tamaño del terreno, tipo de interés hipotecario, tipo del impuesto sobre bienes inmuebles, costes de construcción y número de viviendas en venta en la	El precio de la vivienda.
<b>Tiempo de viaje futuro</b>	Condiciones históricas del tráfico (recopiladas a partir de smartphones, sensores de tráfico, servicios de transporte por carretera y otras aplicaciones de navegación), distancia al destino y condiciones meteorológicas.	El tiempo en minutos y segundos para llegar a un destino.

## Clasificación

Los modelos de clasificación **predicen la probabilidad de que algo pertenezca a una categoría**. A diferencia de los modelos de regresión, cuyo resultado es un número, los modelos de clasificación producen un valor que indica si algo pertenece o no a una categoría determinada. Por ejemplo, los modelos de clasificación se utilizan para predecir si un correo electrónico es spam o si una foto contiene un gato.

Los modelos de clasificación se dividen en dos grupos: **clasificación binaria** y **clasificación multiclase**. Los modelos de **clasificación binaria emiten un valor de una clase que sólo contiene dos valores**, por ejemplo, un modelo que emite lluvia o no lluvia. Los modelos de **clasificación multiclase dan como resultado un valor de una clase que contiene más de dos valores**, por ejemplo, un modelo que puede dar como resultado lluvia, granizo, nieve o aguanieve.

# INTRODUCCIÓN A MACHINE LEARNING

## Aprendizaje no supervisado

Los modelos de **aprendizaje no supervisado** realizan **predicciones a partir de datos que no contienen respuestas correctas**. El objetivo de un modelo de aprendizaje no supervisado es **identificar patrones** significativos entre los datos. En otras palabras, el modelo no tiene pistas sobre cómo categorizar cada dato, sino que debe inferir sus propias reglas.

Un modelo de aprendizaje no supervisado muy utilizado emplea una técnica denominada **agrupación o clusterización**. El modelo encuentra puntos de datos que delimitan agrupaciones naturales.

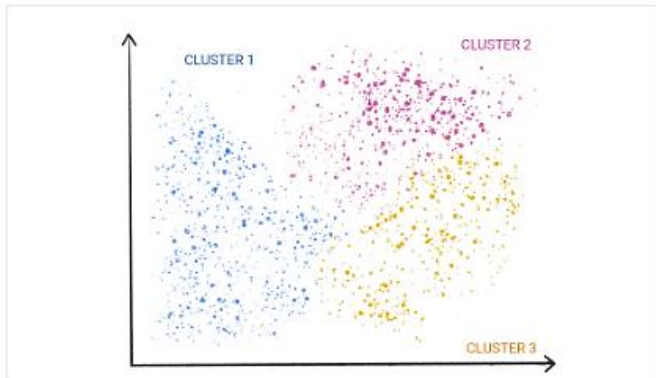


Ilustración 1. Un modelo ML que agrupa puntos de datos similares.

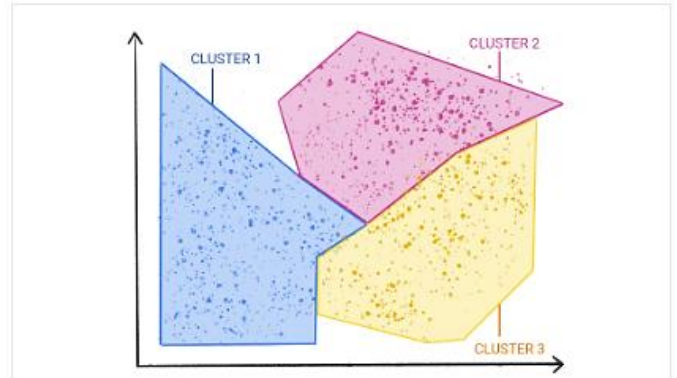


Ilustración 2. Grupos de clusters con demarcaciones naturales.

La **agrupación difiere de la clasificación en que las categorías no están definidas por el usuario**. Por ejemplo, un modelo no supervisado puede agrupar un conjunto de datos meteorológicos en función de la temperatura, revelando segmentaciones que definen las estaciones. A continuación, podría intentar nombrar esos grupos basándose en su comprensión del conjunto de datos.

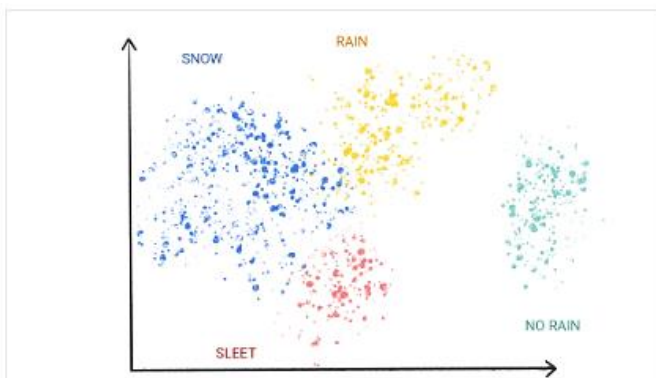


Ilustración 3. Un modelo ML que agrupa patrones meteorológicos similares.

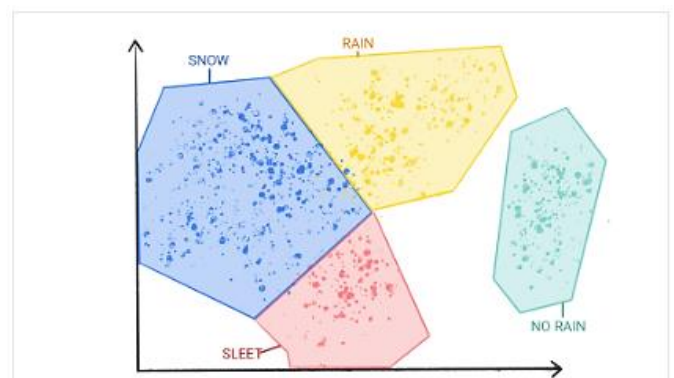


Ilustración 4. Agrupaciones de patrones meteorológicos etiquetados como nieve, aguanieve, lluvia y sin lluvia.

## Aprendizaje por refuerzo

Los modelos de **aprendizaje por refuerzo** hacen **predicciones obteniendo recompensas o penalizaciones en función de las acciones realizadas en un entorno**. Un sistema de aprendizaje por refuerzo genera una política que define la mejor estrategia para obtener el mayor número de recompensas.

# INTRODUCCIÓN A MACHINE LEARNING

El aprendizaje por refuerzo se utiliza para entrenar robots a realizar tareas, como caminar por una habitación, y programas de software como [AlphaGo](#) para jugar al Go.

## IA Generativa


La IA generativa **es una clase de modelos que crea contenidos a partir de la información introducida por el usuario**. Por ejemplo, la IA generativa puede crear imágenes novedosas, composiciones musicales y chistes; puede resumir artículos, explicar cómo realizar una tarea o editar una foto.

La IA generativa **puede tomar una serie de entradas y crear una variedad de salidas**, como texto, imágenes, audio y vídeo. También puede tomar y crear combinaciones de estos elementos. Por ejemplo, un modelo puede tomar una imagen como entrada y crear una imagen y un texto como salida, o tomar una imagen y un texto como entrada y crear un vídeo como salida.

Podemos hablar de modelos generativos por sus entradas y salidas, que suelen escribirse como "tipo de entrada"- "tipo de salida". Por ejemplo, la siguiente es una lista parcial de algunas entradas y salidas de los modelos generativos:

- Texto a texto
- Texto a imagen
- Texto a vídeo
- Texto a código
- Texto a voz
- Imagen y texto a imagen

En la siguiente tabla se enumeran ejemplos de modelos generativos, su entrada y un ejemplo de su posible salida:

Modelo	Entrada	Ejemplo de Salida
Texto a texto	¿Quién inventó la carrera de Le Mans?	Las 24 Horas de Le Mans fueron creadas por el Automobile Club de l'Ouest (ACO), una organización automovilística fundada en 1906. La primera carrera se celebró los días 26 y 27 de mayo de 1923 en el Circuito de la Sarthe, un circuito de carretera de 13,426 kilómetros (8,345 millas) situado cerca de la ciudad de Le Mans, en Francia. La carrera se concibió originalmente como una
Texto a imagen	Un pulpo alienígena flota a través de un portal leyendo un periódico.	

# INTRODUCCIÓN A MACHINE LEARNING

## Texto a video

Un oso de peluche fotorrealista está nadando en el océano en San Francisco. El osito se sumerge en el agua. El oso de peluche sigue nadando bajo el agua con peces de colores. Un oso panda nada bajo el agua.

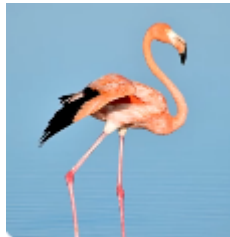


## Texto a código

Escribe un bucle en Python que recorra una lista de números e imprima los números primos.

```
for number in numbers:
    # Check if the number is prime.
    is_prime = True
    for i in range(2, number):
        if number % i == 0:
            is_prime = False
            break
    # If the number is prime, print it.
    if is_prime:
        print(number)
```

## Imagen a texto



Esto es un flamenco. Se encuentran en el Caribe.

¿Cómo funciona la IA generativa? A alto nivel, los modelos generativos aprenden patrones en los datos con el objetivo de producir datos nuevos pero similares. Los modelos generativos son como los siguientes:

- Los cómicos que aprenden a imitar a otros observando los comportamientos y el estilo de hablar de la gente.
- Los artistas que aprenden a pintar en un estilo concreto estudiando muchos cuadros de ese estilo.
- Grupos de versiones que aprenden a sonar como un grupo musical concreto escuchando mucha música de ese grupo.

Para producir resultados únicos y creativos, **los modelos generativos se entrenan inicialmente con un enfoque no supervisado**, en el que el modelo aprende a imitar los datos con los que se ha entrenado. A veces, el modelo se entrena aún más utilizando el aprendizaje supervisado o de refuerzo sobre datos específicos relacionados con las tareas que se le pueden pedir al modelo, por ejemplo, resumir un artículo o editar una foto.

La IA generativa es una tecnología que evoluciona rápidamente y en la que se descubren constantemente nuevos casos de uso. Por ejemplo, los modelos generativos están ayudando a las empresas a perfeccionar las imágenes de sus productos de comercio electrónico eliminando automáticamente los fondos que distraen o mejorando la calidad de las imágenes de baja resolución.

# INTRODUCCIÓN A MACHINE LEARNING

## APRENDIZAJE SUPERVISADO

Las tareas del aprendizaje supervisado están bien definidas y pueden aplicarse a multitud de escenarios, como la identificación de spam o la predicción de precipitaciones.

### Conceptos básicos del aprendizaje supervisado

El aprendizaje automático supervisado se basa en los siguientes conceptos fundamentales:

- Datos
- Modelo
- Entrenamiento
- Evaluación
- Inferencia (Deducción)

#### Datos

Los datos son la fuerza motriz del ML. Los datos se presentan en forma de palabras y números almacenados en tablas, o como valores de píxeles y formas de onda capturados en imágenes y archivos de audio. Los datos relacionados se almacenan en conjuntos de datos. Por ejemplo, podríamos tener un conjunto de datos con lo siguiente:

- Imágenes de gatos
- Precios de viviendas
- Información meteorológica

Los conjuntos de datos están formados por ejemplos individuales que contienen **características (features)** y una **etiqueta (label)**. Un ejemplo es análogo a una fila de una hoja de cálculo. **Las características son los valores que un modelo supervisado utiliza para predecir la etiqueta. La etiqueta es la "respuesta" o el valor que queremos que el modelo prediga.** En un modelo meteorológico que predice las precipitaciones, las características podrían ser la latitud, la longitud, la temperatura, la humedad, la cobertura de nubes, la dirección del viento y la presión atmosférica. La etiqueta sería la cantidad de lluvia.

Los ejemplos que contienen características y una etiqueta se denominan ejemplos etiquetados.

#### Dos ejemplos etiquetados:

Features								Label
date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example



# INTRODUCCIÓN A MACHINE LEARNING

Por el contrario, los ejemplos no etiquetados contienen características, pero no una etiqueta. Después de crear un modelo, éste predice la etiqueta a partir de las características.

**Dos ejemplos no etiquetados:**

Features							
date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

Example

## Características del conjunto de datos

Un conjunto de datos se caracteriza por su **tamaño** y su **diversidad**. El tamaño indica el número de **ejemplos**. La diversidad indica la gama que abarcan esos ejemplos. Los buenos conjuntos de datos son a la vez grandes y muy diversos.

Algunos conjuntos de datos son a la vez grandes y diversos. Sin embargo, algunos conjuntos de datos son grandes, pero tienen poca diversidad, y otros son pequeños pero muy diversos. En otras palabras, un conjunto de datos grande no garantiza una diversidad suficiente, y un conjunto de datos muy diverso no garantiza un número suficiente de ejemplos.

Por ejemplo, un conjunto de datos puede contener datos de 100 años, pero sólo del mes de julio. Utilizar este conjunto de datos para predecir las precipitaciones en enero daría lugar a predicciones deficientes. Por el contrario, un conjunto de datos puede abarcar sólo unos pocos años, pero contener todos los meses. Este conjunto de datos puede dar lugar a predicciones erróneas porque no contiene suficientes años para tener en cuenta la variabilidad.

Un conjunto de datos **también puede caracterizarse por el número de sus características**. Por ejemplo, algunos conjuntos de datos meteorológicos pueden contener cientos de características, desde imágenes de satélite hasta valores de cobertura de nubes. Otros conjuntos de datos pueden contener sólo tres o cuatro características, como humedad, presión atmosférica y temperatura. Los conjuntos de datos con más características pueden ayudar a un modelo a descubrir patrones adicionales y hacer mejores predicciones. Sin embargo, los conjuntos de datos con más características no siempre producen modelos que hagan mejores predicciones porque algunas características pueden no tener una relación causal con la etiqueta.

## Modelo

En el aprendizaje supervisado, **un modelo es la compleja colección de números que definen la relación matemática entre patrones de características de entrada específicos y valores de etiquetas de salida específicos**. El modelo descubre estos patrones a través del entrenamiento.

# INTRODUCCIÓN A MACHINE LEARNING

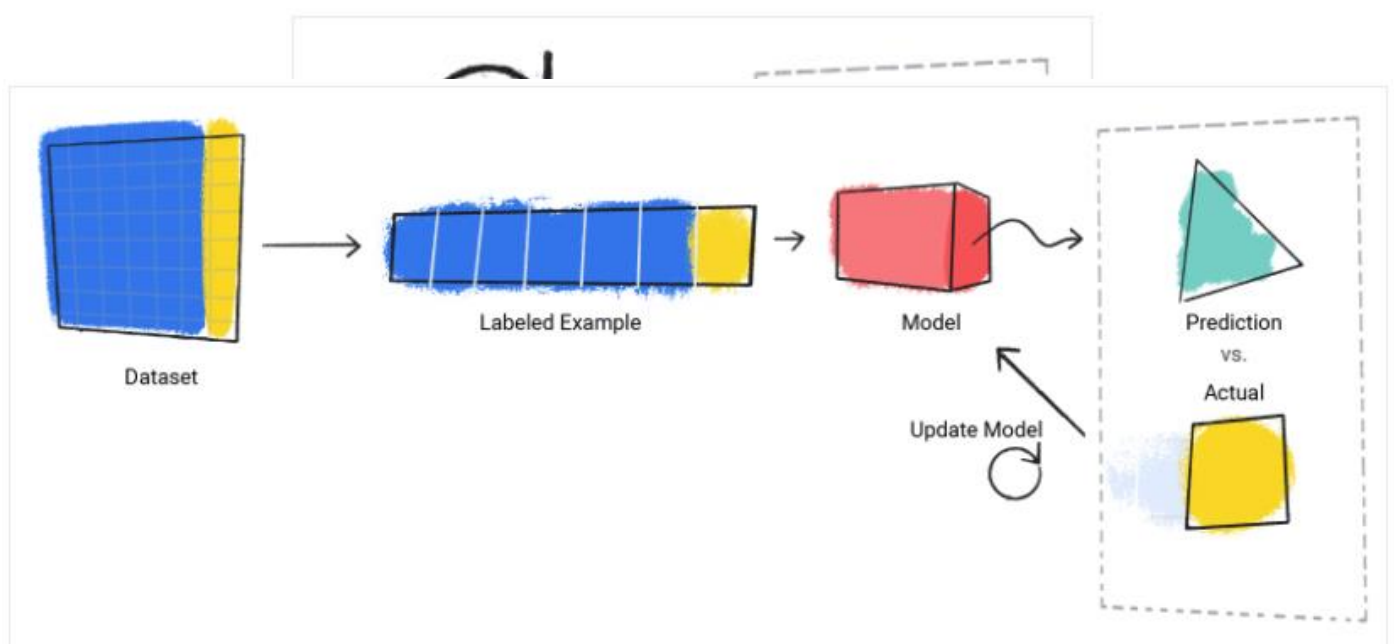
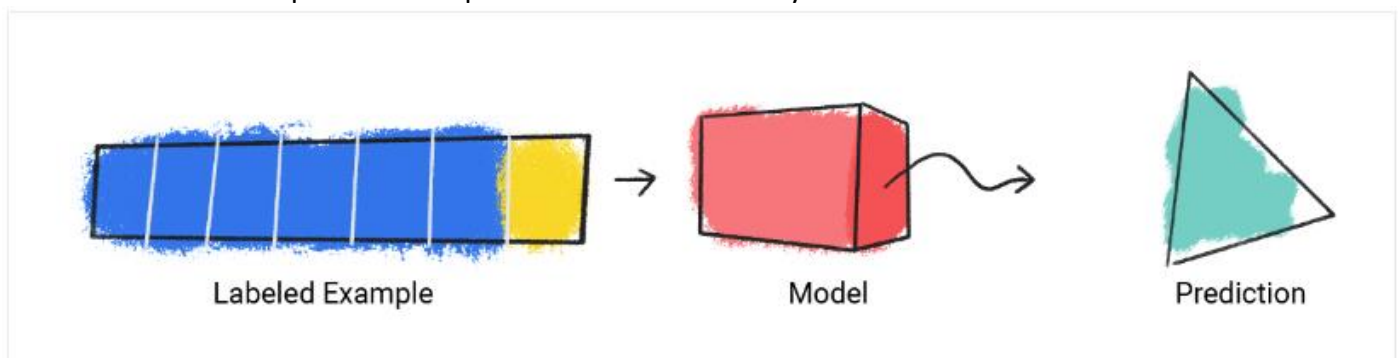
## Entrenamiento

Antes de que un modelo supervisado pueda hacer predicciones, **debe ser entrenado**. Para entrenar un modelo, le damos un conjunto de datos con ejemplos etiquetados. El objetivo del modelo es encontrar la mejor solución para predecir las etiquetas a partir de las características. El modelo encuentra la mejor solución comparando el valor predicho con el valor real de la etiqueta. En función de la diferencia entre los valores predichos y los reales -definida como pérdida-, el modelo actualiza gradualmente su solución. En otras palabras, **el modelo aprende la relación matemática entre las características y la etiqueta para poder hacer las mejores predicciones sobre datos no vistos**.

Por ejemplo, si el modelo predijo 1,15 pulgadas de lluvia, pero el valor real fue de 0,75 pulgadas, el modelo modifica su solución para que su predicción se acerque más a 0,75 pulgadas. Una vez que el modelo ha examinado cada ejemplo del conjunto de datos (en algunos casos, varias veces), llega a una solución que realiza las mejores predicciones, de media, para cada uno de los ejemplos.

A continuación, se muestra el entrenamiento de un modelo:

1. El modelo toma un único ejemplo etiquetado y proporciona una predicción.
2. El modelo compara su valor previsto con el valor real y actualiza su solución.



3. El modelo repite este proceso para cada ejemplo etiquetado del conjunto de datos.



# INTRODUCCIÓN A MACHINE LEARNING

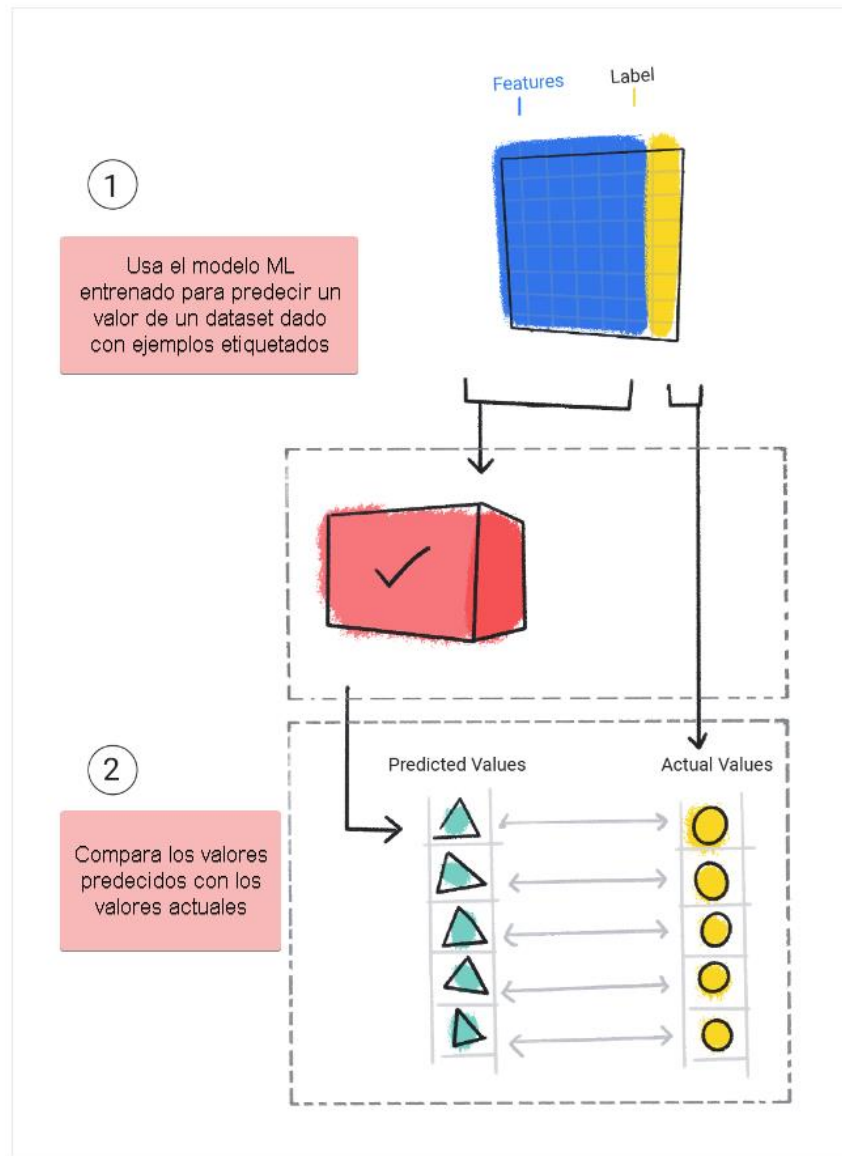
De este modo, el modelo aprende gradualmente la relación correcta entre las características y la etiqueta. Esta comprensión gradual es también la razón por la que los conjuntos de datos grandes y diversos producen un modelo mejor. El modelo ha visto más datos con una gama más amplia de valores y ha refinado su comprensión de la relación entre las características y la etiqueta.

Durante el entrenamiento, los profesionales del ML **pueden realizar ajustes sutiles en las configuraciones y características que el modelo utiliza para hacer predicciones**. Por ejemplo, algunas características tienen más poder predictivo que otras. Por lo tanto, los profesionales del ML pueden seleccionar qué características utiliza el modelo durante el entrenamiento. Por ejemplo, supongamos que un conjunto de datos meteorológicos contiene la hora\_del\_día como característica. En este caso, un experto en ML puede añadir o eliminar hora\_del\_día durante el entrenamiento para ver si el modelo hace mejores predicciones con o sin ella.

## Evaluación

Evaluamos un modelo entrenado para determinar lo bien que ha aprendido. Cuando evaluamos un modelo, utilizamos un conjunto de datos etiquetados, pero sólo damos al modelo las características del conjunto de datos. A continuación, comparamos las predicciones del modelo con los valores reales de la etiqueta.

# INTRODUCCIÓN A MACHINE LEARNING



Dependiendo de las predicciones del modelo, podríamos realizar más entrenamientos y evaluaciones antes de desplegar el modelo en una aplicación del mundo real.

## Inferencia

Una vez que estamos satisfechos con los resultados de la evaluación del modelo, podemos utilizarlo para hacer predicciones, denominadas inferencias, sobre ejemplos no etiquetados. En el ejemplo de la aplicación meteorológica, daríamos al modelo las condiciones meteorológicas actuales: temperatura, presión atmosférica y humedad relativa - y predeciría la cantidad de lluvia.