Resumen del capítulo: Primer modelo entrenado

Dataset para el entrenamiento

En realidad, programas y especialistas se capacitan de manera similar: recopilan y clasifican conocimientos, descubren dependencias y adquieren experiencia. Tanto el proceso de aprendizaje de un humano como el machine learning incluyen cierto material de estudio. En el caso del machine learning, los modelos aprenden de conjuntos de datos de entrenamiento.

En el análisis de datos, las filas se llaman instancias, mientras que las columnas son las variables. En el machine learning, las filas y columnas representan **observaciones** y **características**, respectivamente. La característica que necesitamos predecir se llama **objetivo**.

Aprendizaje supervisado

Tienes un conjunto de datos de entrenamiento y una característica objetivo (precio de venta de la propiedad) que necesitas predecir usando el resto de las características. Esta es una tarea de **aprendizaje supervisado**. El "maestro" plantea **preguntas** (características) y da **respuestas** (el objetivo). No se da ninguna explicación sobre cómo las características conducen a la respuesta exactamente; la máquina tiene que resolverlo por sí misma. El aprendizaje supervisado resulta conveniente para resolver múltiples tareas comerciales.

También hay otras clases:

- **aprendizaje no supervisado**: sin objetivo
- **aprendizaje semisupervisado**: solo una parte de los datos de entrenamiento conoce el objetivo
- **Recomendación**: los usuarios y los elementos reemplazan las funciones y las observaciones (algo que puedas recomendar, por ejemplo, películas o vecindarios).

Veamos los tipos de aprendizaje supervisado.

Todas las variables y características son categóricas o numéricas, y el objetivo no es una excepción.

Las tareas de clasificación se ocupan de objetivos categóricos (por ejemplo, determinar especies de animales en una imagen). Cuando solo tenemos dos categorías (por ejemplo, si un cliente volverá a visitar el sitio web o no), se le denomina **clasificación binaria**.

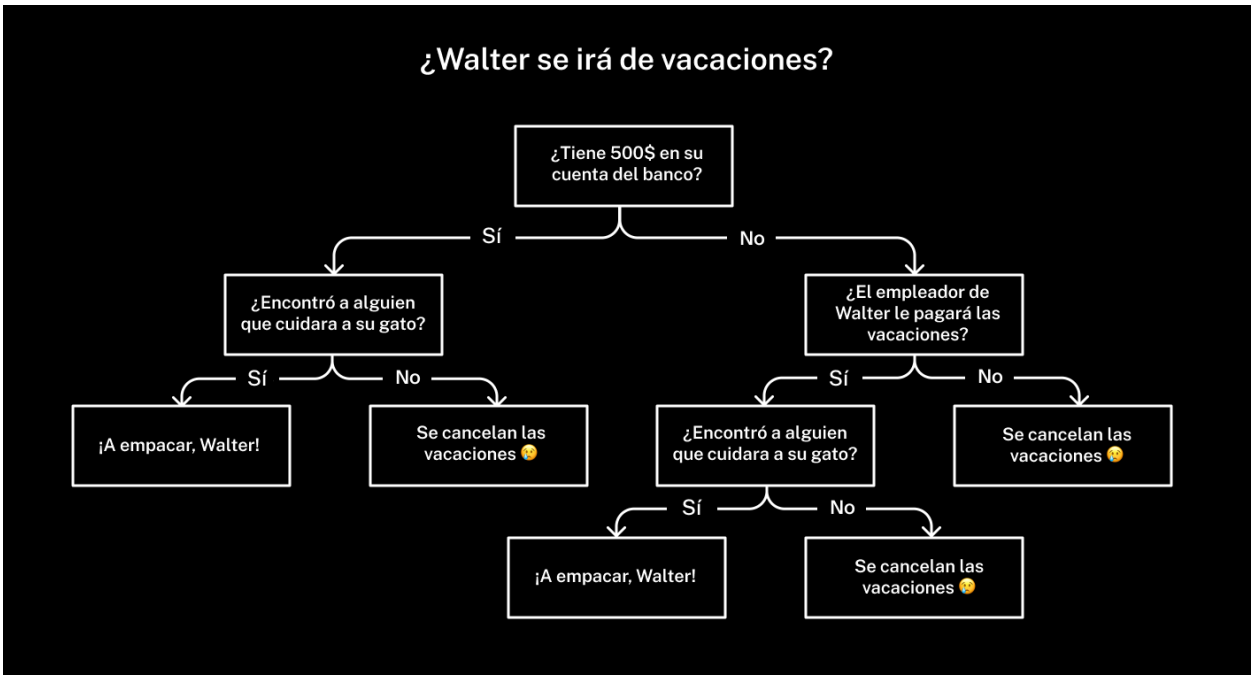
Si el objetivo es numérico, entonces es una **tarea de regresión**. Los datos se utilizan para encontrar relaciones entre las variables y hacer predicciones basadas en la información, como el pronóstico del tiempo o la predicción de los precios del mercado de valores para los próximos días.

Modelos y algoritmos

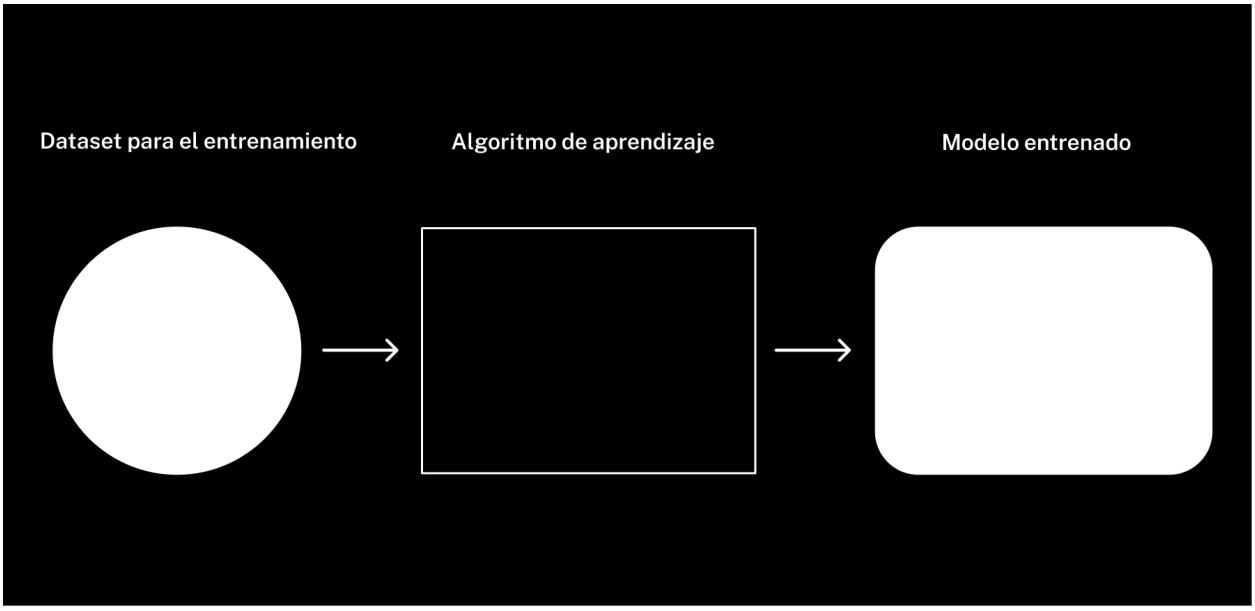
Supongamos que existe algún tipo de relación entre las características y el objetivo. Para hacer predicciones, la máquina tiene que entender cuál es esa relación. Pero no es posible tener en cuenta todas las posibles razones por las que el objetivo haya salido tal y como es. Por ello, tenemos que simplificar esta compleja relación y recurrir a los **modelos de machine learning**.

Existen muchos modelos diferentes que pueden utilizarse para reflejar cómo las características se transforman en el objetivo, y cada uno de ellos conlleva sus propias suposiciones sobre cómo se estructura dicha relación. Quien trabaja en la ciencia de datos acepta estas suposiciones eligiendo un modelo, para luego utilizarlo con el fin de hacer predicciones. Si estas predicciones coinciden con la realidad, significa que las suposiciones eran lo suficientemente precisas y que el modelo elegido era el correcto. Este enfoque se denomina **modelado**.

Un modelo popular se llama **árbol de decisiones**. Este puede describir el proceso de toma de decisiones en casi cualquier situación. Así es como hacemos un árbol de decisiones con respuestas sí/no y diferentes escenarios.



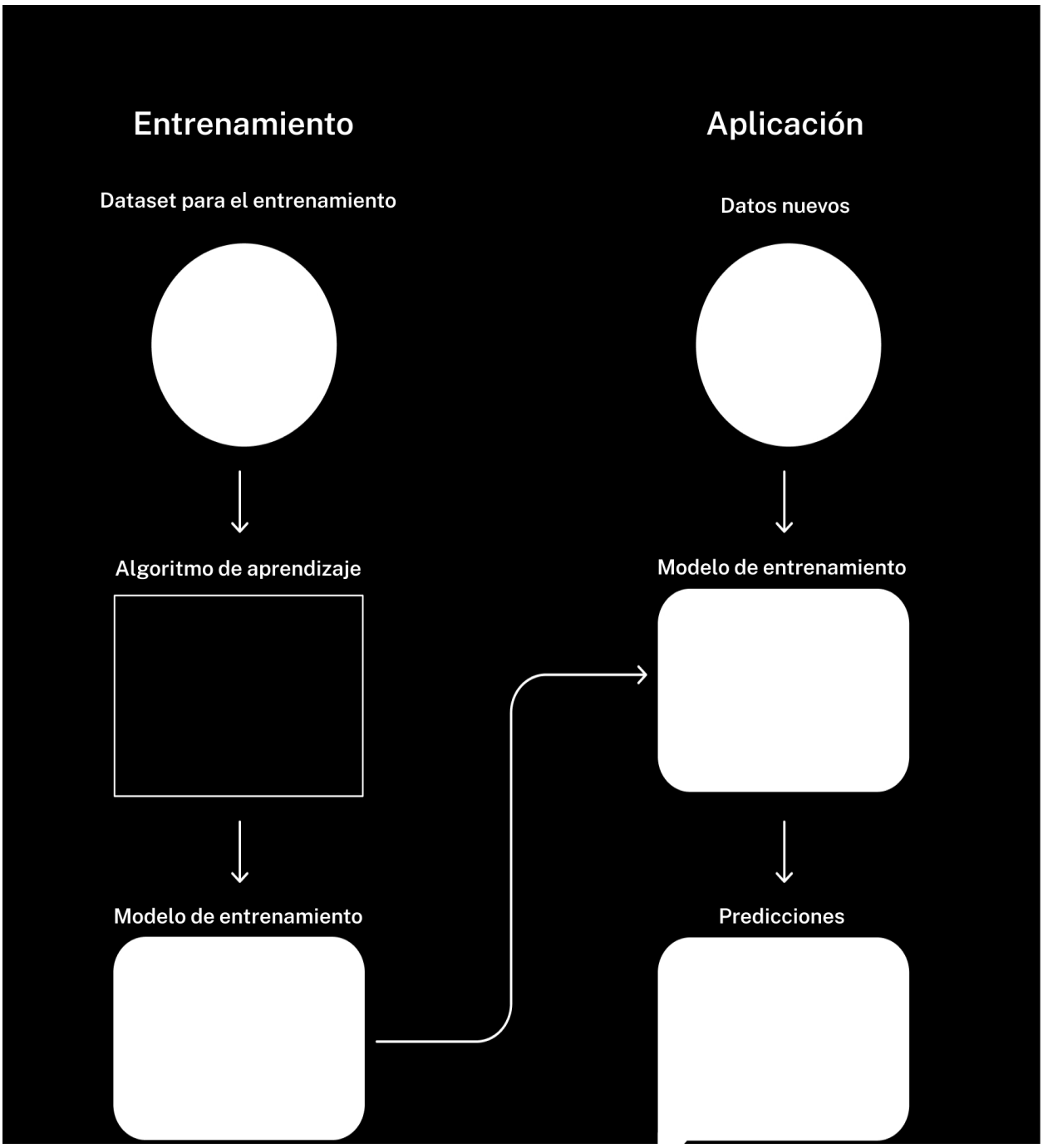
Cada árbol sale diferente. Vamos a **entrenar el modelo** para construir el más adecuado. Además del conjunto de datos, necesitaremos un **algoritmo de aprendizaje**. El conjunto de datos se procesa a través de nuestro algoritmo de aprendizaje, produciendo un **modelo entrenado**.



Una vez entrenado, el modelo está listo para hacer **predicciones**, es decir, para tomar nuevas características de entrada y generar respuestas (el objetivo), sin necesidad de

recurrir a algoritmos ni a conjuntos de datos de entrenamiento.

Es importante recordar que el proceso de machine learning consta de dos pasos: entrenamiento del modelo y aplicación del mismo.



Librería Scikit-learn

Los algoritmos de aprendizaje suelen ser más complejos que los modelos. Así que, por ahora, considéralos como **cajas negras** y no pienses demasiado en lo que está pasando dentro. Céntrate mejor en lo que hay que poner ahí y en lo que hay que hacer con el resultado.

Las librerías de Python ofrecen muchos algoritmos. En esta lección trabajaremos con la librería popular **scikit-learn**, o **sklearn** (*scientific kit for learning). Sklearn* es una gran fuente de herramientas para trabajar con datos y modelos. Para mayor comodidad, la librería está dividida en módulos. Los árboles de decisión se almacenan en el módulo **tree**.

Cada modelo corresponde a una clase separada en *sklearn.* **DecisionTreeClassifier** es una clase para clasificaciones de árboles de decisión. Vamos a importarla de la librería:

```
from sklearn.tree import DecisionTreeClassifier
```

Luego creamos una instancia de la clase:

```
model = DecisionTreeClassifier()
```

Ahora la variable `model` almacena nuestro modelo, y tenemos que ejecutar un algoritmo de aprendizaje para entrenar al modelo para que haga predicciones.

Para iniciar el entrenamiento, llama al método **fit()** y pásale nuestras variables como argumento.

```
model.fit(features, target)
```

Ahora tenemos un modelo entrenado en la variable `model`. Para predecir respuestas, llama al método **predict()** y pásale la tabla con las características de las nuevas observaciones.

```
answer = model.predict(new_features)
```