Actividad 3. Entrenar con datos sesgados: cuando la IA aprende “mal"

En esta propuesta, además de **afianzar como funciona el aprendizaje supervisado**, tiene como objetivo comprender cómo pueden llegar a **formarse sesgos** en las respuestas de la IA ya sean de un modo voluntario o involuntario.

Si el alumnado aun está interiorizando esta clase de conceptos, se recomienda realizar la propuesta didáctica de primer ciclo

Entrenando a nuestra IA con sesgos

Objetivos:

- Comprender que los sistemas de IA aprenden de los datos que les damos
- Reconocer que si los datos están sesgados, la IA puede tomar decisiones injustas o incorrectas.
- Desarrollar pensamiento crítico sobre la tecnología.

Recursos materiales:

- Ordenadores con cámara web o tablets, con acceso a EchidnaML (local)
- Cartulinas o tarjetas con imágenes de diferentes especies, en este caso perros blancos, perros marrones, primates y perros negros.
- Ficha de reflexión

Temporalización: 60 minutos

Introducción

En primer lugar, explicaremos brevemente qué es una Inteligencia Artificial y cómo aprende, lo que viene a ser el aprendizaje supervisado. Por ejemplo podríamos decirles que "*la IA es como un robot que aprende viendo muchos ejemplos, igual que vosotros y vosotras aprendéis a reconocer animales viendo fotos o videos*".

Una vez llegado a este punto, introduciremos el concepto de sesgo: "*si le damos al robot solo fotos de gatos blancos, puede pensar que todos los gatos son blancos. Eso es un sesgo.*"

Preparando el experimento

Para comenzar con el experimento, haremos agrupaciones de 4-5 alumnos. Cada uno de los grupos se encargará de entrenar a una IA para clasificar animales según especie como la que hemos incluido en el apartado recursos. La mitad de los grupos de alumnos recibirá un set de cartas, la otra mitad otro.

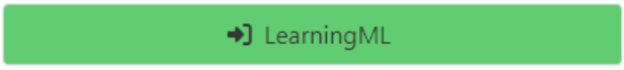
- **SET A:** Perros blancos, perros marrones y primates.
- **SET B:** Perros, blancos, perros marrones, **perros negros** y primates.

> **Importante:** la mitad de los grupos recibirán datos sesgados a propósito (por ejemplo, sólo animales de un tipo) y otros datos más variados (SET A vs SET B)

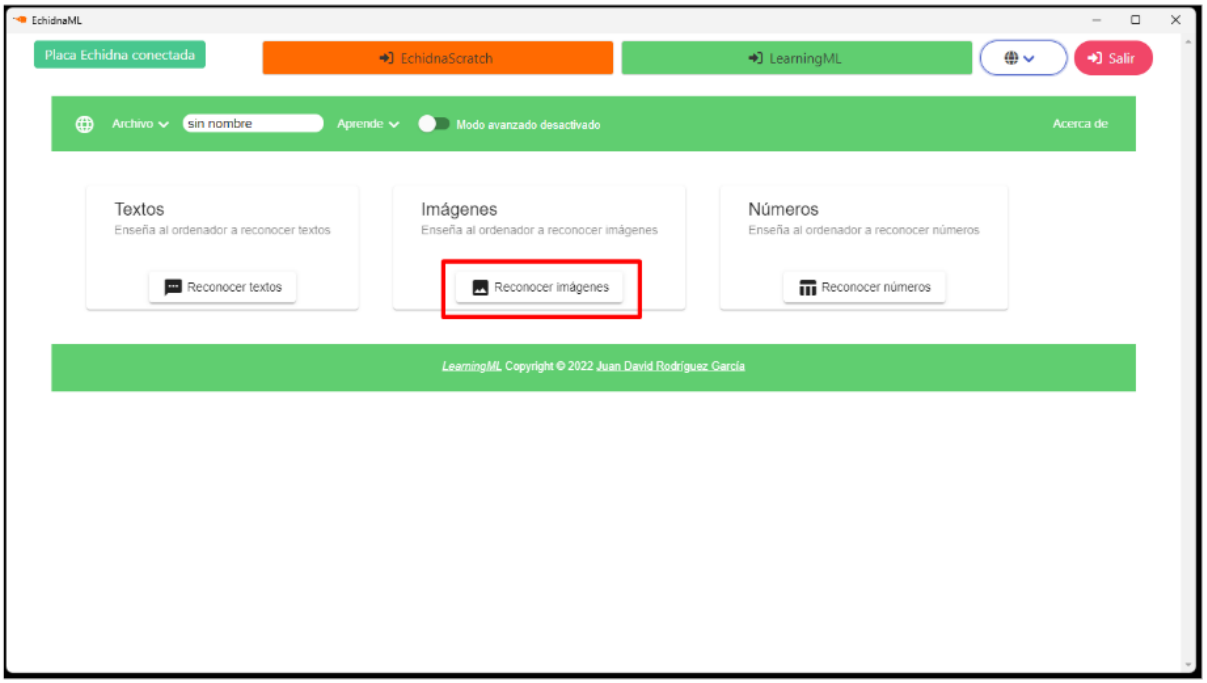
Alimentando a la IA

Los grupos, cargarán sus imágenes en LearningML.

Para ello, abriremos EchidnaML y entraremos en su apartado LearningML (para esta práctica no es necesaria la placa echidna).

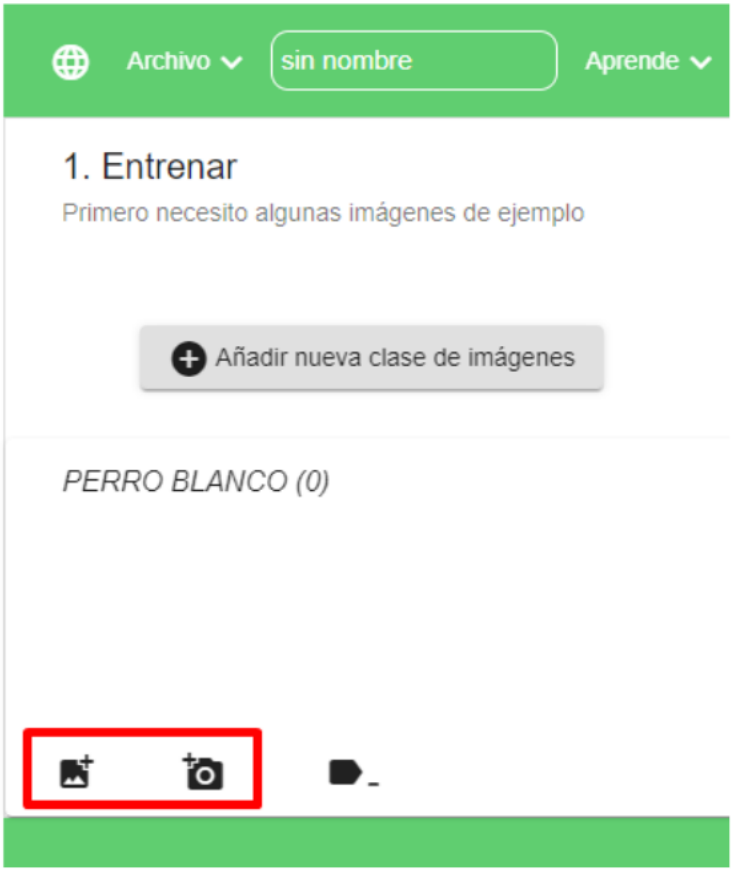


Escogemos el entrenamiento de imágenes



Entrenan el modelo con las categorías en función de las cartas recibidas: PERRO BLANCO, PERRO MARRÓN, PRIMATE y si es el caso, PERRO NEGRO.

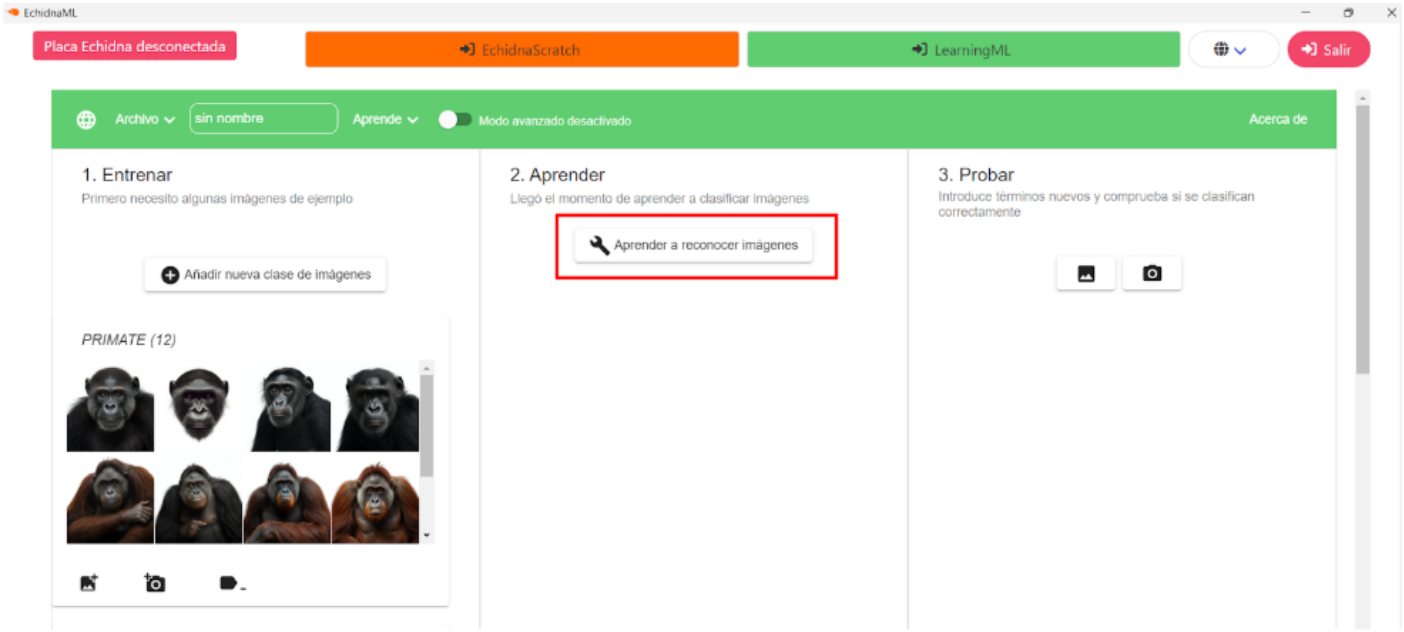
Escogemos un nombre para esta categoría, por ejemplo “PERRO BLANCO” y ahora tenemos dos opciones para alimentarla, o bien activar la webcam para capturar las tarjetas impresas o bien subir las imágenes en formato digital directamente.



Una vez están subidas las imágenes de esa categoría, crearíamos una nueva categoría y repetiríamos el proceso.

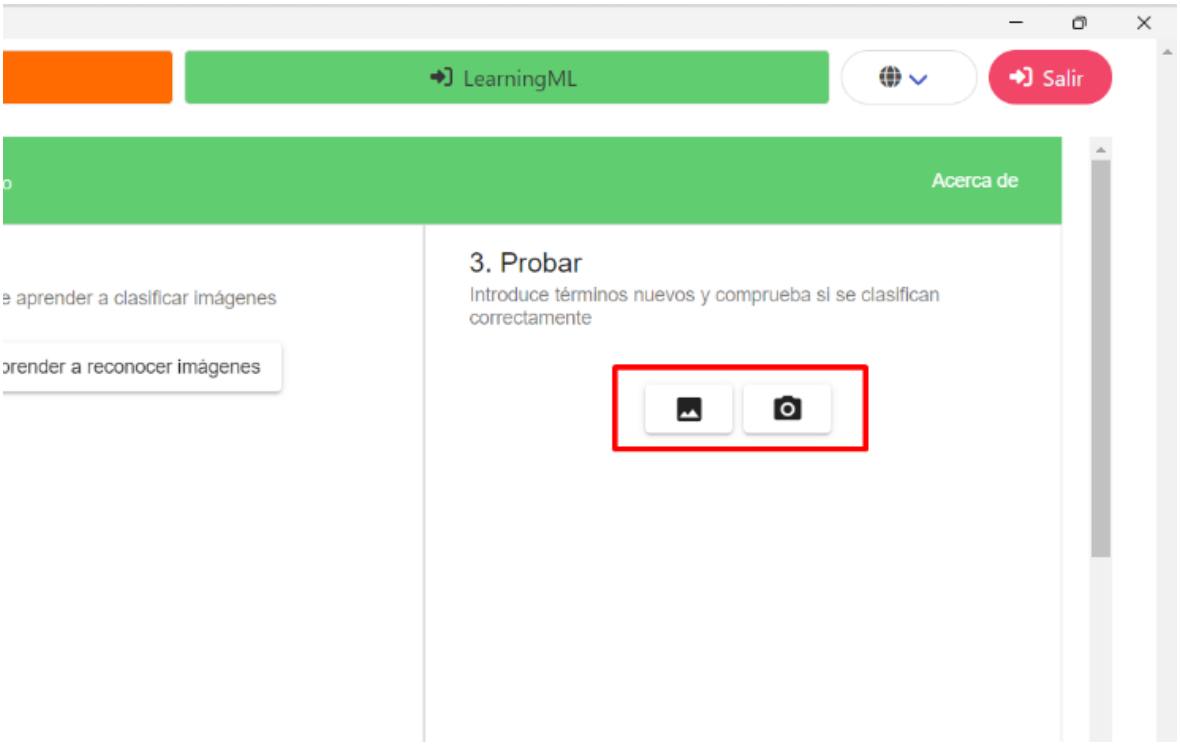
> **NOTA***: En ocasiones es recomendable crear una categoría en la que no haya ningún elemento a aprender, sobre todo cuando se utiliza la opción webcam para evitar confusiones. Se puede llamar a esta categoría como "NADA" para que todo lo que no identifique con las otras categorías vaya ahí.

Cuando hemos terminado el entrenamiento (es decir, hemos terminado de alimentar a la IA con imágenes), es el momento de hacerle que aprenda el modelo:



Finalmente, observamos cómo la IA clasifica nuevas imágenes de prueba.

Para ello, nos vamos a **PROBAR** y tenemos dos opciones, subirle nuevas imágenes o mostrárselas a la webcam.



En este caso vamos a subir una imagen de un perro negro (categoría que no ha sido creada, pues era un set de tarjetas del grupo A).

Aquellos grupos que hayan recibido tarjetas de perros negros y hayan decidido crear la categoría perro negro, obtendrán respuestas acertadas mientras que aquellos que únicamente recibieron fotografías de perros blancos, marrones y primates pero no de perros negros, interpretará a estos como primates (por similitud -color-)

Observación y discusión (10-15 min)

1. Cada grupo prueba su IA con imágenes fuera de su conjunto de entrenamiento.
2. Comparan los resultados:
    - ¿Qué pasó cuando las imágenes de prueba eran diferentes de las que entrenaron?
    - ¿La IA acertó siempre? ¿Falló? ¿Por qué creen que pasó eso?
3. Reflexionan sobre sesgos:
    - “Si entrenamos solo con fotos de perros blancos, ¿la IA reconocerá a los perros negros?”
    - ¿Por qué uno de los grupos fallaba?
    - ¿Cómo se puede solucionar este problema?

Revision #7
Created 2026-01-21 16:37:24 CET by Maria
Updated 2026-03-16 19:14:58 CET by Juan José Mejías