Actividad 2: ¡Crea una IA justa! El laboratorio de sesgos

¡Crea una IA justa! El laboratorio de sesgos

JUSTIFICACIÓN

Esta actividad es útil para que el alumnado conozca que “aprender” en IA significa **crear un modelo** a partir de **datos** y luego **probarlo** con casos nuevos. El alumnado detecta **sesgos** cuando ciertos grupos quedan poco representados y entiende por qué eso puede ser injusto. En esta actividad se Integra pensamiento crítico, convivencia y conocer el lenguaje de IA sin usar pantallas.

Datos de la actividad

- **Curso:** 5º Primaria (10–11 años)
- **Tiempo:** 55–70 min ((recomendable realizar en 2 sesiones))
- **Agrupación:** equipos de 4–5 + puesta en común
- **Espacio:** aula (mesas)

RELACION CURRICULAR

La vinculación curricular de esta actividad, aunque trabajamos de forma desenchufada el funcionamiento de la IA, nos permite realizar la siguiente vinculación curricular:

AREA: **Educación en valores Cívicos y Éticos**

· **CE.EVCE.1**: Deliberar y argumentar sobre problemas de carácter ético referidos a sí mismo y su entorno, buscando y analizando información fiable y generando una actitud reflexiva al respecto, para promover el autoconocimiento y la autonomía moral. *(Deliberar y argumentar sobre problemas éticos; construir posición moral autónoma, incluyendo cuestiones sobre el uso responsable y crítico de medios/redes)*

· **CE.EVCE.2**: Actuar e interactuar de acuerdo con normas y valores cívicos y éticos, reconociendo su importancia para la vida individual y colectiva, y aplicándolos de manera efectiva y argumentada en distintos contextos, para promover una convivencia democrática, justa, respetuosa y pacífica. *(promover convivencia democrática y excluir arbitrariedad, injusticia y discriminación; analizar conflictos también en entornos virtuales y proponer medidas)*.

AREA: **Ciencias de la Naturaleza** (dimensión digital y de datos)

· **CE.CN.1**: Utilizar dispositivos y recursos digitales de forma segura, responsable y eficiente, para buscar información, comunicarse, trabajar de manera individual, en equipo y en red y, para reelaborar y crear contenido digital de acuerdo a las necesidades digitales del contexto educativo.

> Esta actividad curricularmente nos permite en tercer ciclo combinar: el análisis crítico y la deliberación ética en contextos cercanos (IA y decisiones automatizadas) y el uso responsable de información y datos en el apartado digital.

OBJETIVO DIDÁCTICO

Construir y evaluar un modelo sencillo para tomar decisiones, en este caso centrada en la toma de un solo tipo de decisión, usando datos de ejemplo, y comprobar cómo aparece el **sesgo** por representación y por “reglas mal diseñadas”.

DESARROLLO

1) Pregunta detonante

“Si quisiéramos una ‘máquina’ que nos ayudará a decidir nuestro **Robot ayudante del aula**… ¿cómo le enseñaríamos a decidir?”

2) Organización

- **Diseñador/a del modelo** (escribe reglas/puntos)
- **Entrenador/a de datos** (analiza ejemplos)
- **Probador/a** (aplica el modelo a nuevos casos)
- **Auditor/a de sesgo** (revisa a que casos perjudica y por qué)
- (Opcional) **Portavoz**

Materiales
• 30 “fichas de caso” con Robots y rasgos neutros del aula (sin datos sensibles) que queramos que analicen, por ejemplo: respeta turnos, interrumpe explicaciones, cuida el material, etc...

En este caso como ejemplo, los rasgos elegidos serian los siguientes:

POSITIVOS

- Respeta turnos
- Cuida el material
- Trabaja bien en equipo
- Respeta la fila
- Ayuda a ordenar

NEGATIVOS

- Interrumpe
- No cuida el material
- Le cuesta trabajar en equipo
- Olvida los materiales
- Se enfada cuando no gana

A partir de estos datos, que en clase puede ser un **trabajo previo** donde el alumnado seleccione los rasgos que piensan, tanto positivos como negativos, que debería tener el robot ayudante de su clase. Se crean combinaciones para crear las fichas que vamos a utilizar para desarrollar esta actividad.

Como ejemplo tenemos estas fichas "ROBOTS EJEMPLO"

> Podemos determinar el nivel de dificultad de la tarea añadiendo mayor o menor número de rasgos a cada robot

• 3 tipos de sobres:

1. **Fichas de entrenamiento**. estarían ya etiquetadas por el docente como Sí es elegido o NO es elegido.
2. **Fichas de prueba**: Estas fichas irían sin etiqueta para el alumnado. El docente si tendría la solución elegida para comparar los resultados obtenidos por el alumnado.
3. **Fichas difíciles**: Estas tarjetas promueven que se origine debate, como plantea la actividad: se trata de crear casos mixtos como por ejemplos 2 rasgos positivos y uno negativo, 2 negativos y 1 positivo o incluir rasgos no contemplados anteriormente. Aquí es donde puede aparecer el sesgo del modelo.

• Plantilla “Nuestro modelo” (tabla de puntos o reglas)
• Hoja de registro: 12 predicciones señalando CORRECTA/INCORRECTA

• Post-its: DATOS / ALGORITMO / MODELO / SESGO / PRUEBA de esta manera pueden ir identificando en cada momento el momento del proceso en el que se encuentran.

> Nota: se puede utilizar en vez de robots, cualquier tipo de ejemplo que resulte más motivante al grupo como animales, superhéroes, personajes de dibujos, plantas, deportistas, Pokemons,... Imaginación al poder.

3) Datos (entrenamiento)

Cada equipo recibe 10 tarjetas del sobre **ENTRENAMIENTO**, ya etiquetadas por el docente:

El equipo analiza:

- ¿Qué rasgos aparecen mucho en los “SÍ”?
- ¿Qué rasgos aparecen mucho en los “NO”?

> Se nombra: **DATOS = ejemplos con respuesta correcta**.

4) Construcción del algoritmo (modelo por puntos)

El equipo crea un **algoritmo** tipo “puntos” (muy manipulativo y fácil de aplicar)como por ejemplo:

- +2 "respeta los turnos"
- +1 "ayuda a ordenar"
- −2 “interrumpe”
- −1 "olvida los materiales"

Puede ser una decisión de equipo o de clase determinar los valores de cada rasgo. En este caso sería recomendable que fuera una decisión a nivel de clase ya que favorece la comprobación de la tarea de forma igualitaria por todos los grupos.

Finalmente se establece una **Regla final** por ejemplo:

- 3 o más puntos → **SÍ**
- Menos de 3**-> NO**

> Se nombra: ALGORITMO **= las reglas/pasos para sumar y decidir** y **MODELO = la tabla final de puntos (lo aprendido con los datos)**.

5) Prueba (validación) con datos nuevos

Se entrega el sobre **PRUEBA** con 8 tarjetas nuevas (sin solución).
El equipo aplica su modelo, predice SÍ/NO y registra.

Después el docente revela la etiqueta real (o una hoja de soluciones común).

6) Mini-laboratorio de sesgos

Ahora entra en juego el sobre **DIFÍCILES** con 4 tarjetas pensadas para mostrar el sesgo.

Ejemplos:

- Casos con rasgos “mixtos” que no aparecían en entrenamiento (“no sigue las instrucciones” o "escucha con atención")
- Casos con un rasgo que el modelo penaliza demasiado (“lInterrumpe”)

El auditor responde en la hoja:

- ¿Qué tipo de caso recibe más “NO”?
- ¿Ese rasgo estaba poco representado en los datos?
- ¿La regla está exagerando una pista?

> Se nombra: **SESGO = cuando el modelo se equivoca más con ciertos casos porque los datos eran pobres o las reglas no eran justas.**

7) Mejora del modelo

Cada equipo elige UNA mejora (obligatorio justificarla):

- **Mejora de datos:** añadir 2 tarjetas de entrenamiento que faltaban (más variedad de casos)
- **Mejora de reglas:** bajar rasgos de No (−2 → −1), subir un criterio positivo, o añadir “NO SÉ” si está en zona gris
- **Mejora de prueba:** crear una regla de “revisión humana” para casos límite, en este caso la decisión final es tomada por el grupo.

Como ejemplo se puede crear el siguiente **modelo mejorado** teniendo en cuenta solo la puntuación:

- 3 o más puntos ->**SÍ**
- 1 o 2 puntos -> **REVISAR/ NO SÉ o DECISIÓN HUMANA**
- 0 o menos puntos ->**NO**

Se repiten las 4 tarjetas DIFÍCILES y se comparan resultados.

8) Debate final

· "¿Qué mejoró más: cambiar datos o cambiar reglas?"

· "¿Cuándo sería peligroso usar este modelo sin revisar?"

· "¿Qué significa que un modelo sea 'justo'?"

DUA (adaptaciones)

- **Representación:** tarjetas con iconos/pictos (arasaac) y colores por categorías lo que reduce la carga lectora y mejora la comprensión.
- **Acción/expresión:** para construir el modelo por puntos, usad fichas: una ficha = 1 punto (o fichas dobles para +2). Así el alumnado "ve" la suma y puede demostrar el aprendizaje aunque escriba menos; después explican el resultado oralmente o con un mini-esquema "si suma ≥ 3 → SÍ".
-

EVALUACIÓN

A modo de evaluación de la actividad proponemos como ejemplo la siguiente rúbrica.

Indicador (criterios)	1 · Inicial	2 · En proceso	3 · Adecuado	4 · Avanzado
1) Comprensión del funcionamiento de la IA (desenchufada) (CE CN 1)	Confunde datos/reglas/resultados; necesita guía constante.	Identifica partes del proceso con ayuda, con algunos errores.	Explica con claridad: **datos → reglas/modelo → prueba → errores/sesgo**.	Además, relaciona causas del sesgo y anticipa cómo prevenirlo.
2) Uso y organización crítica de datos/información (CE EVCE **1**)	Usa ejemplos/datos sin criterio; no justifica decisiones.	Organiza de forma básica; justifica poco o de manera confusa.	Selecciona y organiza datos con criterio; justifica decisiones de forma clara.	Detecta desequilibrios/ ausencias en los datos y propone cómo corregirlos con buena justificación.
3) Deliberación ética: justicia, igualdad y no discriminación (CE EVCE **1** y **2**)	Opina sin razones; dificultad para escuchar y respetar turnos.	Aporta razones simples; escucha parcialmente; necesita recordatorios.	Argumenta con razones y ejemplos; escucha y dialoga con respeto.	Considera otros argumentos, ajusta su postura y ayuda a construir acuerdos justos.

4) Propuestas de mejora para una IA más justa y convivencia(CE EVCE **2**)	Propone cambios vagos o irrelevantes; no concretas medidas.	Propone una mejora concreta, pero poco viable o poco relacionada con el sesgo.	Propone mejoras **concretas y viables** (mejorar datos, ajustar reglas, revisión humana) y las justifica.	Propone varias mejoras, prevé efectos y define cómo comprobar si el sistema es más justo.

La asociación de los criterios con la actividad seria la siguiente:

· **Registro de entrenamiento/prueba** (Indicadores 1 y 2)

· **Observación del debate** (Indicador 3)

· **Mini-informe “auditor/a de sesgo”**: qué sesgo, a quién afecta, mejora propuesta (Indicador 4)

> **Aprender sobre la IA no es "usar pantallas", sino comprender como decide un sistema basado en datos y que riesgos éticos puede tener.**

Revision #11
Created 2026-01-21 16:37:23 CET by Maria
Updated 2026-03-20 14:54:56 CET by David Cañete