CAPÍTULO I

De la precisión a la responsabilidad: nuevas métricas para la inteligencia artificial

Amparo Alonso Betanzos*

La evaluación de los sistemas de inteligencia artificial ha estado tradicionalmente dominada por métricas de rendimiento técnico, como la precisión (proporción de predicciones positivas que son correctas), el *recall* o sensibilidad (proporción de casos positivos reales correctamente identificados) o el área bajo la curva (AUC, *Area Under Curve*, que mide la capacidad del modelo para discriminar entre clases a distintos umbrales de decisión). Sin embargo, este enfoque resulta hoy claramente insuficiente para capturar los efectos reales de la IA en contextos sociales complejos, como la discriminación de grupos, la falta de confianza de los usuarios o el elevado consumo energético de los algoritmos.

Este capítulo analiza cuatro casos de uso distintos con un objetivo común: mostrar por qué es imprescindible ampliar el marco de evaluación más allá del desempeño predictivo. Esta necesidad no es solo técnica —medir dimensiones como la sostenibilidad, la personalización o el impacto social—, sino también regulatoria. El AI Act europeo introduce la obligación de evaluar riesgos, efectos sociales y posibles daños que las métricas clásicas no reflejan. Sin métricas adecuadas, la propia aplicación de esta regulación se vuelve problemática: aquello que no puede medirse difícilmente puede gobernarse.

La ausencia de métricas integrales conlleva el riesgo de desplegar sistemas técnicamente excelentes pero éticamente deficientes, como modelos de reconocimiento facial con alta precisión y sesgos discriminatorios, o asistentes conversacionales fiables en apariencia pero opacos y propensos a alucinaciones. A través de los cuatro casos estudiados —la

* Este capítulo no hubiese sido posible sin las contribuciones de mis compañeros del grupo de investigación LIDIA (Laboratorio de I+D en IA) del CITIC de la Universidade da Coruña: Jorge Paz Ruza, Bertha Guijarro Berdiñas, Brais Cancela y Carlos Eiras Franco. Asimismo, agradezco la financiación de los proyectos nacionales PID2019-109238GB-C22 y PID2021-128045OA-I00 y de la Xunta de Galicia ED431C 2022/44, con fondos europeos ERDF.

predicción de toxicidad en plataformas digitales, las recomendaciones personalizadas con criterios de sostenibilidad, el uso ético del aprendizaje *Positive Unlabeled* para tratar datos sin etiquetar (cuando solo se conocen ejemplos positivos y el resto de los datos no están etiquetados) y la introducción de una métrica diádica (EAUC, *Excentricity Area Under the Curve*) para capturar sesgos como el de excentricidad—, el capítulo demuestra que lo que no se mide también importa. Justicia, transparencia y sostenibilidad ambiental deben incorporarse explícitamente a la evaluación, como condición necesaria para construir sistemas de IA que generen valor no solo técnico, sino también social.

Palabras clave: métricas en IA, sostenibilidad, sesgos, personalización.

1. INTRODUCCIÓN

La evaluación de modelos de inteligencia artificial (IA) ha estado tradicionalmente centrada en métricas de rendimiento, como la precisión, exactitud, error cuadrático medio, tasa de verdaderos positivos/negativos o área bajo la curva (AUC, que mide la capacidad del modelo para discriminar entre clases a distintos umbrales de decisión), entre otros. Estas métricas siguen siendo útiles, pero en la práctica actual resultan insuficientes. La rápida adopción de sistemas de IA en ámbitos críticos, como la educación, salud, finanzas, interacción social o gobernanza, ha puesto de manifiesto que el rendimiento técnico por sí solo no captura los efectos reales que los modelos producen en los usuarios, las instituciones y la sociedad (como podrían ser la discriminación, la falta de confianza o el consumo energético de los mismos, entre otros). De hecho, varios estudios han mostrado que modelos con altos niveles de precisión pueden simultáneamente perpetuar sesgos (Glickman and Sharot, 2025), producir decisiones opacas (Schilke and Reimann, 2025) o imponer costes ambientales y computacionales desproporcionados (Bolón-Canedo *et al.,* 2024), con importantes repercusiones sociales y económicas. Hoy en día, los modelos de IA no solo predicen, clasifican o recomiendan, sino que también interactúan con personas, influyen en comunidades, redistribuyen visibilidad, condicionan flujos de información y modifican comportamientos. Este cambio de escala y de función revela riesgos que no pueden evaluarse únicamente mediante las tradicionales métricas de rendimiento. Factores como la presencia de sesgos, la opacidad en la toma de decisiones, la posibilidad de generar o amplificar toxicidad, la falta de trazabilidad, la desigualdad en el acceso o las crecientes demandas computacionales y medioambientales son factores que deben integrarse en cualquier marco evaluativo contemporáneo. Paralelamente, el contexto regulatorio refuerza esa necesidad. La reciente aprobación del Reglamento de IA[1] en la Unión Europea exige que los sistemas sean verificables en dimensiones como equidad, transparencia, robustez, seguridad, sostenibilidad y responsabilidad social. Esto necesariamente obliga a ampliar la noción de evaluación para abarcar propiedades que no son estrictamente algorítmicas, sino también sociotécnicas: es decir, propiedades que emergen de la interacción entre modelos, datos, usuarios, instituciones y contextos de aplicación.

En este marco, la llamada IA Responsable (Goellner *et al.*, 2024) avanza hacia enfoques de evaluación multidimensionales que combinan rendimiento con criterios de impacto, sesgo, explicabilidad, responsabilidad y sostenibilidad. Estos nuevos enfoques buscan capturar aspectos tan diversos como la mitigación del daño, la explicabilidad operativa, el comportamiento emergente en sistemas interactivos, la adecuación contextual, la equidad en la distribución de resultados, la eficiencia energética, la reproducibilidad o la gobernanza del ciclo de vida del modelo. Este cambio de paradigma no implica descartar las métricas tradicionales, sino integrarlas en sistemas de evaluación más completos, capaces de reflejar dimensiones técnicas, sociales y éticas. Disponer de métricas

[1] Reglamento de IA: https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

robustas es, por lo tanto, esencial para garantizar que la IA se alinee con los valores y necesidades de la sociedad, minimizando riesgos y maximizando beneficios.

A pesar de estos avances, la brecha persiste. La mayoría de los estudios que proponen métricas para una IA ética se concentran en la equidad (Palumbo *et al.*, 2025) (paridad demográfica, probabilidades igualadas o impacto dispar) y tienden a limitarse en gran medida a entornos de datos tabulares o estructurados, con dificultades para generalizar a arquitecturas multimodales o complejas (por ejemplo, modelos que procesan texto, imágenes o grafos). Otras dimensiones críticas como la explicabilidad/transparencia (Deters *et al.*, 2025), la rendición de cuentas (*accountability*) (Raji *et al.*, 2020), la robustez, la gobernanza de datos, el bienestar social, la sostenibilidad[2] o la confianza del usuario carecen de métricas objetivas, ampliamente aceptadas y operacionalizables.

Esta carencia limita nuestra capacidad de construir sistemas de IA verdaderamente confiables. Sin métricas rigurosas y aplicables para la evaluación de la explicabilidad, el sesgo, la sostenibilidad, la robustez o el impacto social, los desarrolladores y reguladores carecen de herramientas para auditar y garantizar que los sistemas de IA cumplan con los estándares éticos y legales. Es urgente, por tanto, expandir la caja de herramientas de evaluación: definir, validar y adoptar un conjunto de métricas que capturen no solo qué hace un modelo, sino cómo y por qué lo hace, y qué consecuencias sociales puede implicar su comportamiento. Si aspiramos a que la IA no sea solo eficiente, sino también ética, humana y sostenible, debemos pasar de métricas puramente técnicas a sistemas de medición integrales que capturen esos pilares éticos. Este reto constituye tanto un desafío técnico como un imperativo social.

En este capítulo presentamos ejemplos concretos en los que hemos desarrollado o aplicado métricas que traducen principios éticos y legales en medidas cuantificables, útiles para el desarrollo y auditoría de sistemas de IA. En concreto, detallaremos cuatro casos: una medida de toxicidad predictiva para usuarios en redes sociales, un modelo de personalización explicable y sostenible para ser utilizado en un sistema de recomendación, un modelo sostenible que mejora la utilización de datos y la explicabilidad, así como el rendimiento de un sistema, y finalmente proponemos una métrica para sesgos en casos de regresión diádica. Si bien estos ejemplos no constituyen un marco completo multimodal de métricas, representan pasos significativos hacia la construcción de un sistema de evaluación integral para una IA responsable.

2. MODELO DE TOXICIDAD PREDICTIVA PARA USUARIOS DE REDES SOCIALES

Las plataformas digitales, como Reddit, Facebook o X, han intentado frenar la toxicidad en redes sociales utilizando un enfoque reactivo: detectar comentarios ofensivos y eliminarlos, con repercusiones como el bloqueo a los usuarios que los producen. Sin embargo,

[2] https://codecarbon.io/

la moderación tiene sus detractores y durante los últimos años ha habido cambios en el enfoque, en gran medida debido a que consume recursos que las compañías prefieren dedicar a otras cuestiones. Reddit, por ejemplo, en su último informe de transparencia de 2025[3] manifiesta que continúa usando una combinación de herramientas automatizadas y moderación humana para detectar y eliminar contenido que viola sus normas. Meta, por su parte, puso fin a su programa de verificación de hechos mediante terceros (*fact-checking* externo) en sus plataformas, como Facebook o Instagram, sustituyéndolo por un sistema tipo "Community Notes", similar al de X[4]. Esta última plataforma ha recortado sus recursos humanos en moderación, apostando por "moderar con la comunidad"[5]. Tras la eliminación de la moderación reactiva, varios estudios han detectado un aumento de denuncias en las plataformas por acoso, contenido violento y polarización[6].

En cualquier caso, este enfoque reactivo presenta limitaciones evidentes: llega tarde (al no evitar la propagación inicial de contenido nocivo), es costoso para las plataformas, genera desgaste psicológico en los moderadores y contribuye a que los usuarios más tóxicos migren a comunidades cada vez más extremas. En los últimos años, el problema de la toxicidad en redes sociales se ha agravado, en parte por el argumento, cada vez más utilizado por algunas plataformas, de que endurecer la moderación limitaría la "libertad de expresión" o frenaría la innovación. Sin embargo, medir la toxicidad de forma objetiva es muy complejo: depende del contexto cultural, del tono, de la intención y de matices lingüísticos que los sistemas actuales todavía no captan adecuadamente. De hecho, muchos modelos de moderación introducen sesgos sistemáticos que penalizan ciertas lenguas, estilos comunicativos o colectivos vulnerables, contribuyendo a desigualdades que, lejos de resolver el problema, lo amplifican. Además, los usuarios no son uniformes: su comportamiento varía notablemente según el espacio en el que participan, lo que demuestra que la toxicidad no es una propiedad fija del individuo, sino de la interacción entre usuario y comunidad.

Ante esta complejidad, necesitamos métricas transparentes, auditables y adaptables que permitan evaluar el fenómeno sin comprometer derechos ni reproducir discriminaciones. En este contexto, nuestra propuesta plantea un enfoque completamente distinto: predecir la toxicidad antes de que ocurra, bajo la asunción ya mencionada de que esta no depende únicamente ni del usuario ni de la comunidad, sino de la interacción entre ambas cuestiones. En lugar de identificar comentarios tóxicos ya publicados, el método anticipa si un determinado usuario va a comportarse de manera tóxica en una comunidad concreta (Paz-Ruza *et al*., 2025). Si esta predicción es correcta (y en nuestro caso de estudio se muestra que lo es en más del 80 % de las veces), las plataformas podrían evitar interacciones conflictivas antes de que tengan lugar, redu-

[3] https://redditinc.com/policies/transparency-report-january-to-june-2025-reddit

[4] https://techcrunch.com/2025/01/07/meta-drops-fact-checking-and-loosens-its-content-moderation-rules

[5] https://www.socialmediatoday.com/news/x-has-significantly-fewer-moderation-staff/714650/

[6] https://www.europapress.es/portaltic/socialmedia/noticia-meta-reduce-errores-moderacion-cambiar-sistema-aumenta-contenido-violento-facebook-20250530174039.html

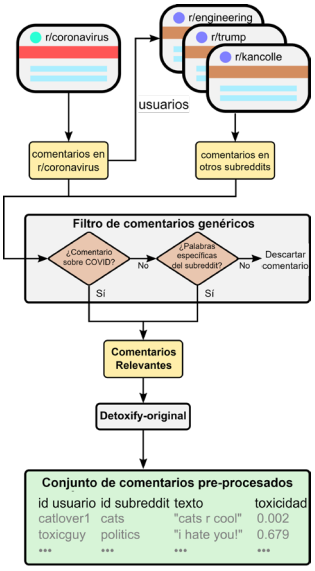
ciendo la exposición a contenido dañino en temas especialmente sensibles como la salud pública, y actuando preventivamente para reducir el riesgo sin recurrir a censuras indiscriminadas ni sobrecargar los sistemas de moderación.

Como caso de uso, hemos elaborado un modelo que hemos probado en la plataforma Reddit sobre comentarios del coronavirus. Para ello, recogemos los comentarios del subreddit coronavirus durante un período coincidente con los inicios de las campañas de vacunación; para cada usuario con comentarios en ese canal, añadimos hasta 1.000 comentarios de otros canales.

Posteriormente preprocesamos el conjunto extraído para, entre otras cuestiones, eliminar comentarios genéricos no útiles, y finalmente obtenemos un conjunto de 620.673 comentarios escritos por 25.893 usuarios en 331 subreddits diferentes. El proceso puede verse en la figura 1. Para cada comentario, nuestro objetivo es predecir su toxicidad, y para ello el método que usamos es el siguiente:

FIGURA 1

PROCESO DE ADQUISICIÓN, FILTRADO Y PREPROCESADO DE DATOS DE COMENTARIOS DE *REDDIT*



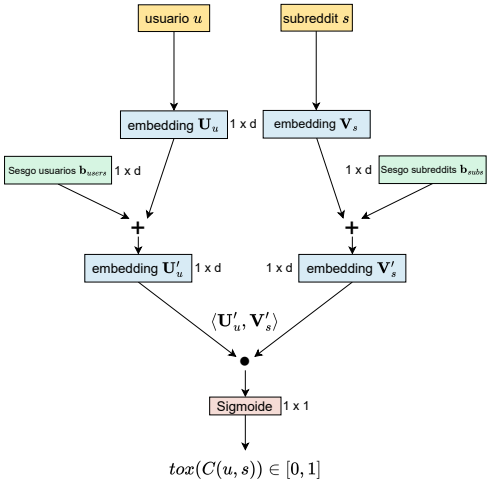
Fuente: Elaboración propia.

- Calculamos un valor [0,1] para la toxicidad utilizando un lenguaje preentrenado, Detoxify original (Hanu and [Unitory team], 2020), para medir la toxicidad de cada comentario.

- Utilizando un modelo de factorización matricial, cuyo esquema se muestra en la figura 2, aprendemos una representación latente de los usuarios y los subreddits, y su interacción predice la probabilidad de toxicidad futura.

FIGURA 2

TOPOLOGÍA DEL MODELO DE APRENDIZAJE AUTOMÁTICO PROPUESTO PARA PREDECIR LA TOXICIDAD DE CONVERSACIONES SOBRE COVID EN INTERACCIONES USUARIO-SUBREDDIT EN LA PLATAFORMA REDDIT. EL MODELO RECIBE LOS IDENTIFICADORES DEL USUARIO $\vec{u}$ Y SUBREDDIT $\vec{s}$, Y PREDICE LA TOXICIDAD ESPERADA $tox(C(\vec{u},\vec{s})) \in [0,1]$ DEL COMPORTAMIENTO DEL USUARIO EN EL SUBREDDIT. d ES EL NÚMERO DE CARACTERÍSTICAS LATENTES UTILIZADAS



Fuente: Elaboración propia.

Para poder predecir la toxicidad futura entre un usuario y un subreddit, es necesario que el modelo haya visto previamente ejemplos de ambos, lo que impide usar particiones tradicionales como la división fija de los datos de entrenamiento y prueba (*holdout*) o la validación cruzada, ya que podrían generar situaciones de ausencia de datos para entrenamiento (*cold start* o problema de inicio en frío). El método habitual en datos diádicos como los que nos ocupan, LOLI (*Leave One Last Item*) reserva para el conjunto de prueba la última interacción temporal (u, s) de cada usuario con dos o más interacciones. Sin embargo, este método es incompatible con nuestro problema, ya que nuestro conjunto de datos no dispone de una única etiqueta temporal por interacción y, además, LOLI está diseñado para tareas evaluadas mediante *rankings*. Por ello, hemos realizado una adaptación novedosa del método LOLI tradicional para tareas de clasificación binaria: seleccionar para cada usuario una de sus interacciones tóxicas y no tóxicas como test, garantizar que cada subreddit tenga ejemplos tanto en entrenamiento como en test y, si se desea, repetir el proceso para obtener un conjunto de validación. De este modo, se construyen particiones equilibradas y realistas para entrenar y evaluar el modelo.

Para evaluar la capacidad predictiva del modelo, utilizamos métricas clásicas de clasificación binaria: sensibilidad, especificidad y la media geométrica (G.Mean) de ambas clases. No se empleó la exactitud, por ser inapropiada en conjuntos desbalanceados como el caso concreto que comentamos, ni el AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*), ya que tiende a ser excesivamente optimista en escenarios muy desequilibrados (Imani *et al.*, 2026) al no reflejar directamente el rendimiento sobre la clase minoritaria. Tampoco se utilizó el F1-Score, definido como la media armónica entre precisión y recall, dado que en escenarios muy desbalanceados puede ser inadecuado al no considerar los verdaderos negativos ni reflejar el impacto global de los errores. La media geométrica resulta más apropiada cuando ambas clases son igualmente relevantes, como ocurre en este caso: tanto bloquear usuarios no tóxicos como permitir la interacción de usuarios tóxicos empeoran la idoneidad de actuación en la plataforma.

Por otra parte, la propuesta introduce un enfoque predictivo novedoso, distinto de los métodos tradicionales basados en detectar y reaccionar, por lo que no existen técnicas directamente comparables en el estado del arte. Por ello, se definieron tres líneas de base de complejidad creciente:

- NON, que predice todas las interacciones como no tóxicas (imitando el comportamiento de no anticipar nada);
- RND, que asigna toxicidad de forma aleatoria según la proporción observada en el conjunto de entrenamiento ($p \approx 0{,}1$);
- USR, que predice la toxicidad de un usuario según la frecuencia de interacciones tóxicas en su historial.

Como podemos ver en la tabla 1, el modelo, que propone un nuevo enfoque predictivo para combatir la toxicidad más eficiente y menos dañino que la moderación reactiva,

TABLA 1

EVALUACIÓN EN EL CONJUNTO DE PRUEBA DE NUESTRO MODELO (MDL), JUNTO CON LOS MODELOS BASE. RESULTADOS DE SENSIBILIDAD (QUÉ PROPORCIÓN DE CASOS TÓXICOS DETECTA EL MODELO), ESPECIFICIDAD (QUÉ PROPORCIÓN DE CASOS NO TÓXICOS CLASIFICA CORRECTAMENTE) Y MEDIA GEOMÉTRICA. LOS MEJORES RESULTADOS SE MARCAN EN AZUL, Y EL SEGUNDO MEJOR RESULTADO ESTÁ EN NEGRITA EN LA TABLA

	Sensibilidad	*Especificidad*	*Media geométrica*
NON	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
RND	0.100 ± 0.011	0.907 ± 0.014	0.301 ± 0.010
USR	**0.241 ± 0.012**	**0.910 ± 0.004**	**0.468 ± 0.010**
MDL	0.849 ± 0.023	0.810 ± 0.017	0.829 ± 0.018

ofrece un rendimiento elevado en un escenario difícil y muy desbalanceado, detectando cuatro veces mejor la toxicidad futura que las estrategias basadas en comportamiento pasado del usuario, evitando el sesgo de predecir todo como el caso general "no tóxico".

Este caso de uso pone de manifiesto que anticipar la toxicidad no es únicamente un problema de modelado, sino también un problema de evaluación adecuada. La capacidad de predecir interacciones tóxicas futuras depende críticamente de cómo construimos los conjuntos de datos, de qué métricas utilizamos y de qué tipo de errores consideramos aceptables. Medir correctamente la toxicidad —y no solo detectarla *a posteriori*— permite diseñar intervenciones más proporcionales, menos punitivas y potencialmente más justas para los usuarios, reduciendo tanto el daño causado por contenidos tóxicos como los efectos colaterales de una moderación excesivamente reactiva. Este ejemplo ilustra, por tanto, que ampliar el marco de evaluación es una condición necesaria para anticipar riesgos sociales en plataformas digitales.

3. MODELO PARA LA EXPLICABILIDAD SOSTENIBLE Y PERSONALIZADA EN SISTEMAS DE RECOMENDACIÓN

La explicabilidad y la sostenibilidad son otros dos de los pilares de la IA responsable. Los sistemas inteligentes deben poder explicar sus razonamientos y recomendaciones para poder crear confianza en los usuarios. Y deben hacerlo en lo posible sin añadir costes computacionales, cada vez más altos dada la creciente complejidad de estos sistemas. En este caso de uso veremos esta problemática dentro de un tipo concreto de sistemas, que son los sistemas recomendadores.

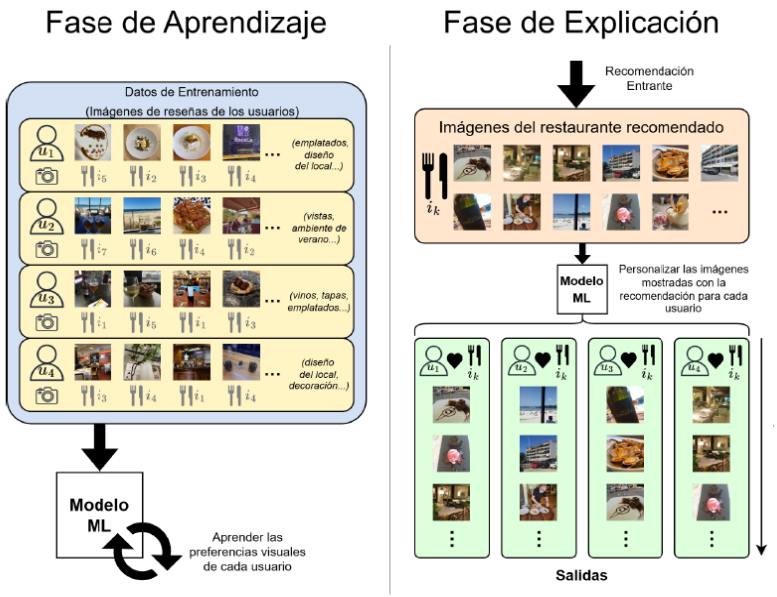
La creciente adopción de sistemas de recomendación en los que se usa IA ha traído enormes beneficios en cuanto a personalización, escalabilidad y eficiencia. Sin embargo, la opacidad de muchos de estos sistemas sigue siendo una preocupación estructural. Por ello, a medida que los recomendadores automatizados influyen en las decisiones cotidianas de personas (qué productos comprar, qué contenidos consumir, qué decisiones culturales o sociales tomar), la necesidad de explicabilidad, trazabilidad y confianza del usuario se vuelve crítica. Esta situación evidencia también la carencia de métricas de transparencia y explicabilidad adaptadas a escenarios reales. Existe, por tanto, una brecha entre lo que optimizan los modelos y aquello que realmente necesitan los usuarios para confiar en un sistema de IA.

En este contexto, nuestro trabajo no introduce una nueva métrica, sino un método que evidencia las limitaciones de las métricas existentes y que, además, muestra que es posible construir sistemas explicables más alineados con el usuario sin aumentar la complejidad del modelo ni el coste computacional y energético. El método propuesto permite generar explicaciones más personalizadas —adaptadas al usuario y al ítem—, ya que los usuarios ya no solo buscan que las recomendaciones que reciben sean personalizadas, sino que también exista cierta explicación que justifique dicha recomendación. Una vía

para lograr esta explicabilidad es el uso de contenido visual, por ejemplo, fotografías de productos, imágenes de lugares, fotos subidas por usuarios, etc., que es bastante común y natural en la percepción humana. Este enfoque natural promueve explicaciones que no son generadas artificialmente ni dependen de metadatos adicionales, lo que puede aumentar la confianza del usuario. La aproximación que aquí interesa es utilizar imágenes subidas a la plataforma en concreto por los propios usuarios como medio de explicación. Este método es intrínsecamente más orgánico y confiable que generar imágenes sintéticas o depender de textos auxiliares, ya que las explicaciones visuales son más representativas de las características del ítem (ver figura 3).

FIGURA 3

ESQUEMA GENERAL DE LA UTILIZACIÓN DE IMÁGENES SUBIDAS POR LOS USUARIOS PARA UNA EXPLICABILIDAD BASADA EN LO VISUAL Y RECOMENDACIONES PERSONALIZADAS. LOS USUARIOS SUBEN IMÁGENES A UNA PLATAFORMA PARA JUSTIFICAR SUS EXPERIENCIAS CON UN ÍTEM. SE CONSIDERA QUE LAS CARACTERÍSTICAS DE ESTAS IMÁGENES CONSTITUYEN BUENAS EXPLICACIONES PARA ESE USUARIO, Y LA IDEA ES QUE UN MODELO DE APRENDIZAJE PUEDE APRENDER ESTAS CARACTERÍSTICAS PARA CADA USUARIO (PARTE IZQUIERDA DE LA FIGURA). LUEGO, LA APARIENCIA DE CUALQUIER RECOMENDACIÓN ENTRANTE PUEDE PERSONALIZARSE (EN ESTE CASO, CON UNA IMAGEN DEL RESTAURANTE RECOMENDADO) UTILIZANDO LA IMAGEN YA EXISTENTE EN LA PLATAFORMA QUE MEJOR REFLEJE LAS PREFERENCIAS EXPLICATIVAS QUE SE HAN APRENDIDO DEL USUARIO; LA IMAGEN ELEGIDA COMO EXPLICACIÓN VARIARÁ ENTRE USUARIOS SEGÚN SUS PREFERENCIAS (LADO DERECHO)



Fuente: Elaboración propia.

Sin embargo, emplear imágenes subidas por el usuario a una plataforma plantea ciertos desafíos importantes, como es el caso de la dispersión de los datos —mientras que las explicaciones textuales pueden ser compartidas por varios usuarios o ítems, las imágenes son únicas para ese par (usuario, ítem)—, lo que hace que el modelo tenga más dificultad para aprender, y además tengamos el problema de que los datos no específicamente subidos por el usuario son normalmente tratados como datos negativos (algo que claramente no es cierto necesariamente, ya que a un usuario determinado puede gustarle una imagen que haya subido otro usuario diferente sobre el mismo ítem o incluso sobre otro diferente). Estas cuestiones tienen además mayor influencia en los modelos que hasta ahora han sido el estado del arte (Paz-Ruza *et al.*, 2022; Diez *et al.*, 2020), en los que el problema se aborda como una tarea de clasificación de autoría, es decir, prediciendo qué imagen de un ítem tiene más probabilidades de haber sido subida por el usuario específico. No obstante, este enfoque de modelar una tarea de *ranking* como una clasificación binaria es subóptimo y conlleva altos costes computacionales y modelos voluminosos. Al centrarse en un *ranking* eficiente y no en modelos complejos de clasificación, nuestro modelo BRIE (Paz-Ruza *et al.*, 2024) consigue reducciones sustanciales de coste computacional y energético, una dimensión que rara vez está integrada en los marcos métricos tradicionales, pese a ser esencial para una IA sostenible (Schwartz *et al.*, 2020; Bolón-Canedo *et al.*, 2024). Este cambio de perspectiva es importante porque revela un desajuste profundo: muchas métricas hoy utilizadas para evaluar explicabilidad no capturan ni la relevancia subjetiva de las explicaciones, ni su adecuación al usuario, ni su impacto en la confianza en el sistema.

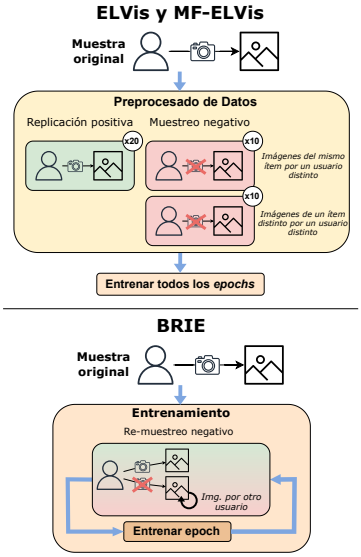
Como mencionamos anteriormente, los enfoques del estado del arte, llamados ELVis (Diez *et al.,* 2020) y MF-ELVis (Paz-Ruza *et al.,* 2022), abordan la selección de explicaciones visuales formulando el problema como una tarea de clasificación binaria. En estos métodos, lo que ocurre en términos generales es que el sistema aprende a distinguir entre explicaciones "buenas" y"malas" para, de forma indirecta, aproximar el problema real de ordenar explicaciones según su adecuación para un usuario concreto. Nuestra propuesta adopta una estrategia más directa basada en aprendizaje por *ranking*, utilizando el algoritmo Bayesian Pairwise Ranking (BPR). Este método está diseñado específicamente para aprender funciones de ordenación, ajustando el modelo para que las explicaciones consideradas como adecuadas aparezcan sistemáticamente por encima de las menos relevantes.

Para ello, BRIE asume que las imágenes subidas por un usuario reflejan de forma consistente los aspectos que este considera relevantes al justificar sus opiniones. Por tanto, todas las imágenes aportadas por un usuario se tratan como ejemplos positivos, bajo la hipótesis de que existe coherencia en los criterios visuales que cada usuario emplea. Como ejemplos negativos, se consideran imágenes no subidas por dicho usuario, una simplificación habitual en escenarios donde no se dispone de ejemplos negativos claramente definidos, un aspecto que en la actualidad estamos mejorando utilizando aproximaciones que nos ayuden a una identificación más precisa de los ejemplos negativos

para cada usuario concreto. BRIE introduce un mecanismo de muestreo más sencillo y flexible: en cada iteración de entrenamiento se generan nuevos ejemplos negativos de forma aleatoria, sin necesidad de preseleccionarlos ni almacenarlos, como se muestra en la figura 4. Esto reduce la complejidad del proceso de aprendizaje y permite un entrenamiento más eficiente, manteniendo al mismo tiempo la capacidad del modelo para aprender preferencias visuales personalizadas.

FIGURA 4

SELECCIÓN DE MUESTRAS NEGATIVAS DE LOS MÉTODOS DEL ESTADO DEL ARTE, LLAMADOS ELVIS Y MF-ELVIS (ARRIBA) Y DEL MÉTODO PROPUESTO, BRIE (ABAJO), USADAS PARA CADA MUESTRA (u, i, p) DEL CONJUNTO DE ENTRENAMIENTO $\mathcal{D}$

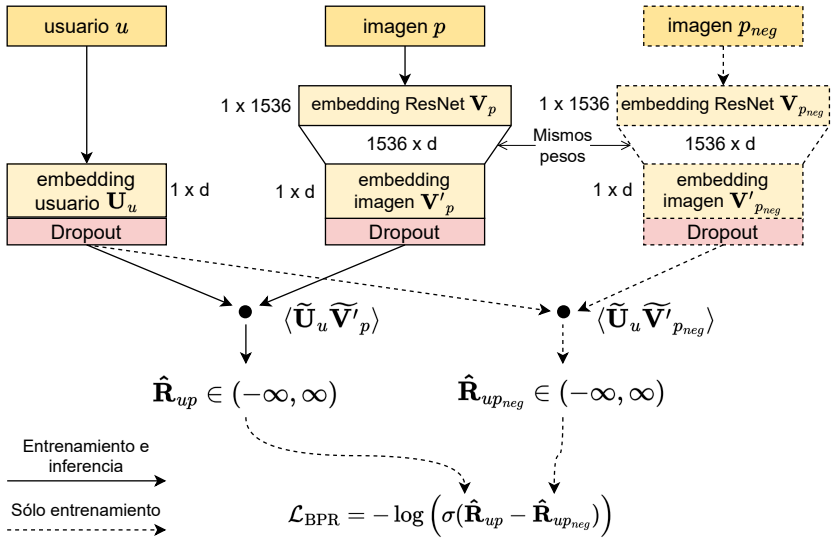


Fuente: Elaboración propia.

Como se muestra en la figura 5, BRIE mantiene la estructura general de modelos anteriores como los ya mencionados ELVis y MF-ELVis, pero introduce mejoras en la forma en que se usan los datos de entrenamiento. El modelo aprende representaciones (*embeddings*) tanto de los usuarios como de las imágenes. Durante el entrenamiento, BRIE recibe un usuario, una imagen subida por ese usuario (considerada adecuada) y otra imagen negativa (considerada poco adecuada). El objetivo es que el sistema aprenda a dar una puntuación más alta a la imagen correcta que a la negativa para ese usuario. En la fase de uso real (inferencia), el modelo solo necesita un usuario y una imagen y devuelve una puntuación que indica lo adecuada que es esa imagen como explicación para dicho usuario. Este enfoque basado en *ranking* simplifica el modelo y evita tratar el problema como una clasificación binaria.

FIGURA 5

TOPOLOGÍA DE BRIE. EN EL PROCESO DE ENTRENAMIENTO, RECIBE EL ID DE USUARIO u, UNA IMAGEN p TOMADA POR u, Y OTRA p_{neg} CONSIDERADA UNA EXPLICACIÓN INADECUADA PARA u, Y DEBE MAXIMIZAR LA DIFERENCIA ENTRE SUS AUTORÍAS PREDICHAS ($\hat{\mathrm{R}}_{up} - \hat{\mathrm{R}}_{up_{neg}}$) PARA MINIMIZAR LA PÉRDIDA $\mathcal{L}_{BPR}$. EN INFERENCIA, RECIBE EL ID DE u Y UNA IMAGEN p, Y CALCULA LA AUTORÍA PREDICHA $\hat{\mathrm{R}}_{up}$, I.E., LA IDONEIDAD DE p PARA EXPLICAR A u UNA RECOMENDACIÓN DEL ÍTEM



Fuente: Elaboración propia.

Para comprobar el rendimiento computacional del modelo, usamos cuatro conjuntos de datos con reseñas de diferentes restaurantes con imágenes tomadas por usuarios de la plataforma TripAdvisor. Los conjuntos de datos pertenecen a distintas ciudades de diferente tamaño u contexto cultural y geográfico (Gijón, Barcelona, Nueva York y Londres) (ver tabla 2), comparando nuestra aproximación con los dos modelos estado del arte, ELVis y MF-ELVis, y con dos modelos de referencia sencillos, el modelo aleatorio RND (Random) —que asigna una preferencia aleatoria a cada par (usuario,imagen)— y un modelo basado en distancias, CNT, que estima la preferencia

TABLA 2

CARACTERÍSTICAS BÁSICAS DE CADA CONJUNTO DE DATOS

Ciudad	*Usuarios*	*Restaurantes*	*Imágenes*	*Imágenes/usuario*
Gijón	5.139	598	18.679	3,64
Nueva York	61.019	7.588	231.141	3,79
Londres	134.816	13.888	479.798	3,56

del par (usuario,imagen) como la distancia negativa entre el embedding ResNetv2 Vp de la imagen p y el centroide de los embeddings ResNetv2 de las imágenes tomadas por el usuario u. Como métricas, usamos Recall@k, NDCG@k y AUC en el *ranking* f(u, i) creado por el modelo para cada caso de prueba, e informamos de las métricas medias (MRecall@k, MNDCG@k, MAUC) de todos los casos de prueba para un conjunto de datos y un modelo (tabla 3).

TABLA 3

RENDIMIENTO DE LOS MODELOS EN CADA CONJUNTO. EL MEJOR Y SEGUNDO MEJOR RESULTADO SE RESALTAN EN NEGRITA Y SUBRAYADO, RESPECTIVAMENTE

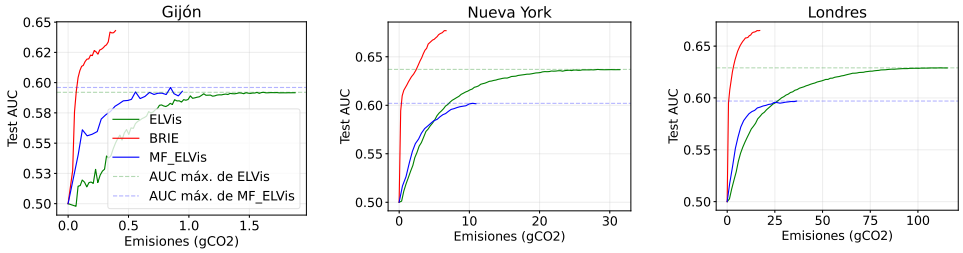
		Métricas		
Ciudad	*Modelo*	*MRecall@10*	*MNDCG@10*	*MAUC*
*Gijón	RND	0.373	0.185	0.487
	CNT	0.464	0.218	0.546
	ELVis	0.521	0.262	**0.596**
	MF-ELVis	**0.538**	**0.285**	0.592
	BRIE	0.607	0.333	0.643
*Nueva York	RND	0.374	0.168	0.502
	CNT	0.431	0.217	0.563
	ELVis	**0.553**	**0.304**	**0.637**
	MF-ELVis	0.516	0.276	0.602
	BRIE	0.598	0.341	0.677
*Londres	RND	0.342	0.155	0.500
	CNT	0.400	0.200	0.562
	ELVis	0.530	**0.293**	**0.629**
	MF-ELVis	**0.531**	0.267	0.597
	BRIE	0.563	0.318	0.665

El modelo BRIE no solo mejora el rendimiento respecto a los métodos anteriores del estado del arte, sino que también reduce de forma significativa el impacto medioambiental. Durante el entrenamiento, BRIE emite aproximadamente un 75 % menos de CO_2 que ELVis y un 50 % menos que MF-ELVis, manteniendo además mejores resultados en todas las métricas evaluadas (Recall, NDCG y AUC) en los tres conjuntos de datos analizados (Gijón, Nueva York y Londres), como vemos en la figura 6.

Los experimentos muestran que BRIE logra un mejor compromiso entre rendimiento y sostenibilidad, de modo que, independientemente del objetivo de reducción de emi-

FIGURA 6

COMPARACIÓN DEL AUC EN TEST VS. EMISIONES DE CO_2 DURANTE EL ENTRENAMIENTO PARA CADA MODELO Y CIUDAD

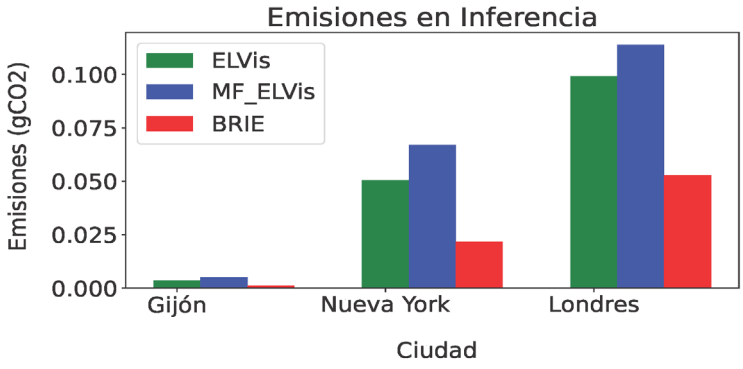


Fuente: Elaboración propia.

siones que se persiga, es posible alcanzarlo sin sacrificar calidad predictiva. En fase de inferencia, BRIE también resulta más eficiente, con una reducción cercana al 50 % en las emisiones de CO_2 frente a ELVis y MF-ELVis en escenarios de uso real, como podemos ver en la figura 7.

FIGURA 7

COMPARACIÓN DE EMISIONES DE CO_2 EN INFERENCIA DE LA PARTICIÓN DE TEST DE CADA CONJUNTO; PROMEDIO SOBRE 50 EJECUCIONES Y CON ACELERACIÓN GPU



Fuente: Elaboración propia.

Esta mejora se debe principalmente a una política de entrenamiento más adecuada y a un uso más eficiente de los datos, evitando arquitecturas innecesariamente complejas. Aunque los experimentos se han realizado con aceleración GPU, lo que mitiga parcialmente el coste de los modelos más grandes, en escenarios basados únicamente en CPU la ventaja energética de BRIE sería aún mayor.

Como vemos, es posible diseñar modelos explicativos eficaces para sistemas de recomendación basados en imágenes de usuarios, que optimicen rendimiento y eficiencia de datos sin sacrificar sostenibilidad. Resaltamos la necesidad de nuevas métricas que evalúen no solo precisión, sino también eficiencia energética y sostenibilidad ambiental de los modelos. Nuestro enfoque demuestra que es posible avanzar en esas direcciones incluso sin métricas nuevas, pero también evidencia la urgencia de desarrollarlas. Sin ellas, la evaluación de sistemas explicables seguirá siendo parcial y no reflejará los valores —personalización, sostenibilidad, confianza— que queremos promover en la IA contemporánea.

4. UN MODELO PARA MEJORAR EL APRENDIZAJE AUTOMÁTICO EN DATOS SIN ETIQUETADO EN PROBLEMAS DE PRIORIZACIÓN GENÉTICA

El tercer caso de uso profundiza en uno de los escenarios más problemáticos y a la vez más habituales para la evaluación de sistemas de IA en biomedicina: aquellos en los que las etiquetas no solo son escasas, sino estructuralmente incompletas. La priorización genética ilustra con claridad cómo los supuestos implícitos de las métricas clásicas fallan cuando la ausencia de evidencia se confunde con evidencia negativa. En estos contextos, evaluar un modelo únicamente en términos de precisión o AUC no solo resulta insuficiente, sino potencialmente engañoso, ya que ignora la incertidumbre inherente al proceso de etiquetado, el coste experimental de los errores y el impacto real de las decisiones de priorización. Este caso pone de relieve que, cuando la IA se utiliza para guiar el descubrimiento científico, las métricas deben capturar no solo la capacidad predictiva del modelo, sino también la calidad del conocimiento generado, la eficiencia computacional y la sostenibilidad del proceso, dimensiones críticas que permanecen invisibles bajo los esquemas de evaluación tradicionales.

La priorización genética constituye una actividad esencial en la investigación biomédica y genómica. Su objetivo es identificar los genes con mayor probabilidad de estar vinculados a un proceso biológico, fenotipo o patología de interés. El empleo de estas metodologías optimiza los esfuerzos del laboratorio, permitiendo que los equipos experimentales se concentren en los candidatos más prometedores, lo cual genera ahorros significativos tanto de tiempo como de recursos económicos y equipamiento. Algunos de los obstáculos con los que se enfrentan los enfoques convencionales de priorización son la escasez y la alta dimensionalidad de los datos genómicos. En este caso de uso, veremos un ejemplo en el que los modelos de aprendizaje automático (AA) han ofrecido una vía prometedora para superar estas dificultades.

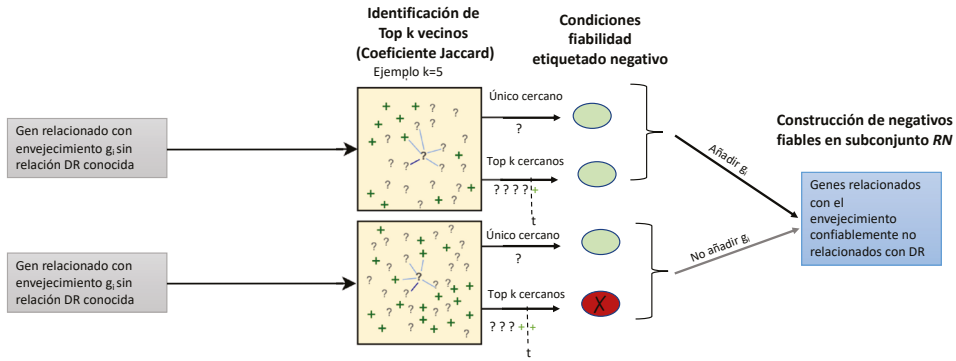
En concreto, nos centraremos en el estudio realizado en (Vega *et al*., 2022), que aplicó AA para hallar genes asociados al envejecimiento que pudieran estar vinculados a la restricción dietética (Dr). La restricción dietética consiste en reducir la ingesta de alimentos sin causar desnutrición y es una de las intervenciones más estudiadas para retrasar

el envejecimiento y mejorar la salud a largo plazo. Se sabe que influye en procesos celulares clave y en numerosos genes, pero identificar exactamente qué genes están implicados es lento y costoso, ya que requiere experimentos de laboratorio complejos. El mencionado estudio utilizó todos los genes sin evidencia conocida de relación con la DR como ejemplos negativos de entrenamiento. Sin embargo, está claro que la carencia de evidencia no equivale a la prueba de la no relación. Esta situación genera un paradigma de datos positivos y sin etiquetar (PU-Positive Unlabeled), donde una parte de los datos tiene una etiqueta positiva confirmada, mientras que el resto (sin etiquetar) se asume como negativo, aunque en realidad algunos podrían ser positivos no detectados. Precisamente, los genes positivos sin etiquetar son el foco de la priorización.

Este trabajo aborda este desafío de datos PU mediante una nueva técnica de aprendizaje PU "basada en distancias" (Paz *et al.*, 2024). El objetivo es perfeccionar los modelos de AA existentes para la priorización genética. El método propuesto, cuyo funcionamiento general se muestra en la figura 8, opera en dos pasos principales:

FIGURA 8

SELECCIÓN DE GENES NEGATIVOS FIABLES BASADA EN SIMILITUD EN EL MÉTODO PROPUESTO



Fuente: Elaboración propia.

- Identificación de negativos confiables: en esta primera etapa, se recurre a un clasificador pseudo-k NN para identificar genes que no están relacionados con la restricción dietética, para ser usados como referencias fiables (ejemplos negativos robustos). Estos son aquellos en los que la mayoría de sus genes vecinos no están relacionados con la DR.

- Entrenamiento del clasificador: Los genes no-DR identificados y los genes DR previamente confirmados se usan para entrenar un ensemble de árboles binarios como modelo de clasificación. De esta forma, el método genera una lista priorizada de nuevos genes candidatos para futuras investigaciones biológicas.

Los resultados son muy positivos: el método no solo mejora la precisión respecto a enfoques anteriores, sino que además reduce el coste computacional y el impacto ambiental, llegando a disminuir las emisiones de CO_2 asociadas al cálculo hasta en un 40 %. Gracias a este enfoque, el estudio identifica varios genes nuevos con un papel potencial en los mecanismos de la restricción dietética, respaldados por evidencias en la literatura científica (tabla 4). Es interesante mencionar que, tras la publicación de nuestro modelo, los genes IRS1 e IRS2 fueron validados como relacionados con la restricción dietética en Drosophila melanogaster.

TABLA 4

GENES CON MAYOR PROBABILIDAD PREDICHA DE RELACIÓN CON LA RESTRICCIÓN DIETÉTICA (DR), COMPARANDO EL MÉTODO EXISTENTE SIN PU (IZQUIERDA) Y EL MÉTODO PROPUESTO CON APRENDIZAJE PU (DERECHA)

Método existente (no-PU)		*Método con aprendizaje PU*	
Gen	**Prob. DR**	**Gen**	**Prob. DR**
GOT2	0.86	*TSC1*	0.97
GOT1	0.85	*GCLM*	0.94
TSC1	0.85	IRS1	0.93
CTH	0.85	PRKAB1	0.92
GCLM	0.82	PRKAB2	0.90
IRS2	0.80	PRKAG1	0.90
SENS2	0.80	*IRS2*	0.90

Este trabajo demuestra cómo una IA mejor diseñada, más consciente de la incertidumbre de los datos, es capaz de mitigar algunas de las limitaciones inherentes a los mismos (naturaleza PU, escasez, alta dimensionalidad) y, por lo tanto, puede convertirse en una aliada clave para la biología del envejecimiento: acelera el descubrimiento científico, reduce costes y hace la investigación más sostenible. Además, logra reducir la complejidad computacional de los métodos previos al mismo tiempo que mantiene su capacidad de explicabilidad.

De forma análoga a los casos de estudio anteriores, este trabajo pone de manifiesto que evaluar sistemas de IA únicamente en términos de precisión o rendimiento predictivo resulta insuficiente. La incorporación de criterios como la calidad del etiquetado, la eficiencia computacional y el impacto medioambiental revela dimensiones del desempeño que permanecen invisibles bajo las métricas tradicionales. En contextos científicos y biomédicos, donde los datos son incompletos, costosos y energéticamente exigentes, se hace imprescindible avanzar hacia nuevas métricas que integren rendimiento, fiabilidad y sostenibilidad, permitiendo comparar modelos no solo por lo que predicen, sino por cómo, con qué coste y con qué impacto lo hacen. Este cambio de perspectiva es clave para una IA verdaderamente responsable y orientada al uso real, especialmente en entornos críticos.

5. MÉTRICA PARA SESGOS EN REGRESIÓN DIÁDICA

El último caso de uso aborda de manera directa uno de los problemas más sutiles y menos visibles de la evaluación en inteligencia artificial: qué ocurre cuando una métrica ampliamente aceptada optimiza el promedio a costa de fallar sistemáticamente en los casos más relevantes.

Definamos primero el contexto en el que trabajaremos: el de los datos diádicos. Los datos diádicos representan información asociada a un par ordenado o no ordenado de entidades (por ejemplo, usuario–usuario, documento–documento, droga–proteína), capturando propiedades de la relación entre ellas más que características aisladas de cada entidad. Ejemplos típicos son las interacciones entre usuarios en redes sociales, las opiniones de usuarios en plataformas como la que vimos anteriormente en TripAdvisor, o las relaciones entre genes y proteínas.

La regresión diádica se refiere a tareas de predicción en las que la variable objetivo está definida sobre pares de entidades y se estima a partir de atributos de cada entidad y/o de su relación. En este tipo de problemas, donde las predicciones se asocian a pares de entidades y tienen consecuencias directas en ámbitos sensibles, las métricas de error global han funcionado durante años como un estándar incuestionado. Sin embargo, este caso muestra que dichas métricas no solo son insuficientes, sino que pueden ocultar comportamientos profundamente problemáticos desde el punto de vista de la equidad, el riesgo y el impacto social. A diferencia de los casos anteriores, aquí no se trata únicamente de qué se mide, sino de qué sesgos estructurales permanecen invisibles cuando no disponemos de la métrica adecuada.

Los modelos de regresión diádica, diseñados para predecir valores reales asociados a pares de entidades, constituyen un componente fundamental de numerosos sistemas de IA actuales. Su aplicación más conocida se encuentra en los sistemas de recomendación (de los que ya hemos visto un caso de uso anteriormente), donde se utilizan para estimar valoraciones o preferencias usuario–ítem, pero su alcance se ha extendido progresivamente a dominios de alto impacto, como la farmacología personalizada, la biomedicina o la modelización de fenómenos económicos y financieros. En estos contextos, las predicciones generadas por los modelos no solo informan decisiones, sino que pueden tener consecuencias directas sobre personas, tratamientos o recursos, lo que hace especialmente relevante analizar cómo se comportan dichos modelos.

A pesar de la complejidad inherente a los datos diádicos —caracterizados por su alta dispersión, distribuciones locales heterogéneas y problemas como el comienzo en frío—, la evaluación de los modelos que operan sobre ellos se ha apoyado históricamente, como ya explicamos anteriormente, en métricas de error global, principalmente RMSE y MAE. Estas métricas, ampliamente consolidadas tras su adopción en hitos como el Netflix Prize (Bennett and Lanning, 2007), ofrecen una visión agregada

del error medio del sistema y han servido como estándar de comparación durante años. Sin embargo, esta perspectiva global oculta aspectos esenciales del comportamiento del modelo cuando se analizan las predicciones a nivel de usuario, ítem o par concreto (Chen, 2017). Los modelos de regresión diádica de última generación tienden a sesgar sus predicciones hacia los valores medios observados de las entidades implicadas, un fenómeno que denominamos sesgo de excentricidad (Paz-Ruza *et al.*, 2025). Este comportamiento permite reducir el error global, pero lo hace a costa de degradar severamente la calidad predictiva en los casos menos frecuentes o más atípicos, que son precisamente los más relevantes en escenarios sensibles desde el punto de vista ético o social. En situaciones extremas, el modelo puede llegar a ofrecer un rendimiento peor que el aleatorio en estos casos críticos, sin que las métricas tradicionales sean capaces de detectarlo.

Nuestros experimentos en ámbitos como la recomendación de contenidos, la predicción de respuesta a fármacos o la estimación de valor económico ponen de manifiesto que RMSE y MAE son insuficientes para revelar estos sesgos estructurales, ya que no distinguen entre errores cometidos en situaciones triviales y errores en predicciones excéntricas con alto impacto potencial. Para abordar esta limitación, proponemos una nueva medida, que llamamos EAUC (*Eccentricity Area Under the Curve*), una métrica diseñada específicamente para capturar cómo varía el error del modelo en función de la excentricidad de los ejemplos, permitiendo así detectar y cuantificar de forma fiable la injusticia algorítmica inducida por el sesgo de excentricidad.

Brevemente, nuestro trabajo, que puede leerse en detalle en (Paz-Ruza *et al.*, 2025), realiza las aportaciones siguientes, que pueden verse resumidas en la figura 9.

- Mostramos que los modelos actuales para regresión diádica sesgan sus predicciones al valor medio observado de las interacciones de cada usuario e ítem, una cuestión que puede ser grave especialmente en entornos críticos. A este fenómeno le llamamos *sesgo de excentricidad*.

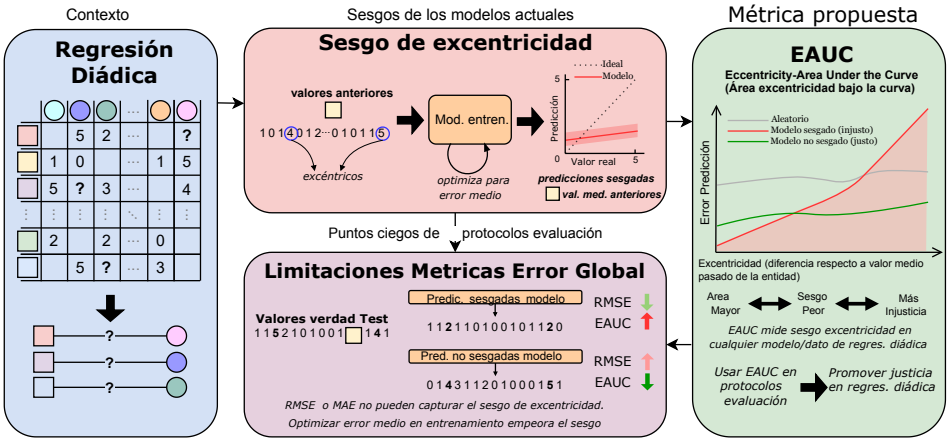
- Demostramos que este sesgo provoca injusticia algorítmica y riesgos en el uso de estos sistemas en tres ámbitos de uso críticos concretos: recomendación de contenidos, farmacología y biomedicina, y mercado y finanzas.

- Demostramos que las medidas de error clásicas, como RMSE o MAE, no detectan estos sesgos en conjuntos de datos reales en los ámbitos anteriores.

- Proponemos una nueva métrica de evaluación, EAUC (*Eccentricity-Area under the Curve*) y un protocolo de evaluación basado en el *sesgo de excentricidad* que permite de forma fiable detectar y cuantificar dicho sesgo en todos los modelos y conjuntos de datos estudiados.

FIGURA 9

RESUMEN DE LAS CONTRIBUCIONES DEL TRABAJO, DE IZQUIERDA A DERECHA: QUÉ ES LA REGRESIÓN DIÁDICA, DEFINICIÓN DEL *SESGO DE EXCENTRICIDAD* EN LA REGRESIÓN DIÁDICA, LAS LIMITACIONES DE LAS MÉTRICAS DE ERROR GLOBALES COMO RMSE O MAE PARA EVALUAR DE MANERA INTEGRAL DICHAS TAREAS, Y LA PROPUESTA DE EAUC COMO UNA MÉTRICA NOVEDOSA QUE PUEDE CUANTIFICAR EL SESGO DE EXCENTRICIDAD (Y LA CONSIGUIENTE INJUSTICIA ALGORÍTMICA) QUE SUFRE UN MODELO DE REGRESIÓN DIÁDICA



Fuente: Elaboración propia.

- Finalmente, proponemos técnicas para prever la tendencia de cualquier conjunto de datos de regresión diádica a inducir este sesgo en modelos entrenados sobre el mismo, así como potenciales aproximaciones para mitigarlo en el futuro.

Este último caso de estudio refuerza una idea central que atraviesa todos los ejemplos analizados en este trabajo: la evaluación de sistemas de IA no puede seguir basándose exclusivamente en métricas de rendimiento promedio. En tareas complejas y de alto impacto, resulta imprescindible avanzar hacia nuevas métricas que incorporen dimensiones como el sesgo, la equidad y el comportamiento del modelo en casos críticos, complementando —y en algunos casos cuestionando— los criterios tradicionales de optimización. Solo mediante este cambio de enfoque será posible desarrollar sistemas de inteligencia artificial realmente fiables, justos y alineados con los requisitos éticos y sociales de sus contextos de aplicación. Este caso de estudio ilustra con especial claridad una conclusión clave de este capítulo: las métricas no son neutrales y elegir mal cómo evaluamos un sistema puede legitimar comportamientos algorítmicos injustos bajo una apariencia de buen rendimiento. La introducción de EAUC demuestra que es posible diseñar métricas que hagan visibles sesgos sistemáticos ignorados por los indicadores clásicos, permitiendo una evaluación más alineada con los riesgos reales de aplicación.

En el contexto del AI Act y de la creciente exigencia de análisis de impacto y equidad, este tipo de métricas no debe entenderse como un complemento opcional, sino como una condición necesaria para una IA responsable. Evaluar no es solo medir el error medio, sino comprender quién falla el sistema, cuándo falla y con qué consecuencias, especialmente en aquellos casos que más importan.

6. CONCLUSIONES

En conjunto, los cuatro casos analizados ponen de manifiesto una idea fundamental: la inteligencia artificial no es neutral y no puede serlo. Los datos que alimentan los modelos reflejan el mundo real, con todas sus desigualdades, asimetrías y sesgos, y los sistemas de IA encarnan necesariamente las decisiones, prioridades e intereses de las organizaciones que los diseñan y despliegan. En este contexto, una evaluación basada casi exclusivamente en métricas de rendimiento técnico —como la precisión, la exactitud o el error medio— resulta claramente insuficiente para comprender el impacto real de estos sistemas sobre las personas, la sociedad y el entorno. Aquello que no medimos no desaparece: la toxicidad en plataformas digitales, los sesgos inducidos por datos incompletos o desbalanceados, la opacidad en la toma de decisiones automatizadas o los costes energéticos y materiales asociados al entrenamiento y uso de modelos de IA continúan existiendo, aunque permanezcan invisibles para las métricas tradicionales. Esta invisibilidad no es inocua: puede traducirse en daños sociales, discriminación, pérdida de confianza o impactos ambientales significativos. Por ello, se hace imprescindible avanzar hacia nuevas métricas que permitan una evaluación integral y responsable de la IA, incorporando dimensiones como la equidad, la explicabilidad, la trazabilidad de las decisiones, la sostenibilidad y la anticipación de riesgos.

La ausencia de estándares claros y, en lo posible, consolidados, en estas dimensiones constituye un desafío adicional que no es únicamente técnico, sino también ético, social y político. Diseñar y desplegar sistemas de IA justos y sostenibles requiere, por tanto, no solo innovación metodológica y tecnológica, sino un compromiso deliberado y coordinado de toda la comunidad: investigadores, reguladores y sociedad deben colaborar para que la IA deje de ser una "caja negra" optimizada únicamente para el rendimiento, y se convierta en una tecnología alineada con valores humanos y orientada al bien común.

Referencias

Bennett, J., and Lanning, S. (2007). The Netflix prize. In Proc. KDD Cup Workshop, 35.

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., and Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096.

Chen, M. (2017). Performance evaluation of recommender systems. *Int. J. Performability Eng.*, *vol. 13*, no. 8, 1246.

Deters, H., Droste, J., Obaidi, M., and Schneider, K. (2025). Exploring the means to measure explainability: Metrics, heuristics and questionnaires. *Information and Software Technology*, 181, 107682, https://doi.org/10.1016/j.infsof.2025.107682.

Diez, J., Perez-Nunez, P., Luaces, O., Remeseiro, B., and Bahamonde, A. (2020). Towards explainable personalized recommendations by learning from users' photos. *Inform. Sci.*, 520, 416-430.

Glickman, M., Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional, and social judgements. *Nat Hum Behav,* 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2.

Goellner, S., Tropmann-Frick, M., and Brumen, B. (2024). Responsible Artificial Intelligence: A Structured Literature Review. ArXiv, abs/2403.06910.

Hanu, L. and [Unitary team]. (2020). Detoxify. Howpulished= github. https://github.com/unitaryai/detoxify

Imani, M., Joudaki, M., Bagheri, A., and Arabnia, H. R. (2026). Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers. *Technologies*, 14, 54. https://doi.org/10.3390/technologies14010054

Palumbo, G., Carneiro, D., and Alves, V. (2025). Objective metrics for ethical AI: a systematic literature review. *Int J. Data Sci. Anal.,* 20, 247–267. https://doi.org/10.1007/s41060-024-00541-w.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., and Eiras-Franco, C. (2024). Sustainable transparency on recommender systems: Bayesian ranking of images for explainability. *Information Fusion*, Volume 111, 102497. https://doi.org/10.1016/j.inffus.2024.102497.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., and Eiras-Franco, C. (2025). Beyond RMSE and MAE: Introducing EAUC to Unmask Hidden Bias and Unfairness in Dyadic Regression Models. *IEEE Transactions on Neural Networks and Learning Systems*, *vol. 36*, no. 11, 19619-19630, Nov. doi: 10.1109/TNNLS.2025.3593059.

Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Eiras-Franco, C. (2025). Predictively combating toxicity in Health-related online discussions through Machine Learning. Proc. International Joint Conference on Neural Networks, IJCNN 2025. DOI: 10.1109/IJCNN64981.2025.11228479.

Paz-Ruza, J., Eiras-Franco, C., Guijarro-Berdiñas, B., and Alonso-Betanzos, A. (2022). Sustainable Personalisation and Explainability in Dyadic Data Systems. *Procedia Computer Science*, Volume 207, 1017-1026, 2022. https://doi.org/10.1016/j.procs.2022.09.157.

Paz-Ruza,J., Freitas, A. A., Alonso-Betanzos, A., and Guijarro-Berdiñas, B. (2024). Positive-Unlabelled learning for identifying new candidate Dietary Restriction-related genes among ageing-related genes. *Computers in Biology and Medicine*, 180, 108999. https://doi.org/10.1016/j.compbiomed.2024.108999.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency , FAT* '20, (New York, NY, USA)*, 33–44, Association for Computing Machinery.

Schilke, O., and Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. https://doi.org/10.1016/j.obhdp.2025.104405.

Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. (2020). Green AI. *Commun. ACM, 63*(12), 54–63.

Vega, M., Gustavo, D., Bespalov, V., Zheng, Y., Freitas, A., and De Magalhaes, J. P. (2022). Machine learning-based predictions of dietary restriction associations across ageing-related genes. *BMC bioinformatics*, 23, 1–28.