Un marco de evaluación Green AI para sistemas de recomendación basado en rendimiento y huella medioambiental

Jorge Paz-Ruza, Xesús Patiño,
Bertha Guijarro Berdiñas, and Amparo Alonso Betanzos

Universidade da Coruña, Grupo LIDIA - CITIC
{j.ruza, x.patinom, amparo.alonso.betanzos, berta.guijarro}@udc.es

Resumen La creciente complejidad de los sistemas de recomendación basados en aprendizaje automático ha incrementado de forma significativa su coste computacional y su huella de carbono, generando la necesidad de enfoques que integren simultáneamente rendimiento y sostenibilidad. En este trabajo se propone una metodología experimental unificada para la evaluación de sistemas de recomendación que incorpora de forma explícita métricas de calidad predictiva y emisiones de CO_2 durante el entrenamiento. La propuesta define tres protocolos de evaluación complementarios (predicción de rating, entrenamiento basado en historial y recomendación Top-K) permitiendo analizar de manera consistente el compromiso entre precisión y eficiencia energética en distintos modelos representativos. Los resultados muestran que modelos más simples pueden competir con arquitecturas complejas en términos de rendimiento, ofreciendo en muchos casos una mejor relación entre precisión y coste ambiental, lo que refuerza la importancia de incorporar criterios de sostenibilidad en el diseño de sistemas de recomendación.

1. Introducción

En la última década, los sistemas de recomendación se han consolidado como herramientas fundamentales en la sociedad digital moderna. Su capacidad para filtrar grandes volúmenes de información y ofrecer sugerencias personalizadas a cada usuario ha transformado sectores estratégicos como el comercio electrónico, el entretenimiento, el turismo o los servicios financieros. Al adaptar los contenidos a las preferencias individuales, estas plataformas no solo mejoran la satisfacción del consumidor, sino que también optimizan los modelos de negocio empresariales [14], de forma que compañías como Netflix, Amazon, Booking o Spotify han declarado que la mayoría de las interacciones en sus plataformas, así como los incrementos en sus ventas, provienen de recomendaciones personalizadas. Sin embargo, el éxito de estos sistemas viene de la mano de una creciente sofisticación, lo que presenta una contrapartida crítica. El entrenamiento y despliegue de modelos cada vez más complejos, desde técnicas clásicas de factorización matricial hasta redes neuronales profundas o arquitecturas basadas en grafos, conllevan un consumo

masivo de recursos computacionales. Este incremento en la demanda de hardware especializado (como GPUs y CPUs de alto rendimiento) se traduce directamente en un elevado gasto energético y, por consiguiente, en una preocupante huella de carbono debido a las emisiones de gases de efecto invernadero asociadas [3]. Tradicionalmente, la investigación en el ámbito de los sistemas de recomendación se ha centrado casi exclusivamente en maximizar la precisión predictiva y la calidad de las clasificaciones. La precisión y el ranking han sido las métricas que determinaban si un modelo era adecuado. Pero la crisis climática y el marco regulatorio europeo en materia de sostenibilidad tecnológica obligan a ampliar ese foco. El enfoque de IA verde promueve que el diseño y la evaluación de los algoritmos no solo consideren su eficacia técnica, sino también su eficiencia energética y su impacto medioambiental [8]. Varios estudios [1,18] sugieren que es urgente transicionar hacia una evaluación multidimensional del aprendizaje automático, donde la eficiencia energética deje de ser una consideración secundaria y se convierta en un parámetro de diseño fundamental. En el caso específico de los sistemas de recomendación, esta necesidad es aún más acuciante debido a la naturaleza dinámica de sus datos, que exige reentrenamientos constantes y un consumo de recursos ininterrumpido. Por ello, el análisis analítico de la sostenibilidad algorítmica no solo permite cuantificar el coste ecológico real de la personalización, sino que también sienta las bases para identificar arquitecturas de compromiso (trade-off) capaces de mitigar el impacto medioambiental sin sacrificar de forma significativa la experiencia del usuario.

2. Antecedentes

A medida que la disciplina ha evolucionado hacia la incorporación de modelos basados en aprendizaje profundo y redes de grafos, se ha agudizado la importancia de estudiar la gestión responsable de los recursos computacionales y el impacto ecológico de dichos algoritmos, y en este contexto existen ya algunos trabajos que realizan diferentes propuestas orientadas a optimizar la precisión, la eficiencia y la adaptabilidad algorítmica de los sistemas de recomendación en múltiples dominios. En [17] se presenta un estudio comparativo de los sistemas presentados en las ediciones de 2013 y 2023 de la conferencia especializada en sistemas de recomendación ACM RecSys [1]. Las conclusiones del estudio revelan un panorama crítico: los algoritmos modernos basados en aprendizaje profundo generan hasta 42 veces más emisiones de CO_2 que las soluciones de hace una década, requiriendo un promedio de 3.297 kg de CO_2 por cada trabajo científico desarrollado (cifra equivalente a la huella de un vuelo de Nueva York a Melbourne). Y lo que es aún más preocupante de este incremento exponencial es que las mejoras en precisión no siempre son proporcionales, dado que estos modelos consumen unas ocho veces más energía que los tradicionales sin garantizar saltos significativos en métricas de ordenación clave, como el NDCG@10. Asimismo, su estudio subraya que factores como la ubicación geográfica de la infraestructura (y la consecuente intensidad de carbono de la red eléctrica local) o el tipo de hardware (GPU frente a CPU)

[1] https://recsys.acm.org/

condicionan de forma sustancial el consumo energético final. En aspectos más generales relativos al desarrollo y despliegue de modelos de creciente complejidad en el campo del aprendizaje automático, existen otros estudios, como [4], que alcanzan las mismas conclusiones. En general, la literatura científica actual coincide en señalar que todavía existe una preocupante carencia de estandarización en las métricas empleadas, así como una falta de estudios empíricos integrales aplicados a la casuística de los sistemas de recomendación industriales. En este trabajo se propone una metodología de evaluación experimental de la sostenibilidad que se aplica a ocho arquitecturas de recomendación pertenecientes a tres familias distintas (factorización matricial, redes profundas y modelos de grafos) bajo tres tipos de entrenamiento diferentes (completo, basado en historial y optimizado para ranking en Top-K). Se integran métricas de evaluación de rendimiento tradicionales (error cuadrático medio raíz-RMSE-, ganancia acumulada descontada normalizada -NDCG-, precisión o recall) y métricas de sostenibilidad, como el consumo energético, el uso de recursos computacionales y la estimación de emisiones de CO_2 sobre varios conjuntos de datos multidominio (entretenimiento, finanzas y farmacología). De esta manera, este estudio no solo cuantifica la huella ecológica, sino que también pretende dotar al desarrollador de una guía empírica para identificar configuraciones que minimicen el coste ambiental sin devaluar la calidad de la experiencia personalizada del usuario.

3. Metodología Propuesta

El objetivo de este trabajo es definir una metodología de evaluación que permita analizar conjuntamente la calidad predictiva, la eficiencia computacional y el impacto ambiental de sistemas de recomendación entrenados bajo diferentes paradigmas de aprendizaje. Mientras que la evaluación tradicional de recomendadores suele centrarse exclusivamente en métricas de precisión o ranking, nuestra propuesta incorpora de forma explícita medidas de sostenibilidad que permiten cuantificar el coste computacional necesario para alcanzar un determinado nivel de rendimiento. Sea

$$\mathcal{D} = \{(u, i, r_{ui})\}$$

el conjunto de interacciones observadas entre usuarios $u \in U$ e ítems $i \in I$, donde r_{ui} representa una valoración explícita o una señal implícita de preferencia. Dado un sistema de recomendación parametrizado por θ, el objetivo consiste en aprender una función capaz de modelar las preferencias de los usuarios para producir predicciones o rankings personalizados.

La metodología propuesta estudia tres paradigmas de entrenamiento ampliamente utilizados en sistemas de recomendación: entrenamiento *pointwise*, entrenamiento basado en historial y entrenamiento orientado a ranking Top-K. Cada paradigma induce un comportamiento diferente tanto en la calidad de las recomendaciones como en el consumo de recursos computacionales. La Figura 1 muestra una representación conceptual de los tres enfoques considerados.



Figura 1: Representación conceptual de los tres paradigmas de entrenamiento considerados en la metodología propuesta.

3.1. Entrenamiento Pointwise

El primer paradigma corresponde al enfoque clásico de predicción de interacciones usuario–ítem. Cada instancia de entrenamiento está formada por un par (u, i) y su valoración asociada r_{ui}. El modelo aprende una función

$$\hat{r}_{ui} = f(u, i; \theta),$$

cuyos parámetros se optimizan minimizando una función de pérdida de regresión. Habitualmente se emplea el error cuadrático medio,

$$\mathcal{L}_{point} = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} (r_{ui} - \hat{r}_{ui})^2.$$

Este paradigma está orientado a maximizar la precisión de las predicciones individuales y constituye el esquema de entrenamiento predominante en tareas de predicción de ratings.

Algorithm 1 Entrenamiento Pointwise

Require: Dataset de entrenamiento $\mathcal{D}$
Require: Modelo f_θ
for $e = 1, \ldots, E$ **do**
  for $B \subset \mathcal{D}$ **do**
    $\hat{r}_{ui} \leftarrow f_\theta(u, i)$ ▷ $\forall (u, i, r_{ui}) \in B$
    $\mathcal{L} \leftarrow \frac{1}{|B|} \sum_{(u,i)\in B} (r_{ui} - \hat{r}_{ui})^2$ ▷ Pérdida pointwise
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  end for
  $\mathcal{M}_e \leftarrow Evaluate(f_\theta, \mathcal{D}_{val})$
  $S_e \leftarrow MeasureResources()$
end for
return $f_\theta, \{\mathcal{M}_e, S_e\}$

3.2. Entrenamiento Basado en Historial

El segundo paradigma aproxima de forma más realista el funcionamiento de un sistema de recomendación desplegado. En lugar de procesar interacciones aisladas, el modelo utiliza el historial completo de cada usuario para inferir sus preferencias futuras. Sea H_u el conjunto de interacciones históricas asociadas al usuario u. El modelo aprende una función

$$\hat{\mathbf{r}}_u = g(H_u; \theta),$$

capaz de generar simultáneamente predicciones para un conjunto amplio de ítems candidatos. La optimización se realiza comparando las preferencias observadas con las preferencias estimadas para cada usuario.

Debido a que el número de ítems disponibles puede ser muy elevado, resulta habitual emplear técnicas de *negative sampling*, incorporando únicamente un subconjunto representativo de ejemplos negativos durante el entrenamiento. Este paradigma permite evaluar el comportamiento de los modelos en escenarios más próximos a las aplicaciones reales, aunque normalmente implica un mayor coste computacional.

Algorithm 2 Entrenamiento Basado en Historial

Require: Historiales $\{H_u\}_{u \in U}$
Require: Modelo g_θ
for $e = 1, \ldots, E$ **do**
  for $u \in U$ **do**
    $\mathcal{P}_u \leftarrow H_u$ ▷ Interacciones observadas
    $\mathcal{N}_u \leftarrow SampleNegatives(u, k)$ ▷ Muestreo negativo
    $\mathcal{C}_u \leftarrow \mathcal{P}_u \cup \mathcal{N}_u$
    $\hat{\mathbf{r}}_u \leftarrow g_\theta(H_u, \mathcal{C}_u)$
    $\mathcal{L} \leftarrow Loss(\mathbf{r}_u, \hat{\mathbf{r}}_u)$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  end for
  $\mathcal{M}_e \leftarrow Evaluate(g_\theta, \mathcal{D}_{val})$
  $S_e \leftarrow MeasureResources()$
end for
return $g_\theta, \{\mathcal{M}_e, S_e\}$

3.3. Entrenamiento Orientado a Ranking Top-K

El tercer paradigma considera la recomendación como un problema de ordenación. En este caso, el objetivo no consiste en estimar valoraciones exactas, sino en garantizar que los ítems más relevantes aparezcan en las primeras posiciones de la lista recomendada.

Sea $s(u, i)$ una función de puntuación que mide la relevancia del ítem i para el usuario u. Dados un ítem relevante i^+ y un ítem no relevante i^-, el modelo aprende una ordenación tal que

$$s(u, i^+) > s(u, i^-).$$

Para ello se emplean funciones de pérdida orientadas a ranking que aproximan el comportamiento de métricas utilizadas habitualmente en recomendación, como Recall@K o NDCG@K. Este paradigma se aproxima al escenario operativo de muchos sistemas de recomendación industriales, donde la calidad de la ordenación resulta más importante que la predicción exacta de una valoración.

Algorithm 3 Entrenamiento Orientado a Ranking Top-K

Require: Conjunto de interacciones $\mathcal{D}$
Require: Modelo de ranking s_θ
for $e = 1, \ldots, E$ **do**
  for $u \in U$ **do**
    $P_u \leftarrow PositiveItems(u)$
    $N_u \leftarrow SampleNegatives(u, k)$
    for $i^+ \in P_u,\ i^- \in N_u$ **do**
      $y^+ \leftarrow s_\theta(u, i^+)$
      $y^- \leftarrow s_\theta(u, i^-)$
      $\mathcal{L} \leftarrow \mathcal{L} + \ell(y^+, y^-)$ ▷ Pérdida de ranking
    end for
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  end for
  $\mathcal{M}_e \leftarrow EvaluateRanking(s_\theta, \mathcal{D}_{val})$
  $S_e \leftarrow MeasureResources()$
end for
return Top-K recommendations

3.4. Evaluación de Sostenibilidad

Los tres paradigmas anteriores se evalúan bajo un protocolo común de monitorización y análisis de sostenibilidad. Durante cada experimento se registran tanto métricas de calidad de recomendación como indicadores de consumo de recursos computacionales.

Formalmente, cada configuración experimental queda caracterizada por un vector de evaluación

$$\mathbf{v} = (Q, S),$$

donde Q representa el conjunto de métricas de calidad del recomendador y S el conjunto de métricas de sostenibilidad. Entre estas últimas se incluyen el tiempo de entrenamiento, el consumo de memoria, la utilización del procesador y

las emisiones equivalentes de dióxido de carbono, comúnmente utilizadas en el ámbito [1].

La monitorización se realiza durante todo el proceso de aprendizaje, registrando tanto valores acumulados como medidas por época. Esto permite analizar la evolución temporal del coste computacional y estudiar posibles compromisos (*trade-offs*) entre rendimiento predictivo e impacto ambiental.

De este modo, la metodología propuesta proporciona un marco unificado para comparar distintas estrategias de entrenamiento desde una perspectiva de IA sostenible, permitiendo identificar configuraciones capaces de ofrecer recomendaciones de alta calidad con un coste computacional y ambiental reducido.

4. Configuración Experimental

Esta sección contiene los detalles de la batería de experimentos diseñada, que combina múltiples dominios de aplicación y arquitecturas de recomendación para estudiar la influencia de distintos factores sobre el consumo de recursos y las emisiones asociadas al entrenamiento e inferencia de los RS.

4.1. Conjuntos de Datos

La evaluación experimental se realizó sobre seis conjuntos de datos pertenecientes a tres dominios distintos:

- Recomendación de productos: se utilizaron los conjuntos *MovieLens 1M* [5] y *Netflix Prize* [13], ampliamente utilizados como referencia en sistemas de recomendación. Ambos contienen valoraciones explícitas de usuarios sobre ítems de tipo película.
- Farmacología: se emplearon los conjuntos *GDSC1* [21] y *CTRPv2* [2], donde las interacciones representan respuestas farmacológicas entre líneas celulares y compuestos químicos.
- Finanzas y economía: se incluyeron los conjuntos *Kiva Microloans* [10] e *IMF Direction of Trade Statistics (DOTS)* [9], pertenecientes al ámbito financiero, donde las interacciones modelan relaciones entre entidades económicas.

En todos los casos, los datos se dividieron en conjuntos de entrenamiento, validación y test mediante una partición 80/10/10. La Tabla 1 resume las principales características de los conjuntos de datos utilizados.

4.2. Modelos Evaluados

Se evaluaron ocho modelos representativos de tres familias principales de sistemas de recomendación:

- Métodos clásicos de factorización matricial: *Matrix Factorization* (MF) [11] y *Non-Negative Matrix Factorization* (NNMF) [12]. Ambos son modelos de referencia populares debido a su simplicidad y eficiencia.

Cuadro 1: Estadísticas de los conjuntos de datos utilizados.

Conjunto	Usuarios	Ítems	Interacciones	Media	Mín.	Máx.	Densidad (%)
MovieLens 1M	6 040	3 706	1 000 209	3.58	1.00	5.00	4.47
Netflix Prize	480 189	17 770	100 480 507	3.60	1.00	5.00	1.18
GDSC1	970	402	333 161	2.07	-10.58	12.35	25.60
CTRPv2	829	544	108 224	12.75	0.11	20.00	24.00
Kiva Microloans	4 005	2 673	110 371	3.16	3.00	5.00	1.03
IMF DOTS	225	233	16 776	6.78	0.00	12.08	32.00

- Modelos basados en redes neuronales: *AutoRec* [15], *Neural Collaborative Filtering* (NCF) [7] y *Latent Relational Metric Learning* (LRML) [19]. Estos métodos, de mayor cccoste computacional, buscan capturar relaciones no lineales complejas entre usuarios e ítems.
- Modelos basados en grafos: *LightGCN* [6], *Graph-based Hybrid Recommendation System* (GHRS) [16] y *Knowledge Graph Neural Network with Label Smoothness* (KGNN-LS) [20]. Estos enfoques explotan explícitamente la estructura relacional del grafo usuario–ítem.

Para todos los modelos se utilizaron los hiperparámetros recomendados por los autores en sus respectivas publicaciones o repositorios de código fuente.

4.3. Protocolo de Evaluación

Cada modelo fue entrenado bajo los tres paradigmas descritos en la sección anterior: entrenamiento pointwise, entrenamiento basado en historial y entrenamiento orientado a ranking Top-K.

La calidad de recomendación se evaluó mediante MAE y RMSE en tareas de predicción de valoraciones, y Recall@K y NDCG@K en tareas de ranking.

Paralelamente, se monitorizaron métricas de sostenibilidad: tiempo de entrenamiento, consumo de memoria, utilización de CPU y emisiones equivalentes de CO_2.

4.4. Entorno de Implementación

Todos los experimentos se implementaron en Python utilizando PyTorch como framework principal. La monitorización de recursos se realizó mediante herramientas de profiling de sistema, mientras que la estimación de emisiones de carbono se llevó a cabo mediante la librería CodeCarbon.

Se registraron métricas tanto globales como por época de entrenamiento, permitiendo analizar la evolución temporal del coste computacional y su relación con la calidad del modelo.

5. Resultados

Con el fin de evaluar la relación entre calidad de recomendación y sostenibilidad ambiental, se analizaron los distintos modelos bajo los tres protocolos de entrenamiento propuestos: predicción punto a punto, entrenamiento basado en historial y recomendación Top-K. Aunque los experimentos se realizaron sobre múltiples conjuntos de datos, la discusión detallada se centra en MovieLens 1M debido a las limitaciones de espacio presentes y su condición como el escenario más representativo y popular de evaluación. Los resultados obtenidos en el resto de conjuntos de datos muestran tendencias consistentes y se emplearon principalmente para validar la generalidad de las conclusiones extraídas.

5.1. Resultados de rendimiento y sostenibilidad

La Tabla 2 resume el rendimiento y las emisiones asociadas a cada modelo en los tres protocolos de entrenamiento considerados.

Cuadro 2: Resultados obtenidos sobre MovieLens 1M para los tres protocolos de entrenamiento. En negrita se destacan los mejores resultados de cada métrica.

Modelo	Punto a punto			Basado en historial			Top-K		
	MAE	RMSE	CO_2 (g)	MAE	RMSE	CO_2 (g)	Recall@50	NDCG@50	CO_2 (g)
MF	**0.6764**	0.8564	2.078	**0.6754**	**0.8616**	14.006	0.9351	**0.9752**	8.922
NNMF	0.7601	0.9980	1.218	0.7690	1.0146	1.211	0.0970	0.0738	1.569
AutoRec	0.9273	1.1287	**0.086**	0.8367	1.0177	1.002	0.0392	0.0176	**0.237**
NCF	0.8224	1.0421	1.002	0.7647	0.9683	0.823	0.1942	0.1277	1.221
LRML	0.7857	**0.8358**	0.719	2.2041	2.3405	**0.494**	0.3807	0.1335	0.999
LightGCN	1.1818	1.4888	8.925	1.0490	1.2330	9.636	0.2396	0.3362	3.577
GHRS	0.8624	1.0445	0.379	0.8933	1.0730	0.498	**0.9408**	0.8705	1.760
KGNN-LS	0.7220	0.9131	7.705	0.8852	1.0876	1.291	0.2184	0.1356	6.523

En el escenario de predicción punto a punto, LRML y MF obtienen los mejores resultados en términos de precisión, alcanzando valores de RMSE de 0.8358 y 0.8564, respectivamente. Sin embargo, ambos modelos presentan perfiles de sostenibilidad diferentes. Mientras que LRML requiere únicamente 0.719,g de CO_2, MF alcanza 2.078,g debido a un entrenamiento considerablemente más costoso. En el extremo opuesto, AutoRec registra las menores emisiones de todo el estudio (0.086,g), aunque con una pérdida apreciable de precisión.

En el protocolo basado en historial, MF vuelve a obtener el menor error (RMSE = 0.8616), confirmando la robustez de los enfoques clásicos de factorización matricial. No obstante, dicho rendimiento implica también el mayor coste ambiental observado, con 14.006,g de CO_2. Por el contrario, NCF y NNMF mantienen niveles competitivos de precisión con emisiones sustancialmente inferiores. Un resultado especialmente llamativo es el deterioro de LRML, cuyo RMSE

aumenta hasta 2.3405, evidenciando una menor adaptación a este protocolo de entrenamiento.

Finalmente, en la tarea de recomendación Top-K, GHRS y MF dominan claramente al resto de alternativas. GHRS obtiene los mejores valores de Recall@50 (0.9408), mientras que MF alcanza el mayor NDCG@50 (0.9752). Estos resultados indican que ambos modelos son especialmente adecuados para tareas de ranking, aunque con perfiles de sostenibilidad distintos.

En conjunto, los resultados muestran que no existe un modelo universalmente dominante. El rendimiento relativo depende significativamente del protocolo de entrenamiento utilizado, lo que refuerza la necesidad de incorporar criterios de sostenibilidad en la selección de modelos de recomendación.

5.2. Relación entre rendimiento y emisiones

La Figura 2 representa la evolución del rendimiento frente a las emisiones acumuladas de CO_2 para los tres protocolos considerados. Esta visualización permite analizar simultáneamente la calidad predictiva alcanzada y el coste ambiental necesario para conseguirla.

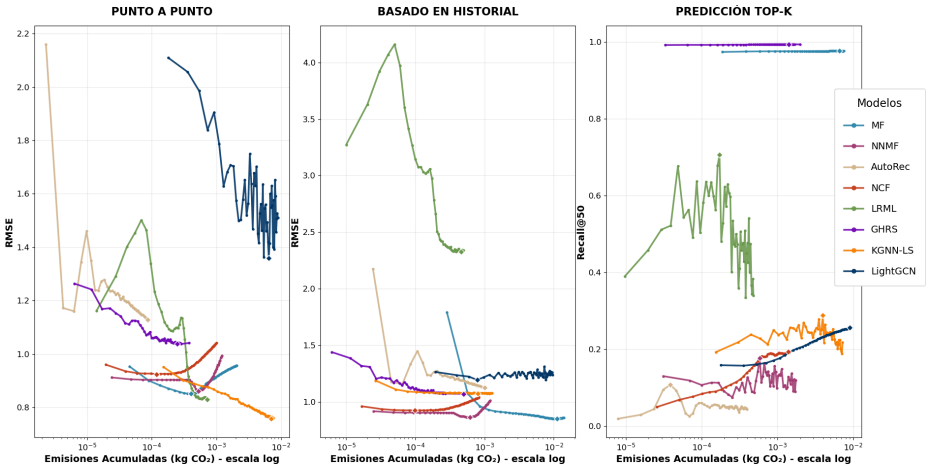


Figura 2: Relación entre rendimiento y emisiones acumuladas de los tres métodos de entrenamiento de los distintos modelos en el conjunto Movielens 1M.

En el escenario punto a punto (Figura 2a), se observa una frontera de compromiso claramente definida. AutoRec ocupa el extremo de máxima eficiencia energética, mientras que LRML y MF alcanzan los menores errores con incrementos moderados de emisiones. Por el contrario, LightGCN y KGNN-LS requieren órdenes de magnitud superiores de emisiones para obtener mejoras relativamente limitadas en RMSE.

La situación cambia en el entrenamiento basado en historial (Figura 2b). Aunque MF sigue proporcionando la mejor precisión, su coste ambiental aumenta notablemente. En este contexto, NCF y NNMF se sitúan en una posición más equilibrada, ofreciendo una combinación favorable entre precisión y sostenibilidad. En contraste, LRML presenta un comportamiento atípico, alcanzando bajas emisiones pero con una pérdida sustancial de rendimiento.

En la tarea Top-K (Figura 2c), GHRS y MF se sitúan claramente en la zona de mayor rendimiento, alcanzando valores de Recall@50 próximos a la saturación. Sin embargo, otras alternativas como LRML ofrecen compromisos interesantes cuando las restricciones energéticas son prioritarias.

Un resultado especialmente relevante es que las arquitecturas basadas en grafos no muestran un comportamiento homogéneo. Mientras que LightGCN y KGNN-LS presentan costes computacionales elevados para niveles de rendimiento moderados, GHRS logra combinar una elevada calidad de recomendación con un coste ambiental considerablemente más contenido.

6. Conclusiones

Este trabajo ha presentado un análisis sistemático de distintos modelos de recomendación bajo múltiples protocolos de entrenamiento, incorporando explícitamente métricas de sostenibilidad basadas en emisiones de CO_2 asociadas al proceso de entrenamiento. Este planteamiento permite evaluar de forma conjunta la capacidad predictiva de los modelos y su coste ambiental, en línea con las directrices actuales de Green AI y la necesidad creciente de métodos de evaluación más responsables.

En conjunto, los resultados de nuestros experimentos confirman la existencia de un claro compromiso entre precisión y sostenibilidad, donde pequeñas variaciones en arquitectura o estrategia de entrenamiento pueden inducir diferencias significativas en la huella de carbono sin mejoras sustanciales en el rendimiento. Este hecho refuerza la necesidad de incorporar métricas energéticas en la evaluación estándar de sistemas de recomendación, así como criterios de eficiencia computacional en la selección de modelos.

Como trabajo futuro, resulta relevante explorar estrategias de entrenamiento multiobjetivo que optimicen simultáneamente rendimiento y consumo energético, Asimismo, sería de interés ampliar el análisis a infraestructuras heterogéneas o a gran escala, donde el impacto de la optimización energética puede ser aún más significativo.

Declaración de intereses Los autores declaran no tener intereses en conflicto.

Agradecimientos Este trabajo está financiado por MICIU/AEI FEDER Una forma de hacer Europa (PID2023-147404OB-I00), Ministerio de Transformación Digital y Función Pública y 'Next-GenerationEU' /PRTR (TSI-100925-2023-1), y Xunta de Galicia (ED431C 2026/38, ED431G 2023/01).

Referencias

1. Bolón-Canedo, V., Morán-Fernández, L., Alonso-Betanzos, A.: Green AI: A review on energy-efficient Machine Learning. Neurocomputing **568**, 127083 (2024). https://doi.org/10.1016/j.neucom.2023.127083
2. Broad Institute: Cancer therapeutics response portal v2 (ctrpv2) (2015), https://portals.broadinstitute.org/ctrp.v2/
3. Carbonell, J., Smith, J.: Green ai: Measuring and reducing the carbon footprint of recommender systems. Journal of Sustainable Computing **45**, 102–115 (2023)
4. Chadli, M., col: Promoting sustainability in the lifecycle of machine learning systems: A systematic mapping. Computer Science Review / TIC Sustainability Workshop (2024), mapeo sistemático de 52 publicaciones entre 2019 y 2023; menciona CodeCarbon y CarbonTracker
5. GroupLens Research: Movielens 1m dataset (2003), https://grouplens.org/datasets/movielens/1m/
6. He, X., Deng, K., et al.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: SIGIR (2020)
7. He, X., Liao, L., Zhang, H.: Neural collaborative filtering. In: WWW (2017)
8. Henderson, P., Hu, J., Romalov, J., von Platen, P., Leshem, G., Pineau, J.: Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research **21**(248), 1–43 (2020), http://jmlr.org/papers/v21/20-312.html
9. International Monetary Fund: Imf direction of trade statistics (dots) (2020), https://data.imf.org/
10. Kiva: Kiva microloans dataset (2017), https://www.kiva.org/
11. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. In: IEEE Computer (2009)
12. Lee, D.D., Seung, H.S.: Non-negative matrix factorization for recommender systems (2001)
13. Netflix Inc.: Netflix prize dataset (2006), https://www.netflixprize.com/
14. Ricci, F., Rokach, L., Shapira, B.: Recommender Systems Handbook. Springer, New York, NY, USA (2015). https://doi.org/10.1007/978-1-4899-7637-6
15. Sedhain, S., et al.: Autorec: Autoencoders meet collaborative filtering. In: WWW (2015)
16. Unknown: Graph-based hybrid recommendation system (2021)
17. Vente, J., Wegmeth, S., col.: The environmental cost of recommendation: A comparative study of RecSys 2013 and 2023. In: Proceedings of the ACM Conference on Recommender Systems (RecSys) (2024), análisis detallado de emisiones de CO2 y métricas de ranking como NDCG
18. Verdecchia, R., Tamburri, D.A., Lago, P.: Energy-efficient Machine Learning: A comprehensive survey of approaches and open research challenges. Journal of Systems and Software **199**, 111612 (2023). https://doi.org/10.1016/j.jss.2023.111612, https://arxiv.org/abs/2301.11047
19. Wan, P., et al.: Latent relational metric learning for collaborative ranking (2018)
20. Wang, X.e.a.: Knowledge graph neural network with label smoothness (2019)
21. Wellcome Sanger Institute: Genomics of drug sensitivity in cancer (gdsc1) (2018), https://www.cancerrxgene.org/