

© International Baccalaureate Organization 2023

All rights reserved. No part of this product may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without the prior written permission from the IB. Additionally, the license tied with this product prohibits use of any selected files or extracts from this product. Use by third parties, including but not limited to publishers, private teachers, tutoring or study services, preparatory schools, vendors operating curriculum mapping services or teacher resource digital platforms and app developers, whether fee-covered or not, is prohibited and is a criminal offense.

More information on how to request written permission in the form of a license can be obtained from <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

© Organisation du Baccalauréat International 2023

Tous droits réservés. Aucune partie de ce produit ne peut être reproduite sous quelque forme ni par quelque moyen que ce soit, électronique ou mécanique, y compris des systèmes de stockage et de récupération d'informations, sans l'autorisation écrite préalable de l'IB. De plus, la licence associée à ce produit interdit toute utilisation de tout fichier ou extrait sélectionné dans ce produit. L'utilisation par des tiers, y compris, sans toutefois s'y limiter, des éditeurs, des professeurs particuliers, des services de tutorat ou d'aide aux études, des établissements de préparation à l'enseignement supérieur, des fournisseurs de services de planification des programmes d'études, des gestionnaires de plateformes pédagogiques en ligne, et des développeurs d'applications, moyennant paiement ou non, est interdite et constitue une infraction pénale.

Pour plus d'informations sur la procédure à suivre pour obtenir une autorisation écrite sous la forme d'une licence, rendez-vous à l'adresse <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

© Organización del Bachillerato Internacional, 2023

Todos los derechos reservados. No se podrá reproducir ninguna parte de este producto de ninguna forma ni por ningún medio electrónico o mecánico, incluidos los sistemas de almacenamiento y recuperación de información, sin la previa autorización por escrito del IB. Además, la licencia vinculada a este producto prohíbe el uso de todo archivo o fragmento seleccionado de este producto. El uso por parte de terceros —lo que incluye, a título enunciativo, editoriales, profesores particulares, servicios de apoyo académico o ayuda para el estudio, colegios preparatorios, desarrolladores de aplicaciones y entidades que presten servicios de planificación curricular u ofrezcan recursos para docentes mediante plataformas digitales—, ya sea incluido en tasas o no, está prohibido y constituye un delito.

En este enlace encontrará más información sobre cómo solicitar una autorización por escrito en forma de licencia: <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

Informática

Estudio de caso: ¿Puedo recomendar lo siguiente?

Para usar en mayo y noviembre de 2023

Instrucciones para los alumnos

- Para la prueba 3 de Nivel Superior se requiere el cuadernillo del estudio de caso.

Página en blanco

Introducción

Usted acaba de terminar una videoconferencia con sus dos amigos, Jungmin y Lijing. Ellos tienen una idea para un negocio en línea y quieren su ayuda. La empresa se llamará *NextStar*. Ofrecerá una aplicación para que los usuarios puedan ver la obra de artistas que aún no han sido descubiertos.

Los artistas pueden ser actores, cantantes, guionistas, cómicos, pintores, escultores o cineastas. De hecho, cualquier artista que quiera demostrar su talento podrá subir archivos a la aplicación. El contenido cargado puede ser calificado por todos los usuarios. En función de estas valoraciones, la aplicación recomienda nuevos contenidos a cada usuario.

- 10 Jungmin y Lijing planean que el sitio web de *NextStar* sea gratuito y creen que podrán ganar dinero con la publicidad cuando haya suficientes usuarios. Son conscientes de que esta aplicación acabará necesitando una gran cantidad de almacenamiento, por lo que están buscando empresas de alojamiento en la nube. Una vez que se hayan añadido suficientes contenidos, la aplicación incorporará un sistema de recomendación.
- 15 La siguiente información ofrece un resumen de lo que ya se ha investigado e incluye algunos desafíos que debe tener en cuenta.

Informática en la nube

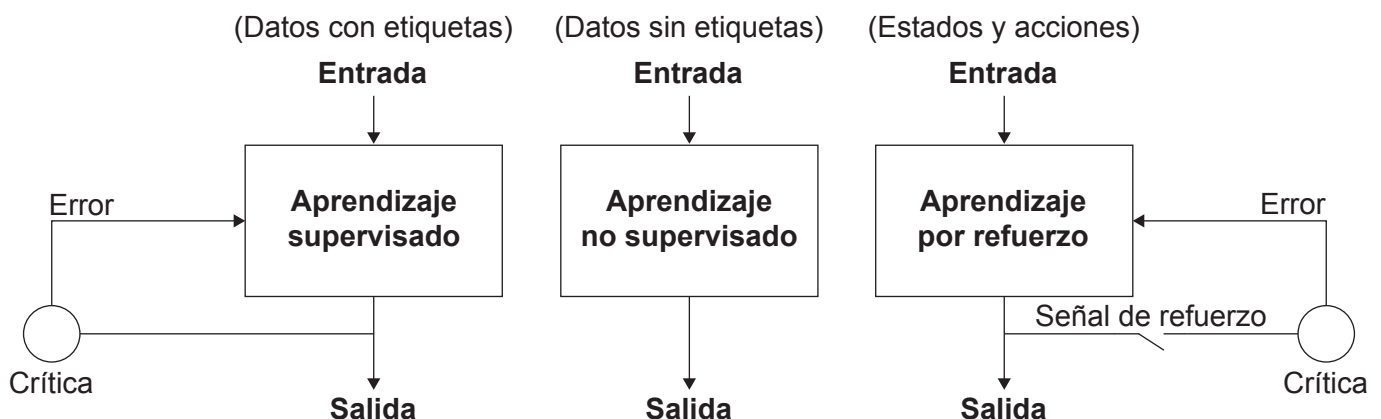
El alojamiento de aplicaciones que utilizan datos a nivel empresarial está ampliamente disponible y es asequible gracias a la informática en la nube. Los usuarios solo pagan por los recursos que utilizan, por lo que pueden empezar con poco y añadir más recursos a medida que estos crecen. Esto hace que la informática en la nube sea ideal para una empresa nueva como *NextStar*.

Existen varios *modelos de implementación en la nube* que podrían utilizarse para alojar los datos de *NextStar*. También hay tres *modelos de servicio en la nube*: *software como servicio (SaaS)*, *plataforma como servicio (PaaS)* e *infraestructura como servicio (IaaS)*. *NextStar* tiene la intención de utilizar IaaS.

Aprendizaje automático

El aprendizaje automático es un subcampo de la inteligencia artificial. Existen tres tipos principales de aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y *aprendizaje por refuerzo* (véase la **figura 1**).

Figura 1: Los tres tipos principales de aprendizaje automático



Véase al dorso

- 30 Un algoritmo de aprendizaje supervisado utiliza *datos de entrenamiento* etiquetados para aprender una función que produce una salida adecuada, cuando se le dan nuevos datos no etiquetados. Normalmente, el aprendizaje supervisado se utiliza para clasificar datos o hacer predicciones.

Un algoritmo de aprendizaje no supervisado aprende patrones a partir de datos no etiquetados. Estos algoritmos toman referencias de las observaciones de los datos de entrada en vivo.

- 35 El sistema puede organizar los datos en subconjuntos, o grupos, que no han sido preclasificados por los programadores.

Un algoritmo de aprendizaje por refuerzo aprende en un entorno interactivo por ensayo y error utilizando la retroalimentación de sus propias acciones y experiencias. Algunos sistemas de recomendación pueden considerarse un tipo de aprendizaje por refuerzo, ya que un comportamiento positivo, como la revisión de contenidos, se recompensa con mejores recomendaciones.

- 40

Sistemas de recomendación

Cuando hay una gran cantidad de contenidos disponibles, un sistema de recomendación dirige al usuario a contenidos que no ha visto pero que pueden ser de su interés. Los sistemas de recomendación utilizan los datos de los contenidos que los usuarios ya han valorado (datos reales) para generar las preferencias predichas de los contenidos que aún no han valorado (datos anticipados).

- 45

La mayoría de los sistemas de recomendación utiliza el aprendizaje supervisado. Es menos común el uso del aprendizaje no supervisado y del aprendizaje por refuerzo.

- 50 Los sistemas de recomendación pueden utilizar el *filtrado basado en el contenido*, el *filtrado colaborativo* o una combinación de ambos. Los sistemas de recomendación híbridos combinan varios algoritmos de aprendizaje automático. Esto se demostró el 21 de septiembre de 2009, cuando el equipo Pragmatic Chaos de BellKor ganó el Premio Netflix y USD 1 000 000 por el mejor sistema de recomendación de películas mediante filtrado colaborativo. Este sistema de recomendación combinó 107 algoritmos diferentes en un modelo híbrido que superó, en un 10,06 %, la puntuación de la *media cuadrática de errores* (RMSE, por sus siglas en inglés) del propio algoritmo de Netflix.
- 55

Filtrado basado en el contenido

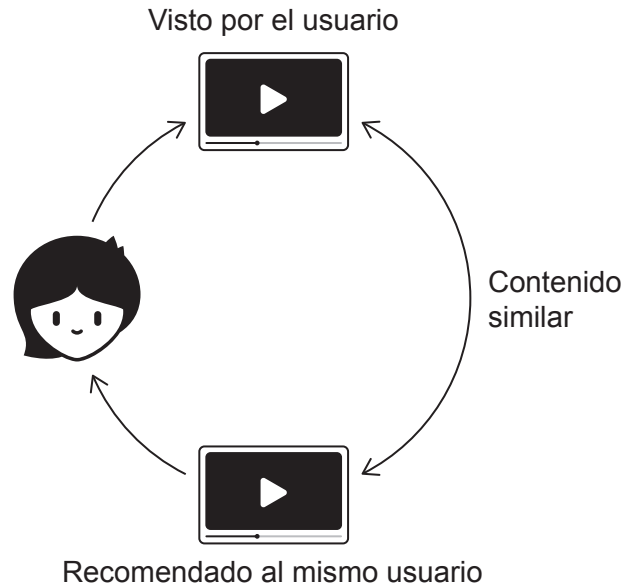
El filtrado basado en el contenido, a veces llamado filtrado elemento por elemento, o artículo por artículo, se centra en los atributos de un elemento en lugar de utilizar las interacciones y la retroalimentación del usuario. El enfoque basado en el contenido es una clasificación específica del usuario, en la que el clasificador aprende lo que le gusta y lo que no le gusta al usuario basándose en los atributos de un artículo.

- 60

Dado que el sistema de recomendación de *NextStar* contendrá videoclips, los atributos podrían incluir el género artístico, la fecha de lanzamiento, el artista, el idioma, el género de la persona y su edad. Por ejemplo, si un usuario valora mucho los videoclips de comedia, es probable que el sistema le recomiende más videos de comedia (véase la **figura 2**).

- 65

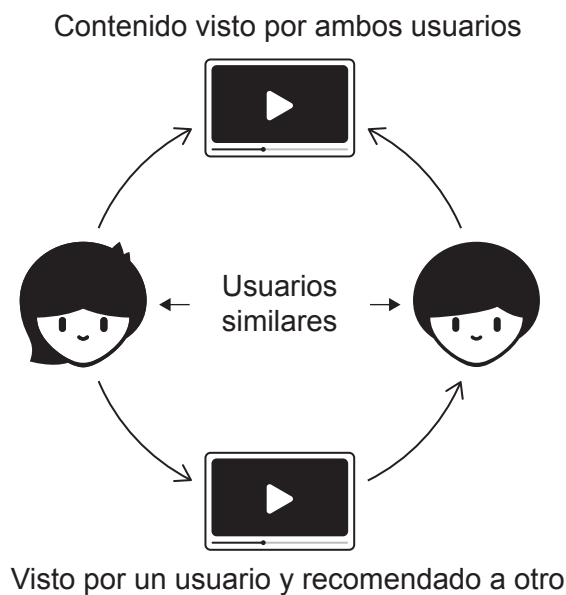
Figura 2: Un ejemplo de filtrado basado en el contenido



Filtrado colaborativo

- Con el filtrado colaborativo, las recomendaciones para cada usuario se generan utilizando la información de valoración de otros usuarios y artículos. La hipótesis central es que los usuarios que han estado de acuerdo en el pasado tienden a estarlo en el futuro. Así, si dos usuarios dan una puntuación similar a un contenido, es probable que el otro usuario disfrute de otro contenido muy valorado por uno de ellos. Así, ese contenido puede recomendarse al segundo usuario (véase la **figura 3**). Una de las limitaciones de los sistemas de recomendación por filtrado colaborativo es el *sesgo de popularidad*, por el que los contenidos populares se recomiendan con demasiada frecuencia.

Figura 3: Un ejemplo de filtrado colaborativo



El filtrado colaborativo puede utilizar diferentes algoritmos para recomendar nuevos contenidos. Dos tipos de algoritmos que se pueden utilizar son el *vecino más cercano k (k -NN)* y la *factorización matricial*.

80 Vecino más cercano k (k-NN)

El algoritmo k-NN utiliza la similitud de características para predecir los valores de cualquier dato nuevo o faltante. Esto significa que a los nuevos puntos de datos se les asigna un valor basado en su parecido con otros puntos de datos del conjunto de entrenamiento.

85 El algoritmo k-NN realiza sus predicciones basándose en los vecinos más cercanos. El aspecto “k” de este algoritmo representa el número de vecinos y es simplemente un *hiperparámetro* que puede ajustarse mediante un enfoque de prueba y error.

Factorización matricial

90 La factorización matricial es una alternativa al algoritmo k-NN. La dificultad de utilizar enfoques de factorización matricial estándar para los sistemas de recomendación es que el conjunto de datos no es completo. Para superar esta limitación, es necesario estimar los valores de las matrices más pequeñas mediante un algoritmo iterativo.

En la **figura 4**, la matriz de interacción usuario-elemento representa la valoración de cada usuario (filas) de cada elemento de contenido (columnas). El usuario 1, representado por US1, ha valorado los tres primeros elementos, pero no ha valorado el elemento 4 ni el 5, representados por EL4 e EL5.

Figura 4: Matriz de interacción usuario-elemento

		Elemento				
		EL1	EL2	EL3	EL4	EL5
Usuario	US1	3	2	3		
	US2		2	2	2	
	US3	3	4		4	2
	US4	2	2	4		1
	US5	2		3	2	1
	US6			3	1	1

Clave:  Contenido desconocido

100 La factorización matricial funciona al dividir la gran matriz de interacción usuario-elemento en dos matrices más pequeñas —una matriz elemento-característica y una matriz usuario-característica— para captar las características más importantes necesarias para el aprendizaje. Si se modifican los valores de la matriz elemento-característica y de la matriz usuario-característica, también cambiarán los valores correspondientes de la matriz de interacción usuario-elemento (véase la **figura 5**).

Figura 5: Factorización matricial

Matriz elemento-característica

Elemento

	EL1	EL2	EL3	EL4	EL5	
Característica	C1	1,2	3,1	2,1	4,5	0,7
	C2	2,6	1,5	4,4	0,4	1,1

Matriz usuario-característica

Característica

	C1	C2
US1	0,3	0,7
US2	0,3	0,4
US3	0,7	0,8
US4	0,4	0,6
US5	0,7	0,4
US6	0,1	0,6

Matriz de interacción usuario-elemento

Elemento

	EL1	EL2	EL3	EL4	EL5
US1	2,18	1,98	3,71	1,63	0,98
US2	1,4	1,53	2,39	1,51	0,65
US3	2,92	3,37	4,99	3,47	1,37
US4	2,04	2,14	3,48	2,04	0,94
US5	1,88	2,77	3,23	3,31	0,93
US6	1,68	1,21	2,85	0,69	0,73

Las matrices se ajustan al generar preferencias anticipadas para contenidos sobre los que ya existen datos de preferencias reales. Una vez que los valores de predicción se acercan a la calificación real, se supone que las matrices podrán predecir eficazmente las preferencias para las que no existen datos reales.

Un proceso llamado *descenso de gradiente estocástico* utiliza una *función de costo* con el fin de ajustar cada celda haciendo un pequeño cambio en las matrices elemento-característica y usuario-característica. Por ejemplo, en la **figura 4**, el valor de la celda de intersección de US1 e EL1 es 3, pero en la **figura 5** este valor es 2,18. Por lo tanto, el error de esta celda es $(3 - 2,18)^2$, es decir 0,6724.

Entrenamiento de los sistemas de recomendación

Un sistema de recomendación puede ser evaluado utilizando divisiones de entrenamiento/prueba. Los datos de calificación se dividen en un conjunto de entrenamiento y un conjunto de prueba. Una división comúnmente utilizada es cuando el 80 % de los datos se asigna al conjunto de entrenamiento y el otro 20 % al conjunto de prueba.

Un sistema de recomendación aprende las relaciones entre los elementos y entre los usuarios. Después de entrenado, hace predicciones sobre cómo un usuario podría valorar un elemento que aún no ha valorado.

Un problema común del entrenamiento de un algoritmo de aprendizaje automático es el *sobreajuste*, en el que el modelo se ajusta demasiado al conjunto de datos de entrenamiento. Cuando el modelo se entrena durante demasiado tiempo sobre los datos de entrenamiento, o cuando el modelo es demasiado complejo, puede empezar a aprender características irrelevantes con respecto al conjunto de datos. En consecuencia, el modelo no se generaliza eficazmente frente a nuevos datos.

125 Evaluación de los sistemas de recomendación

La precisión del sistema de recomendación puede evaluarse mediante dos medidas diferentes: la *media absoluta de errores* (MAE, por sus siglas en inglés) y la media cuadrática de errores (RMSE, por sus siglas en inglés). Estas medidas presentan una indicación del rendimiento de los sistemas de recomendación en los datos de entrenamiento/prueba.

- 130 Sin embargo, la eficacia de un sistema de recomendación no se conoce del todo hasta que lo utiliza el público. Un sistema de recomendación no funciona bien si no recomienda el contenido que le gustaría al usuario o le sugiere un contenido que no le gusta.

- 135 *La precisión y el recuerdo* son métricas de rendimiento utilizadas en datos reales. La precisión es una medida de exactitud, que indica la fracción de instancias relevantes entre las instancias recuperadas. El recuerdo es una medida de completitud. La *medida F* proporciona una única puntuación que equilibra las cuestiones de precisión y recuerdo.

- 140 También es importante la forma en que se muestran las recomendaciones a los usuarios. Una lista puede ser suficiente, o es factible seleccionar el contenido recomendado por grupos o subgrupos. Estos grupos pueden organizarse por género artístico, edad o género de la persona, o cualquier otra categoría posible.

Preocupaciones sociales y éticas

Cuando se construye un modelo a partir del comportamiento de los usuarios, se pueden utilizar dos tipos de *datos de comportamiento*: datos explícitos y datos implícitos.

- 145 Los datos de comportamiento explícitos se refieren a los recogidos a partir de aquellos enviados por los usuarios, como cuando un usuario valora un videoclip, ingresa sus preferencias o busca un artículo. Los usuarios pueden creer que estos son los únicos datos que se utilizan para hacer recomendaciones.

- 150 Los datos de comportamiento implícitos se refieren a aquellos que el usuario no sabe que también se compilan. Esto puede incluir datos de clics, de compras o hasta el uso de un registrador de teclas.

La calidad de los datos de los usuarios es fundamental para el éxito del proyecto *NextStar*, pero existen preocupaciones éticas sobre la compilación, el almacenamiento y el uso de los datos de comportamiento. *NextStar* también debe tener en cuenta el *derecho al anonimato* y *derecho a la privacidad* de sus usuarios.

155 Desafíos a los que se enfrenta

Para ayudar a sus amigos con su nuevo emprendimiento empresarial, hay una serie de desafíos que debe investigar:

- Comprender las similitudes y diferencias entre el aprendizaje supervisado, el no supervisado y el por refuerzo.
- 160 • Comprender cómo el algoritmo k-NN y la factorización matricial pueden utilizarse en los sistemas de recomendación.
- Entender cómo entrenar, probar y evaluar un sistema de recomendación.
- Comparación de los sistemas de recomendación basados en el filtrado de contenidos y en el de tipo colaborativo.
- 165 • Comprender las preocupaciones éticas relacionadas con la compilación, el almacenamiento y la utilización de los datos de comportamiento de los usuarios.

No es necesario que los candidatos conozcan las ecuaciones matemáticas relacionadas con los sistemas de recomendación.

Terminología adicional

Algoritmo del vecino más cercano k (k-NN, *k-nearest neighbour*)
Aprendizaje por refuerzo (*reinforcement learning*)
Datos de entrenamiento (*training data*)
Datos de comportamiento (*behavioural data*)
Derecho al anonimato (*right to anonymity*)
Derecho a la privacidad (*right to privacy*)
Descenso de gradiente estocástico (*stochastic gradient descent*)
Factorización matricial (*matrix factorization*)
Filtrado basado en el contenido (*content-based filtering*)
Filtrado colaborativo (*collaborative filtering*)
Función de costos (*cost function*)
Hiperparámetro (*hyperparameter*)
Media absoluta de errores (MAE, *mean absolute error*)
Media cuadrática de errores (RMSE, *root-mean-square error*)
Medida F (*F-measure*)
Modelos de implementación en la nube (*cloud deployment models*)
Modelos de servicio en la nube (*cloud delivery models*):
 infraestructura como servicio (IaaS)
 plataforma como servicio (PaaS)
 software como servicio (SaaS)
Precisión (*precision*)
Recuerdo (*recall*)
Sesgo de popularidad (*popularity bias*)
Sobreajuste (*overfitting*)

Algunas empresas, productos o personas nombradas en este estudio de caso son ficticios y cualquier semejanza con entidades reales es solamente una coincidencia.

Advertencia:

Los contenidos usados en las evaluaciones del IB provienen de fuentes externas auténticas. Las opiniones expresadas en ellos pertenecen a sus autores y/o editores, y no reflejan necesariamente las del IB.

Referencias:

Figura 1 Jones, M. Tim., 2017. Models for machine learning. [en línea] Disponible en: <https://developer.ibm.com/articles/cc-models-machine-learning/> [Consultado: 15 de octubre de 2021]. Material original adaptado.

Los demás textos, gráficos e ilustraciones: © Organización del Bachillerato Internacional, 2023