

# Recomendación Cinematográfica: Explorando Preferencias con Aprendizaje No Supervisado

## **Resumen**

Netflix, Amazon, Max, Disney+, mubi, el nuevo tubi<sup>1</sup>? Servicios por suscripción masivos, orientados a nichos, algunos gratis? Todos compiten por mantener y atraer a usuarios que cada vez tienen más opciones para escoger. Aún cuando creemos que este negocio ya está completamente inventado, ¿cuántas veces nos sorprendemos por las recomendaciones que recibimos de estos servicios? Una lista realmente curada para cada uno es la aspiración que tenemos todos. El propósito de este trabajo es explorar esas preferencias y tratar de encontrar oportunidades que conduzcan a tales recomendaciones curadas para los usuarios.

Para crear un algoritmo de recomendación de películas hemos extraído los datos disponibles en Kaggle y realizado una revisión inicial para determinar su utilidad en nuestro proyecto. Actualmente, contamos con tres bases de datos en formato csv: keywords, movie\_metadata y ratings\_small. Con estos datos se creará un modelo de recomendación de películas teniendo en cuenta la calificación que le dan los usuarios a las películas además de otras características para ofrecerles recomendaciones interesantes para ver.

## **Introducción**

¿Qué películas me recomienda un sistema de aprendizaje no supervisado sabiendo que he calificado ciertas películas con un puntaje entre 1 y 5?

La clave del éxito de las plataformas de streaming, sobre todo Netflix la precursora de estas, es la capacidad para conocer y adaptarse a las preferencias y gustos de diferentes mercados y audiencias. La inversión que hace Netflix en contenido original ha permitido que el mundo del streaming se haya expandido tanto, pero más que tener un montón de películas para ver sin parar, Netflix sabe en cuáles películas invertir que les genere mayor retorno. Esta pregunta es interesante porque plantea que, al saber qué películas les gustan a los clientes, es posible predecir qué otras películas podrían disfrutar dada su predilección por las primeras. Este es un problema de agrupamiento, donde buscamos crear clústeres de películas infiriendo que la similitud entre las mismas implica que si a una persona le gustó una película dentro de ese cluster podría también gustarle otra película del mismo y por ello recomendaremos en base a ello.

## **Revisión preliminar de antecedentes en la literatura**

Los sistemas de recomendación (Recommender Systems) son aplicaciones que ofrecen a los usuarios recomendaciones personalizadas de productos y servicios que aún no han adquirido, basadas en sus intereses, con el objeto de incrementar las ventas y mantener la base de clientes satisfecha. Estos sistemas son ampliamente usados en diversas áreas como en plataformas de

---

1

[https://www.nytimes.com/2024/08/13/business/media/tubi-movies-tv-streaming.html?unlocked\\_article\\_code=1.G04.C76S.micrXPlvHnCm&smid=url-share](https://www.nytimes.com/2024/08/13/business/media/tubi-movies-tv-streaming.html?unlocked_article_code=1.G04.C76S.micrXPlvHnCm&smid=url-share)

streaming, redes sociales, planeación de viajes, recomendaciones de música, colocación laboral entre otras [1]

Amazon patentó su primera versión de su sistema de recomendación alrededor del 2004. Esta trajo consigo un incremento del 29% en las ganancias alcanzando los US\$12.83 billones en el segundo trimestre fiscal, mientras que Netflix implementó esta herramienta en su aplicación para disminuir el número de suscripciones canceladas y al mismo tiempo aumentar el tiempo promedio que los usuarios interactúan con la aplicación. [2]

Estos sistemas fueron desarrollados para hacer sugerencias con base en las preferencias de los usuarios: su género favorito, actores, directores, entre otros parámetros, por ejemplo el sistema de Netflix también agrega una explicación que ayuda al usuario entender por qué esa película esta siendo recomendada, ayudando a incrementar la credibilidad de la aplicación y la lealtad del usuario. Otro enfoque es la propuesta de películas basada en emociones por ejemplo felicidad, enojo o tristeza.[2]

En el área de diseño e implementación de sistemas de recomendación actualmente existen 4 enfoques: [2]

- Sistemas de recomendación basado en contenido
- Sistemas de recomendación basado en filtrado colaborativo
- Sistemas de recomendación basado en conocimiento
- Sistemas de recomendación híbrido

\* Profundizaremos en estos enfoques a medida que avancemos en el desarrollo del curso y el proyecto

### **Descripción detallada de los datos**

Estas son las 3 bases de datos que elegimos:

**movies\_metadata:** El archivo principal de metadatos de películas. Contiene información sobre 45,000 películas originalmente incluidas en el conjunto de datos completo de MovieLens, un grupo de investigación del Departamento de Ingeniería y Ciencias Computacionales de la Universidad de Minnesota. Los datos se originaron entre 1996 y 2018. Las características incluyen pósters, fondos, presupuesto, ingresos, fechas de lanzamiento, idiomas, países y compañías de producción.

**keywords:** Incluye las palabras clave de la trama de las películas del conjunto de datos MovieLens. Está disponible en forma de un objeto JSON convertido a cadena.

**ratings\_small:** El subconjunto de 100,000 calificaciones de 700 usuarios en 9,000 películas.

En `movies_metadata` encontramos toda la información perteneciente a la película: género, presupuesto, una marca si es para adultos o no, un identificador para cruzar con las otras bases, título original, idioma original, resumen, año de estreno, cuantas personas dieron un rating a la película, entre otras:

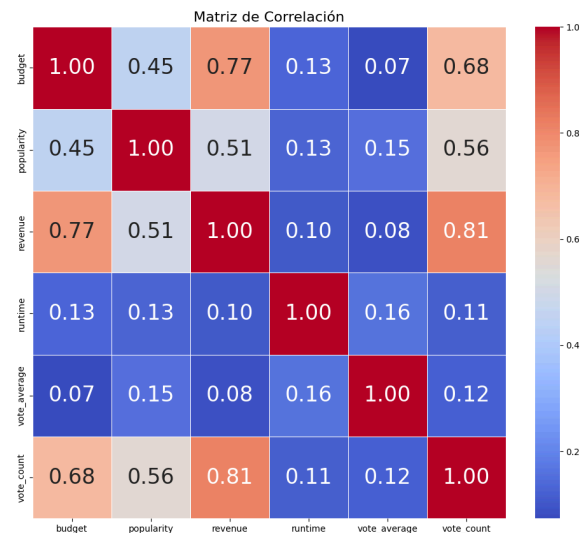
- 'id': 'Identificador único para cada película en el dataset. Puede no ser único a nivel global.',
- 'imdb\_id': 'Identificador único de la película en IMDb. Formato: "ttXXXXXX", donde X son dígitos.',
- 'original\_title': 'Título original de la película tal como se presenta en su país de origen.',
- 'overview': 'Descripción o sinopsis de la película. Proporciona un resumen breve de la trama.',
- 'popularity': 'Puntuación que refleja la popularidad de la película. Generalmente calculada con base en factores como las búsquedas y visualizaciones.',
- 'release\_date': 'Fecha en la que la película fue estrenada en cines. Formato: "YYYY-MM-DD".',
- 'revenue': 'Monto total de ingresos generados por la película en dólares. Puede incluir taquilla y otros ingresos.',
- 'runtime': 'Duración de la película en minutos. Representa el tiempo total que dura la película.',
- 'status': 'Estado actual de la película. Ejemplos incluyen "Released" (estrenada), "Post Production" (en postproducción), etc.',
- 'tagline': 'Slogan o lema de la película, utilizado como herramienta de marketing. Generalmente una frase corta y atractiva.',
- 'title': 'Título de la película en el idioma principal del dataset o el título que se usa comúnmente.',
- 'vote\_average': 'Promedio de las calificaciones recibidas de los usuarios. Un valor numérico que refleja la recepción general de la película.',
- 'vote\_count': 'Número total de votos que la película ha recibido. Utilizado para calcular el promedio de votos.',
- 'budget': 'Monto total del presupuesto de producción de la película en dólares.',
- 'genres': 'Lista de géneros a los que pertenece la película. Generalmente en formato JSON, donde cada género tiene un ID y un nombre.',
- 'production\_companies': 'Lista de compañías productoras involucradas en la realización de la película. En formato JSON, donde cada entrada incluye información como el nombre de la compañía y su ID.',
- 'production\_countries': 'Lista de países en los que se produjo la película. En formato JSON, incluyendo el nombre del país y su código.',
- 'spoken\_languages': 'Idiomas hablados en la película. En formato JSON, donde cada idioma tiene un nombre y un código.',
- 'adult': 'Indica si la película es para adultos. Generalmente un valor booleano.',
- 'belongs\_to\_collection': 'Información sobre la colección a la que pertenece la película, en formato JSON si aplica. Incluye ID y nombre de la colección.',
- 'homepage': 'URL de la página web oficial de la película.',
- 'original\_language': 'Código del idioma original en el que se hizo la película, utilizando el estándar ISO 639-1.',
- 'poster\_path': 'Ruta o URL del cartel (poster) de la película. Generalmente una cadena que representa la ruta al archivo de imagen.'

Cabe destacar que no contamos con información demográfica de los usuarios, lo cual hubiese sido importante para hacer agrupamientos por esas características, y quizás este hecho limite el alcance propuesto.

Se realizó una limpieza preliminar de los datos, descartando algunas variables que no aportan valor debido a un alto porcentaje de datos faltantes, o porque más del 90% de las observaciones pertenecen a una sola categoría.

Las características que resaltamos de la exploración de datos inicial se encuentran a continuación y otras características en los anexos al final.

## Matriz de correlación variables numéricas



La correlación más alta ocurre entre el ingreso generado por las películas y el número de votos recibidos (0.81)

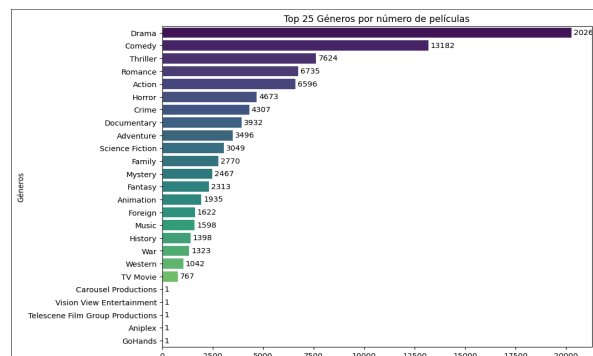
La siguen la correlación entre el presupuesto y los ingresos generados (0.77), y la relación entre el primero y el número de votos (0.68)

Un análisis de componentes principales ayudará a reducir la dimensionalidad.

Los ingresos-popularidad (0.51) se relacionan en mayor medida que el presupuesto con su popularidad (0.45).

Todas las correlaciones con la media de votos son bajas, y no se presentan correlaciones negativas.

## Géneros

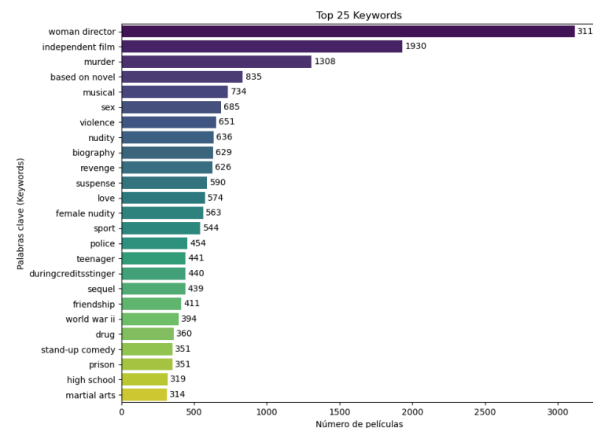


Algo interesante de esta base es que una película puede tener más de un género.

La mayoría son películas dramáticas, seguidas de comedia, suspenso, romance y acción.

También existen datos que parecen ser productoras y no géneros, lo que tendremos en cuenta en la corrección de datos.

## Base Palabras Clave

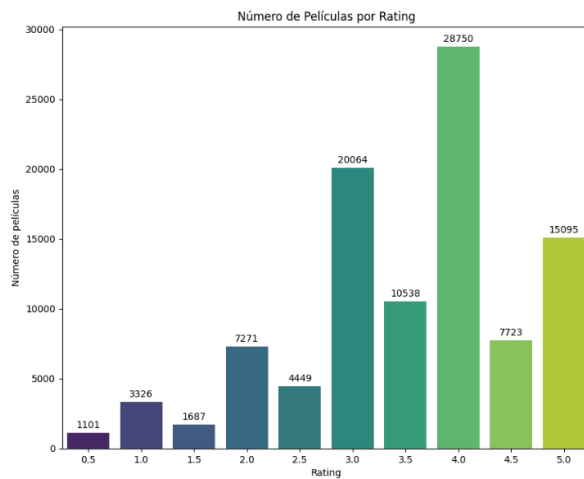


**keywords:** esta base de información es interesante y podría servir para generar clusters.

Tiene dos columnas: id de la película y keywords en formato JSON, palabras clave que describen la película. Por ejemplo para Toy Story las palabras clave son: jealousy, toy, boy, friendship, friends, rivalry, toy comes to life, etc. Por lo que se infiere el contenido de la película.

En los primeros lugares de frecuencia se destacan “woman director” e “independent film”.

### Base Ratings



**ratings:** esta base contiene la calificación que cada usuario (671 usuarios en total) le dio a cada película: se tienen el userId (que persona dio el rating), movieId, IMDbId (que es como se puede buscar la película por la página de IMDB) y el rating que le dio el usuario

En la gráfica de la izquierda se observa el conteo de las películas y el rating que le dieron los usuarios.

Al momento de trabajar los datos será importante verificar que un usuario solo califique una película una vez, para evitar errores más adelante.

### Propuesta metodológica

El alcance inicial definido será desarrollar modelos de clustering con diferentes técnicas de agrupamiento y medidas de distancia.

En primer lugar haremos una análisis de componentes principales PCA para determinar las dimensiones más relevantes en los datos.

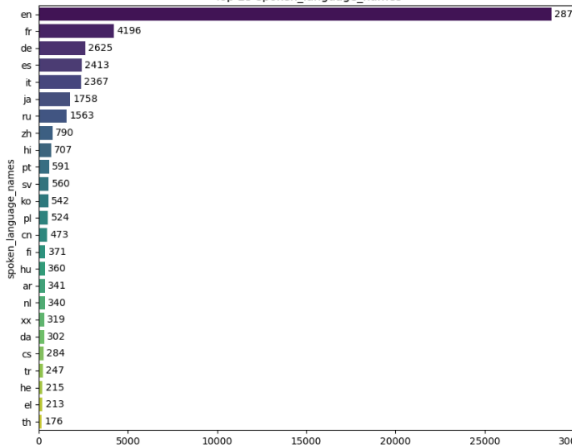
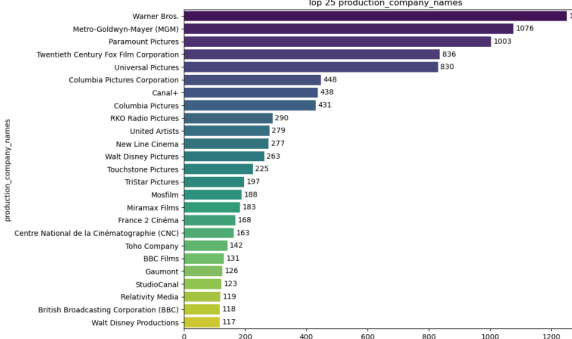
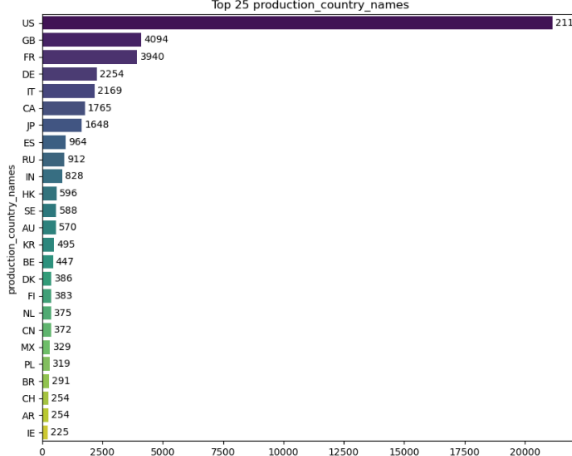
Identificamos variables bastante interesantes, que codificaremos pues son en su mayoría categóricas.

Después implementaremos los métodos de agrupamiento, comenzando con algunas variables de interés, e incluyendo otras paulatinamente para analizar los resultados obtenidos, y finalmente determinar los grupos de películas que se parecen y de este modo hacer recomendaciones según las calificaciones de otras películas que han visto.

Es importante resaltar que esta metodología se enriquecerá a medida que avancemos en los contenidos del curso.

**Repositorio GitHub:** [https://github.com/juankquintana/proyecto\\_recomendacion\\_peliculas](https://github.com/juankquintana/proyecto_recomendacion_peliculas)

**Anexos:**

<p>Lenguajes</p> <p>Top 25 spoken_language_names</p>  <table><thead><tr><th>spoken_language_names</th><th>Número de películas</th></tr></thead><tbody><tr><td>en</td><td>28745</td></tr><tr><td>fr</td><td>4196</td></tr><tr><td>de</td><td>2625</td></tr><tr><td>es</td><td>2413</td></tr><tr><td>it</td><td>2367</td></tr><tr><td>ja</td><td>1758</td></tr><tr><td>ru</td><td>1563</td></tr><tr><td>zh</td><td>790</td></tr><tr><td>hi</td><td>707</td></tr><tr><td>pt</td><td>591</td></tr><tr><td>sv</td><td>560</td></tr><tr><td>ko</td><td>542</td></tr><tr><td>pl</td><td>524</td></tr><tr><td>cn</td><td>473</td></tr><tr><td>is</td><td>371</td></tr><tr><td>hu</td><td>360</td></tr><tr><td>ar</td><td>341</td></tr><tr><td>nl</td><td>340</td></tr><tr><td>xx</td><td>319</td></tr><tr><td>da</td><td>302</td></tr><tr><td>cs</td><td>284</td></tr><tr><td>tr</td><td>247</td></tr><tr><td>he</td><td>215</td></tr><tr><td>el</td><td>213</td></tr><tr><td>th</td><td>176</td></tr></tbody></table>	spoken_language_names	Número de películas	en	28745	fr	4196	de	2625	es	2413	it	2367	ja	1758	ru	1563	zh	790	hi	707	pt	591	sv	560	ko	542	pl	524	cn	473	is	371	hu	360	ar	341	nl	340	xx	319	da	302	cs	284	tr	247	he	215	el	213	th	176	<p>La mayoría de películas son en inglés (46%), seguidas de francés (10%), alemán, español e italiano con porcentajes cercanos al 5% cada uno.</p> <p>Se resalta aquí que la base está focalizada en películas de occidente, pues el número de lenguajes hablados en Asia es pequeño teniendo en cuenta el tamaño de la industria en esa región.</p>
spoken_language_names	Número de películas																																																				
en	28745																																																				
fr	4196																																																				
de	2625																																																				
es	2413																																																				
it	2367																																																				
ja	1758																																																				
ru	1563																																																				
zh	790																																																				
hi	707																																																				
pt	591																																																				
sv	560																																																				
ko	542																																																				
pl	524																																																				
cn	473																																																				
is	371																																																				
hu	360																																																				
ar	341																																																				
nl	340																																																				
xx	319																																																				
da	302																																																				
cs	284																																																				
tr	247																																																				
he	215																																																				
el	213																																																				
th	176																																																				
<p>Compañías productoras</p> <p>Top 25 production_company_names</p>  <table><thead><tr><th>production_company_names</th><th>Número de películas</th></tr></thead><tbody><tr><td>Warner Bros.</td><td>1250</td></tr><tr><td>Metro-Goldwyn-Mayer (MGM)</td><td>1076</td></tr><tr><td>Twentieth Century Fox Film Corporation</td><td>1003</td></tr><tr><td>Universal Pictures</td><td>836</td></tr><tr><td>Columbia Pictures Corporation</td><td>830</td></tr><tr><td>Canal+</td><td>448</td></tr><tr><td>Columbia Pictures</td><td>438</td></tr><tr><td>RKO Radio Pictures</td><td>431</td></tr><tr><td>United Artists</td><td>290</td></tr><tr><td>New Line Cinema</td><td>279</td></tr><tr><td>Walt Disney Pictures</td><td>277</td></tr><tr><td>Touchstone Pictures</td><td>263</td></tr><tr><td>TriStar Pictures</td><td>225</td></tr><tr><td>Mosfilm</td><td>197</td></tr><tr><td>Miramax Films</td><td>188</td></tr><tr><td>France 2 Cinéma</td><td>183</td></tr><tr><td>Centre National de la Cinématographie (CNC)</td><td>168</td></tr><tr><td>Yoko Company</td><td>163</td></tr><tr><td>BBC Films</td><td>142</td></tr><tr><td>Gaumont</td><td>131</td></tr><tr><td>StudioCanal</td><td>126</td></tr><tr><td>Relativity Media</td><td>123</td></tr><tr><td>British Broadcasting Corporation (BBC)</td><td>119</td></tr><tr><td>Walt Disney Productions</td><td>118</td></tr><tr><td></td><td>117</td></tr></tbody></table>	production_company_names	Número de películas	Warner Bros.	1250	Metro-Goldwyn-Mayer (MGM)	1076	Twentieth Century Fox Film Corporation	1003	Universal Pictures	836	Columbia Pictures Corporation	830	Canal+	448	Columbia Pictures	438	RKO Radio Pictures	431	United Artists	290	New Line Cinema	279	Walt Disney Pictures	277	Touchstone Pictures	263	TriStar Pictures	225	Mosfilm	197	Miramax Films	188	France 2 Cinéma	183	Centre National de la Cinématographie (CNC)	168	Yoko Company	163	BBC Films	142	Gaumont	131	StudioCanal	126	Relativity Media	123	British Broadcasting Corporation (BBC)	119	Walt Disney Productions	118		117	<p>Con algunas excepciones (Canal+, Mosfilm, FR2, CNC, BBB, Studio Canal), las mayores productoras de cine se encuentran en los Estados Unidos.</p>
production_company_names	Número de películas																																																				
Warner Bros.	1250																																																				
Metro-Goldwyn-Mayer (MGM)	1076																																																				
Twentieth Century Fox Film Corporation	1003																																																				
Universal Pictures	836																																																				
Columbia Pictures Corporation	830																																																				
Canal+	448																																																				
Columbia Pictures	438																																																				
RKO Radio Pictures	431																																																				
United Artists	290																																																				
New Line Cinema	279																																																				
Walt Disney Pictures	277																																																				
Touchstone Pictures	263																																																				
TriStar Pictures	225																																																				
Mosfilm	197																																																				
Miramax Films	188																																																				
France 2 Cinéma	183																																																				
Centre National de la Cinématographie (CNC)	168																																																				
Yoko Company	163																																																				
BBC Films	142																																																				
Gaumont	131																																																				
StudioCanal	126																																																				
Relativity Media	123																																																				
British Broadcasting Corporation (BBC)	119																																																				
Walt Disney Productions	118																																																				
	117																																																				
<p>País de origen de las productoras</p> <p>Top 25 production_country_names</p>  <table><thead><tr><th>production_country_names</th><th>Número de películas</th></tr></thead><tbody><tr><td>US</td><td>21193</td></tr><tr><td>GB</td><td>4094</td></tr><tr><td>FR</td><td>3940</td></tr><tr><td>DE</td><td>2254</td></tr><tr><td>IT</td><td>2169</td></tr><tr><td>CA</td><td>1765</td></tr><tr><td>JP</td><td>1648</td></tr><tr><td>ES</td><td>964</td></tr><tr><td>RU</td><td>912</td></tr><tr><td>IN</td><td>828</td></tr><tr><td>HK</td><td>596</td></tr><tr><td>SE</td><td>588</td></tr><tr><td>AU</td><td>570</td></tr><tr><td>KR</td><td>495</td></tr><tr><td>BE</td><td>447</td></tr><tr><td>DK</td><td>386</td></tr><tr><td>FI</td><td>383</td></tr><tr><td>NL</td><td>375</td></tr><tr><td>CN</td><td>372</td></tr><tr><td>MX</td><td>329</td></tr><tr><td>PL</td><td>319</td></tr><tr><td>BR</td><td>291</td></tr><tr><td>CH</td><td>254</td></tr><tr><td>AR</td><td>254</td></tr><tr><td>IE</td><td>225</td></tr></tbody></table>	production_country_names	Número de películas	US	21193	GB	4094	FR	3940	DE	2254	IT	2169	CA	1765	JP	1648	ES	964	RU	912	IN	828	HK	596	SE	588	AU	570	KR	495	BE	447	DK	386	FI	383	NL	375	CN	372	MX	329	PL	319	BR	291	CH	254	AR	254	IE	225	<p>Confirmando lo expresado anteriormente, los datos con los que contamos son en su gran mayoría de origen norteamericanos y europeos.</p>
production_country_names	Número de películas																																																				
US	21193																																																				
GB	4094																																																				
FR	3940																																																				
DE	2254																																																				
IT	2169																																																				
CA	1765																																																				
JP	1648																																																				
ES	964																																																				
RU	912																																																				
IN	828																																																				
HK	596																																																				
SE	588																																																				
AU	570																																																				
KR	495																																																				
BE	447																																																				
DK	386																																																				
FI	383																																																				
NL	375																																																				
CN	372																																																				
MX	329																																																				
PL	319																																																				
BR	291																																																				
CH	254																																																				
AR	254																																																				
IE	225																																																				

## **Referencias:**

Bohorquez, Laura Natalia. (2023). Recuperado de:

<https://www.eltiempo.com/cultura/cine-y-tv/netflix-la-historia-del-rechazo-que-termino-en-el-exito-de-la-plataforma-763072#:~:text=La%20compa%C3%B1a%20ha%20invertido%20miles,190%20pa%C3%ADses%20y%2021%20idiomas.>

Banik, Rounak. (2017). Recuperado de:

[https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=movies\\_metadata.csv](https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=movies_metadata.csv)

DataLens. (2018). Recuperado de:

<https://files.grouplens.org/datasets/movielens/ml-latest-small-README.html>

Bagus Murdyantoro (2023). Movie Recommender System: Building Movie Recommendations with Machine Learning. Recuperado de:

<https://medium.com/@bagusmurdyantoro1997/movie-recommender-system-building-smart-movie-recommendations-with-machine-learning-21bfbedb6f3d>

## **Citas:**

1. Fatih Gedikli. The Importance of Recommender Systems: A Key Technology of the World Wide Web. Recuperado de: <https://frontnow.com/post/recommender-systems-key-technology-world-wide-web>
2. Bogdan Walek, Vladimir Fojtik (2020). A hybrid recommender system for recommending relevant movies using an expert system Recuperado de: <https://www.sciencedirect.com/science/article/pii/S0957417420302761>
3. John Koblin (2024). The Little Streamer That Could Recuperado de: [https://www.nytimes.com/2024/08/13/business/media/tubi-movies-tv-streaming.html?unlocked\\_article\\_code=1.G04.C76S.micrXPlvHnCm&smid=url-share](https://www.nytimes.com/2024/08/13/business/media/tubi-movies-tv-streaming.html?unlocked_article_code=1.G04.C76S.micrXPlvHnCm&smid=url-share)