Los gustos en Netflix

¿Qué contenidos gozan de mayor popularidad?

Trabajo hecho por Ivan Kompaniets

Índice

Contexto y audiencia

Contexto

Netflix es una plataforma de streaming en línea que propone series, películas y otros tipos de contenido a sus usuarios a cambio de un pago mensual.

Para tener un ingreso estable, la plataforma necesita fidelizar a los clientes que ya tiene proponiéndoles contenidos de interés y atraer a nuevos usuarios.

Por lo tanto, es esencial entender qué contenidos llaman más la atención de los usuarios y qué características determinan su éxito.

Audiencia

El presente trabajo podría ser útil a la dirección de Netflix para mejorar su oferta de contenidos así como a otras plataformas de streaming que buscan a aumentar su cuota de mercado.

Limitaciones

La principal limitación del presente trabajo es que está basado en datos de 2021. En los últimos dos años los gustos de los usuarios han podido cambiar, por lo tanto habría que actualizar el estudio con nuevos datos.

Preguntas de intéres

Nuestra principal hipótesis es que un contenido con ciertas carácteristicas llama más la atención de los usuarios que los otros contenidos. Para confirmar o refutar la hipótesis vamos a contestar a las siguientes preguntas:

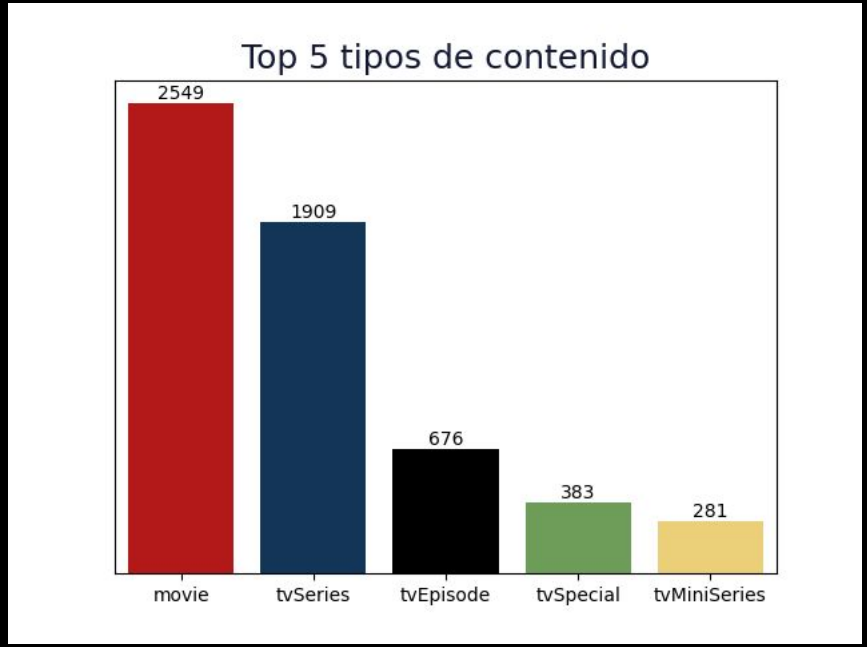
- ¿Hay una relación entre la popularidad y la calidad de los contenidos?
- ¿Qué tipo de contenido es más popular?
- ¿Qué tipo de contenido es de mejor calidad?
- ¿Cuáles son los géneros más populares?
- ¿Cuáles son los contenidos con la mejor nota?

Metadata

Cantidad de contenidos:

6167

País de producción y idioma con más contenidos:



Metadata

Como podemos observar en los gráficos, los dos tipos de contenido más presentes en la plataforma son las series y las películas. También cabe destacar el tipo de miniserie que es parecido a la serie pero tiene un número de episodios más pequeño. Podremos unir los dos tipos al procesar los datos.

En cuanto a la gráfica de líneas nos muestra que la cantidad de contenidos aumenta con los años antes de llegar a una caída en 2021. Puede ser explicada por el hecho que el conjunto de datos fue armado este año y también por el impacto de la pandemia que causó una bajada de producción cinematográfica.

También se puede notar que la cantidad de los contenidos del top 1000 aumenta con la cantidad general pero no es una aumentación tan brusca, por lo cual podemos suponer que muchos de los nuevos contenidos no logran tener una popularidad muy alta.

En las dos siguientes diapositivas serán presentadas las variables incluidas en el conjunto de datos.

Variables

Nombre de variable	Descripción	Tipo de data
imdb_id	Identificación en IMDB	string
title	Nombre del contenido	string
popular_rank	Rango de popularidad (los rangos más pequeños indican mayor popularidad)	int
certificate	Clasificación por edades (no podrá ser usada porque hay muchos valores nulos)	string
startYear	El año de estreno	date
endYear	El año del final (puede ser diferente del año de estreno para las series)	date
episodes	Cantidad de episodios (1 para las películas, otro valor para las series)	int
runtime	La duración (total para las películas y de un episodio para las series)	int
type	Si es una película, serie, miniserie o otro	string

Variables

Nombre de variable	Descripción	Tipo de data
orign_country	País de producción	string
language	Idioma	string
plot	Trama	string
summary	Resumen	string
rating	Rankeo en IMDB	float64
numVotes	Número de votos en IMDB	float64
genres	Géneros (el mismo contenido puede pertenecer a varios géneros)	string
isAdult	Si es contenido adulto (columna suprimida, porque todos los registros tienen el mismo valor)	int64
cast	Reparto	string
image_url	Enlace a la imagen (no será utilizado)	string

Análisis exploratorio de datos

En las siguientes diapositivas será presentado el análisis exploratorio de los datos.

¿Hay una relación entre la popularidad y la calidad de los contenidos?

La calidad de una película a menudo es bien reflejada por la nota que ella recibe en la página IMDB.
Para ver si la popularidad de la película depende de su nota promedio, dividimos los datos en 12 grupos y calculamos la nota promedio por grupo.
A la izquierda están los grupos con la popularidad más altas y a la derecha los grupos menos populares.
Podemos ver que hay una relación entre la popularidad y la calidad. La nota promedio va bajando a medida que baja la popularidad.
La única excepción es el grupo 12 que tiene una nota un poco más alta que el grupo anterior.

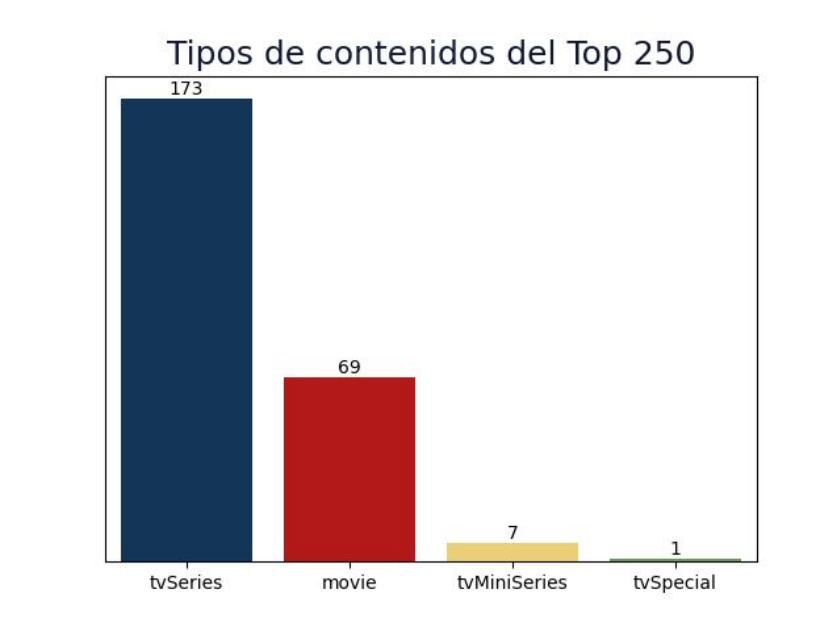
¿Qué tipos de contenido son más populares?

Para identificar los tipos de contenido más populares nos enfocamos en los 250 contenidos más populares.

Podemos observar que el tipo de contenido más popular en el gráfico son las series.

Es un resultado interesante porque las películas son el contenido más común en el set de datos en general.

Significa que la mayor popularidad de las series no es una simple coincidencia pero hay una tendencia detrás.



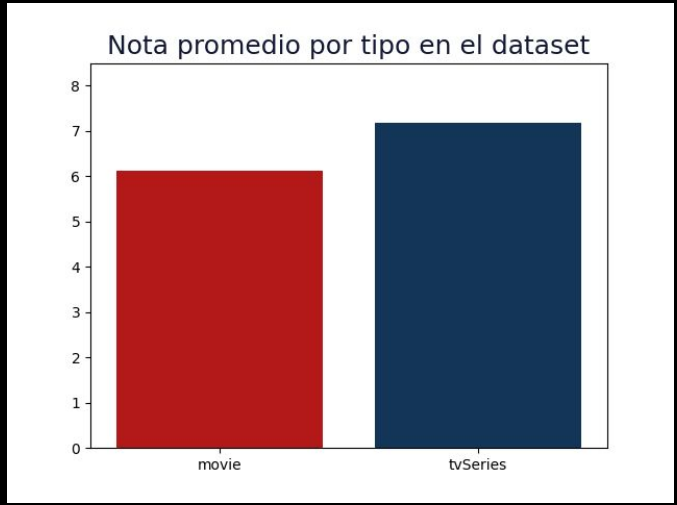
¿Qué tipo de contenido es de mejor calidad?

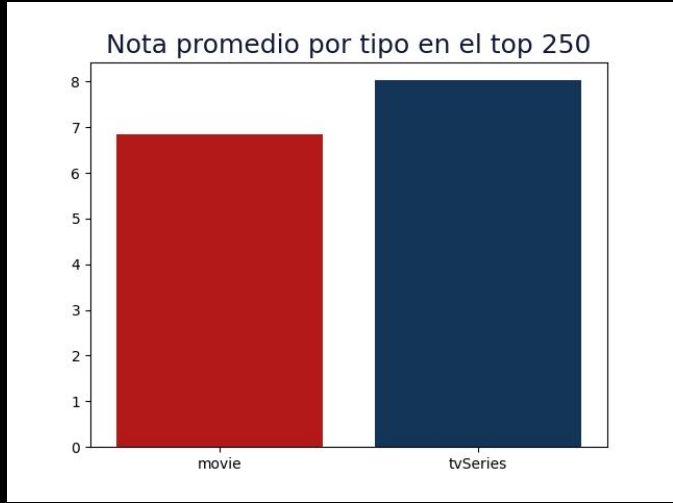
Como vimos que las películas son menos presentes entre los 250 contenidos más populares, nos preguntamos si no hay también una diferencia cualitativa entre los dos tipos de contenido.

Cómo podemos observar, la respuesta es sí!

Podemos ver en los dos gráficos que la calidad de las series es superior en todo el set de datos y en el top 250.

Además, si nos fijamos en los valores, las notas promedio en el top 250 son más altas que en el set de datos entero, lo que demuestra otra vez la relación entre la calidad y la popularidad.



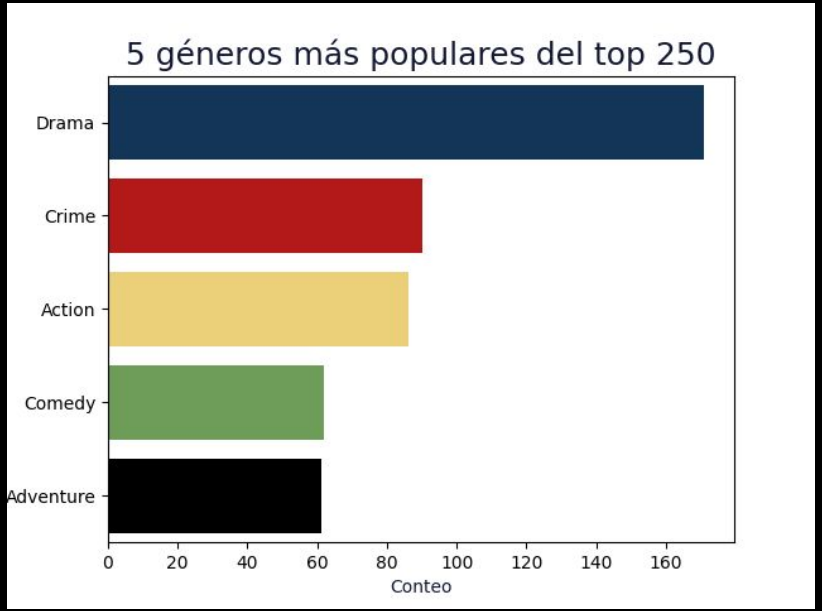


¿Cuáles son los géneros más populares?

De nuevo nos fijamos en los géneros más populares en el top 250 y obtuvimos el presente gráfico.

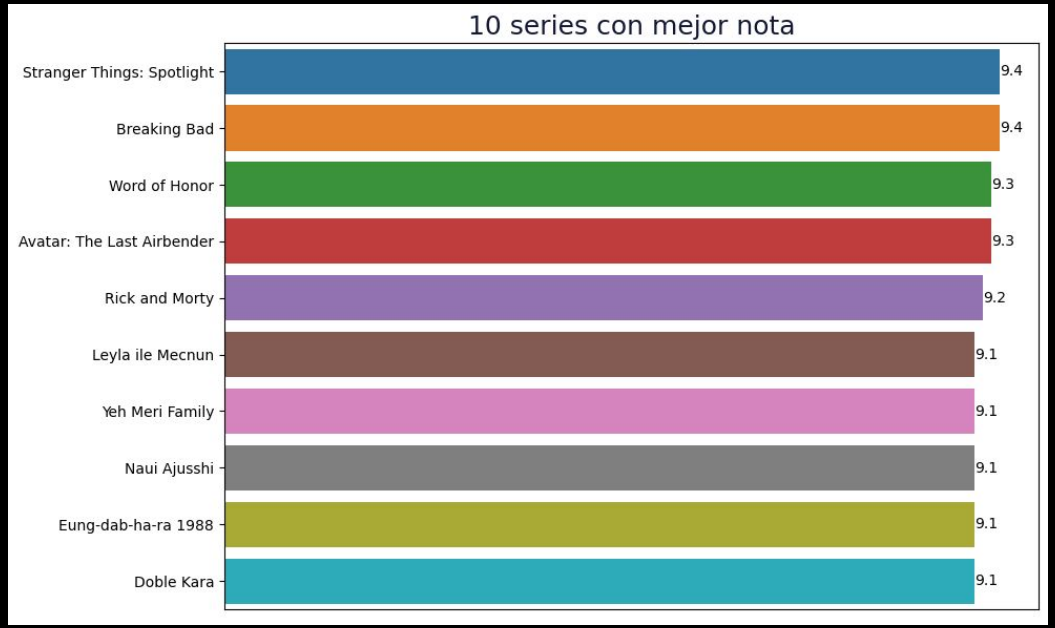
Un detalle interesante es que un contenido puede pertenecer a varios géneros a la vez y los géneros que vemos en el gráfico pueden ser compatibles. Así podemos agrupar los géneros “drama” y “crimen” así como “comedy” y “adventure”.

Estos son los dos grandes tipos de contenidos más rebuscados en Netflix.



¿Cuáles son las series con la mejor nota?

En el presente gráfico podemos ver las series con las mejores notas pero como vemos en la tabla no siempre tienen el mejor índice de popularidad.
Por lo tanto, la calidad no es el único criterio pero también son importantes otras características como el género.



	Título	Rango de popularidad
0	Breaking Bad	15.000000
1	Rick and Morty	60.000000
2	Avatar: The Last Airbender	138.000000
3	Word of Honor	595.000000
4	Eung-dab-ha-ra 1988	1130.000000
5	Leyla ile Mecnun	1194.000000
6	Naui Ajusshi	1226.000000
7	Yeh Meri Family	1418.000000
8	Stranger Things: Spotlight	3673.000000
9	Doble Kara	3904.000000

Insights y recomendaciones

Insights

- Hemos observado que la popularidad de los contenidos de Netflix depende de varias características
- Una característica importante es la nota de IMDB que refleja la calidad del contenido. Sin embargo, no es el único factor.
- Otro factor importante es el género de contenido. Por ejemplo, los usuarios de Netflix parecen ser particularmente interesados por los dramas criminales y las películas de aventura cómicas.

Recomendaciones

- Para mejorar su oferta de contenidos, Netflix debería seguir proponiendo nuevos contenidos (sobre todo, series) en los géneros preferidos de los usuarios.
- Otra estrategia podría consistir en tratar de mejorar la calidad de las películas accesibles en la plataforma para atraer a otros tipos de usuarios.
- Finalmente, podría ser útil proponer series y películas en los géneros menos populares para atraer a usuarios con intereses más diversos.

Selección de modelo

El objetivo del presente trabajo es desarrollar un modelo de agrupación que permita destacar varios tipos de contenidos dentro del conjunto de datos y hacer sugerencias de contenido a los usuarios.

Para la tarea de agrupación crearemos modelo de **KMeans** y **Agglomerative Clustering**. Estos modelos podrán ser utilizado para llevar a cabo un análisis más profundo del conjunto de datos e identificar relaciones entre las entidades.

Luego, necesitaremos encontrar los contenidos más parecidos entre sí para poder hacer sugerencias basándonos en los gustos de los usuarios. Para hacerlo, construiremos un modelo de **NearestNeighbors** que permite buscar los registros más cercanos a un registro escogido.

Finalmente, escribiremos una función para extraer el id, el título, el año y el género de cada sugerencia del conjunto de datos inicial.

Preprocesamiento del conjunto de datos. Selección de las variables

Primero, vamos a seleccionar las variables que usaremos para la construcción del modelo.

Incluiremos las siguientes variables:

- **startYear**: año de estreno. Modificaremos la variable para representar períodos más bien que años exactos
- **type**: tipo de contenido (serie, película u otro)
- **orign_country:** las variables **orign_country** e **language** contienen información parecido. Sin embargo, la variable “país” es más precisa ya que los contenidos en español o inglés pueden ser producidos por muchos países diferentes.
- **rating:** la nota que recibió un contenido en IMDB. Es importante porque refleja su calidad. Por otra parte, la variable “numVotes” que indica el número de notaciones que recibió no será incluida porque no nos dice nada sobre el valor intrínseco de un contenido.
- **genres:** es una variable esencial porque los usuarios a menudo van a escoger películas en función del género.
- **imdb_id:** refleja la popularidad del contenido.

Selección de las variables

Variables **excluidas** del análisis:

- **title:** el título del contenido no puede ser usado en el análisis.
- **episodes, runtime:** las dos describen la duración de un contenido. Las vamos a utilizar para crear una variable **duración total**.
- **plot, summary:** variables textuales. Pueden ser utilizadas pero necesitan procesamiento.
- **cast:** una variable difícil de procesar que contiene listas de actores. Como hemos visto en la parte del análisis exploratorio de datos, ciertos actores no determinan necesariamente el éxito de una serie o película. Además el uso de la variable aumentaría de forma significativa la dimensionalidad de los datos.
- **image_url:** es un enlace a la portada del contenido. No vamos a procesar datos visuales.

Preprocesamiento de variables

La variable **genres** es esencial para nuestro análisis ya que determina las elección de contenidos por los usuarios. Sin embargo, en nuestro conjunto de datos, cada registro puede corresponder a varios géneros. Para solucionar el problema, usamos el algoritmo de One Hot Encoding para codificar cada género. Si un contenido pertenece a un género recibe 1 en la columna correspondiente y si no, recibe 0.

Como era de esperar, los contenidos pueden pertenecer a uno, dos o tres géneros.

También codificamos las variables textuales **type** y **orign_country.** A la diferencia de los géneros, cada registro tiene un valor único en estas columnas, por lo tanto, usamos el algoritmo de Label Encoding que representó cada valor numérico por un valor textual.

Creación de nuevas variables

Basándonos en las variables existentes, agregamos dos nuevas variables **total_rt** (duración total) y **period** (periodo).

La duración total es una variable numérica que fue obtenida como suma de las variable **episodio** y **runtime.** Como resultados, las series que consisten de varios episodios tienen una duración total más larga que las películas y otros contenidos de un episodio.

En cuanto al periodo es una variable categórica obtenida a partir de la variable **startYear.** A menudo los espectadores no van a decidir si los interesa un contenido según su año de producción exacto pero más bien si es un contenido nuevo o viejo. Por lo tanto, destacamos 4 grupos de contenidos que son:

- Grupo 0: contenidos viejos de antes de 2000
- Grupo 1: contenidos producidos entre 2000 y 2010
- Grupo 2: contenidos producidos entre 2010 y 2020
- Grupo 3: contenidos muy nuevos estrenados después de 2020

Luego de crear las nuevas variables, aplicamos el algoritmo de Standard Scaler par estandardizar las variable numéricas **rating** y **total_rt.**

Selección de modelos

Para realizar la agrupación de los datos, seleccionamos los siguientes algoritmos:

- Agglomerative Clustering,
- Birch,
- HDBSCAN,
- OPTICS,
- KMeans,

La principal dificultad de nuestro conjunto de datos fue que contiene datos sobre contenidos muy diversos, así que era de esperar que obtengamos muchos grupos al hacer la agrupación.

Otro riesgo era que al usar un algoritmo con un número de grupos determinado, como KMeans, la mayoría de los datos termine en un solo grupo por no haber sido clasificada. Esto bajaría de forma significativa la utilidad del modelo.

Para evaluar los modelos usamos las métricas Silhouette Score, Davies-Bouldin score y Calinski-Harabasz score. También calcularemos el tamaño promedio de los clusters y la desviación estándar de los tamaños.

Selección de hiperparámetros

Agglomerative Clustering

El primer algoritmo que usamos fue Agglomerative Clustering porque permite agrupar los datos por distancia sin determinar un número de grupos al instanciar el modelo.

Escogimos el método de agrupación de Ward y fijamos el umbral de distancia en 200. Un valor de distancia demasiado alto reduciría el número de grupos pero podría producir algunos grupos demasiado grandes que serían poco útiles para el análisis.

Al contrario, un valor de distancia muy bajo produciría una gran cantidad de grupos muy pequeños que serían difíciles de analizar.

El umbral de 200 permite tener métricas de agrupación bastante altas sin crear grupos de tamaños desproporcionados.

Selección de hiperparámetros

HDBSCAN

Es un algoritmo de agrupación que puede clasificar ciertos registros como ruido. Luego, tal como Agglomerative Clustering genera una cierta cantidad de grupos.

Como hiperparámetro fijamos el tamaño máximo de un grupo en 500 para impedir el algoritmo de clasificar la mayoría de los datos en el mismo grupo.

OPTICS

Es otro algoritmo que elimina el ruido de los datos. Los resultados podrán ser comparados a los de HDBSCAN.

Como hiperparámetro definimos en 5 el parámetro min_samples que corresponde al número mínimo de registros para crear un grupo.

Al **evaluar** estos dos modelos, excluimos los datos que fueron clasificados como ruido para concentrarnos sobre el resto de los grupos.

Selección de hiperparámetros

KMeans

Es un modelo que agrupa los datos en un cantidad de grupos predeterminada. Para escoger la mejor cantidad de grupos usamos el metodo de codo que demostró que la cantidad óptima de grupos es 15.

BIRCH

BIRCH es un algoritmo de agrupación jerárquica que usamos con el hiperparámetro n_clusters para generar una cantidad de grupos determinada. Lo fijamos en 15, el mismo número de clusters que usamos para KMeans. Nos permitirá comparar los dos modelos.

Es importante ver si los grupos son bien equilibrados o hay un grupo más grande que el resto. Lo último puede ocurrir si una parte de los datos que no corresponde a los grupos ya formados es tirada en un grupo de sobra. En este caso, el grupo más grande sería parecido al grupo de ruido formado por los algoritmos de OPTICS y HDBSCAN.

Rendimiento de los modelos

Para evaluar los modelos, usamos las siguientes métricas:

- **Silhouette score**: muestra si los grupos formados están bien separados entre sí. Los valores pueden situarse entre -1 y 1. Cuanto mayor sea la métrica, mejor es la agrupación.
- **Davies-Bouldin score**: evalúa la similaridad de los clusters. Un valor bajo significa que los grupos son bien delimitados y fáciles de distinguir.
- **Calinski-Harabasz score**: evalúa la cohesión dentro de cada cluster. Al comparar los modelos, se debe elegir el que tiene el valor más alto.
- **Tamaño mediano**: calculamos el tamaño mediano de los grupos formados. Junto que la desviación estándar, permite evaluar si los grupos tienen tamaños parecido o hay un desequilibrio
- **Tamaño del cluster más grande**: es útil para los modelos con un número de grupos limitado. En el caso de los modelos de HDBSCAN y OPTICS, el grupo más grande es el grupo de ruido.
- **Desviación estándar de los tamaños**: comparamos la métrica con el tamaño mediano para saber si los son grupos son del tamaño parecido o tenemos grupos muy grandes y muy pequeños.
- **Número de grupos**: es útil para los algoritmos de Agglomerative Clustering, HDBSCAN y OPTICS que tienen un número de grupos determinado.

Tabla de resultados

Creamos una función para evaluar los modelos iniciados y agregar las métricas en una tabla.

	Silhouette	Davies-Bouldin	Calinski-Harabasz	median_size	size_std	largest cluster / noise	n_clusters
AgglomerativeClustering	0.379629	0.827888	19833.465545	297	397.914346	1321	17
HDBSCAN	0.302374	0.922213	4532.665706	9	32.594940	3884	118
OPTICS	0.467651	0.772198	4278.193785	7	5.197511	4918	146
KMeans	0.400343	0.791186	21466.894403	305	420.149476	1462	15
Birch	0.375625	0.789972	20710.588950	338	413.520151	1392	15

Análisis de los resultados

Al comparar los modelos podemos constatar que el algoritmo de OPTICS tiene el mejor Silhouette score y el mejor Davies-Bouldin score. Sin embargo, tanto OPTICS cono HDBSCAN clasificaron más de la mitad de los datos como ruido: 4918 y 3884 datos respectivamente. También podemos ver que separaron el resto de los datos en muchos grupos pequeños. Eso por un lado explica el valor bajo del Calinski-Harabasz score de ambos modelos y denota por otro lado la gran diversidad de los datos.

En cuanto al algoritmo de Agglomerative Clustering que no tenía una cantidad de clusters predeterminada, dividió los datos en 17 grupos de tamaños muy variados, lo que se puede deducir de la gran diferencia entre el grupo más grande y el grupo mediano, asi como de la desviación estándar importante de los tamaños.

Finalmente los algoritmos de KMeans y BIRCH tienen métricas bastante parecidas. El modelo de KMeans tiene un Silhouette score y Calinski-Harabasz score levemente más altos, mientras que el model de BIRCH llegó a un mejor valor del Davies-Bouldin score.

Basándonos en las métricas, elegiremos KMeans como nuestro algoritmo. Primero, formó un número de grupos menos grande que los algoritmos de OPTICS y HDBSCAN lo que nos permite destacar tendencias más generales en los datos. Segundo, obtuvo mejores métricas que los modelos de agrupación jerárquica, es decir OPTICS y HDBSCAN.

Modelo final

Al volver a correr el modelo, obtuvimos las siguientes métricas:

	Resultado
Silhouette score	0.422180
Davies-Bouldin score	0.781554
Calinski-Harabasz score	22027.153799
Tamaño de grupo mediano	305.000000
Tamaño de grupo más grande	1693

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Tamaño	32	1693	674	403	55	103	80	335	1074	124	567	29	392	305	301

Podemos observar que tenemos dos grupos grandes de más de mil unidades y 4 grupos pequeños de menos de 100 unidades. Ahora procederemos al análisis de los grupos obtenidos.

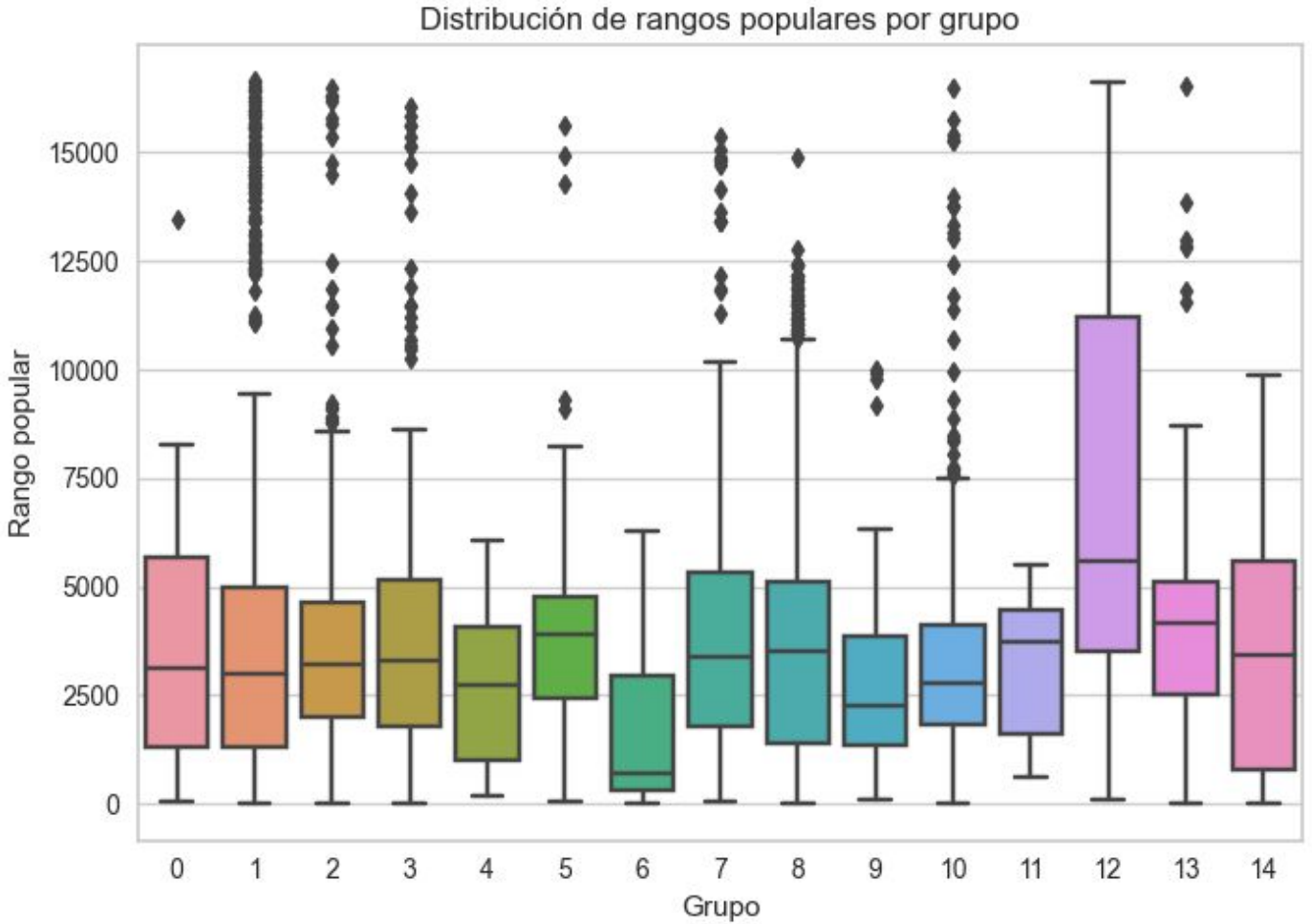
Análisis de los grupos

Para llevar a cabo un análisis de los grupos, agregamos las etiquetas de grupo al conjunto de datos inicial que contiene la información sobre los rangos populares de los contenidos. Luego generamos un diagrama de caja para destacar los grupos con un rango popular mayor. La gráfica obtenida se puede ver en la siguiente diapositiva.

En la gráfica, se destaca inmediate el grupo número 6 que tiene un mejor rango popular que el resto de los grupos (recordamos que los rangos más bajos denotan una mayor popularidad).

Los dos otros grupo con un rango popular promedio elevado son los 4 y 9. Utilizando la tabla de los tamaños podemos notar que son grupos relativamente pequeños. El grupo 6 contiene solamente 80 contenidos. El grupo 4 contiene 55 y el grupo 9 contiene 124.

Analizaremos cada uno de los grupos en detalle para identificar sus características particulares.

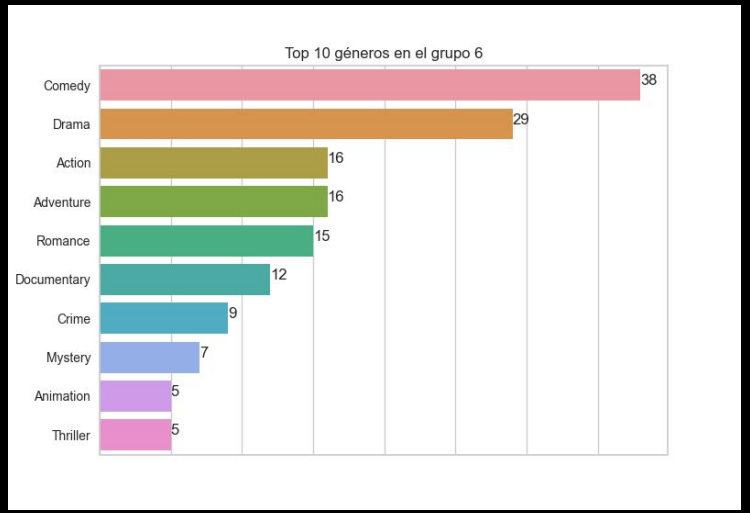

Distribución de rangos populares por grupo
Rango popular
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Grupo

Análisis del grupo 6

Realizamos un análisis de los contenidos del grupo 6 por país, tipo y género. Podemos notar que el grupo contiene tanto películas como series. Sin embargo, son contenido producidos sobre todo en dos países: Reino Unido y Estados Unidos. También se destacan ciertos géneros en el gráfico de géneros. Podemos constatar una predominancia de comedia y dramas. Otros géneros importante son acción y aventura. Esto parece confirmar nuestras conclusiones del análisis de los datos sobre los géneros más populares.

Los 3 países del grupo 6

	Conteo
United States	65
United Kingdom	13
Sweden	2



Tipos de contenido en el grupo 6

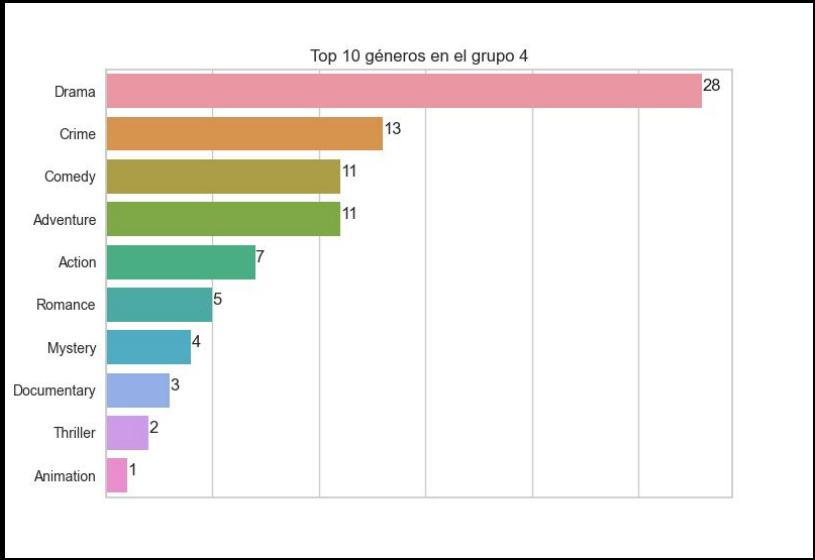
	Conteo
movie	37
tvSeries	27
tvSpecial	7
video	5
tvMovie	2
tvMiniSeries	1
tvEpisode	1

Análisis del grupo 4

Este grupo tiene un tipo de contenido sobresaliente que son películas de género drama producidos en los Estados Unidos. El segundo género más representado es crimen, por lo tanto podemos suponer que se trata de dramas criminales.
Por otro lado, cabe señalar que también son presentes comedias y aventuras, mientras los dos otros países incluidos son el Reino Unido y Suecia que también podemos ver en el precedente grupo.

Los 3 países del grupo 4

	Conteo
United States	41
United Kingdom	7
Sweden	7



Tipos de contenido en el grupo 4

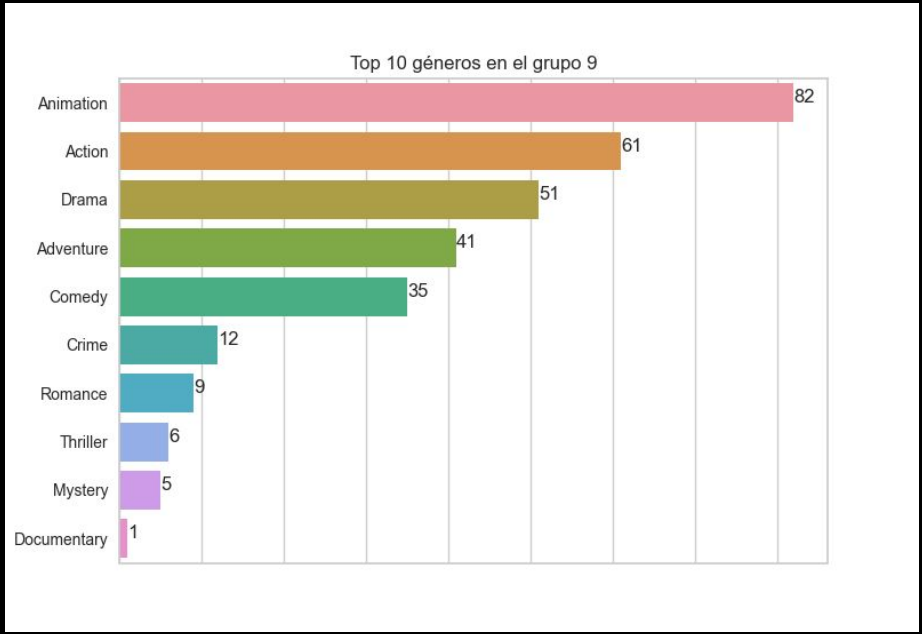
	Conteo
movie	48
tvSeries	2
short	2
tvSpecial	2
tvMiniSeries	1

Análisis del grupo 9

En el presente grupo también se destaca un tipo de contenido en particular que son películas y series de animación japonesas, es decir animés, que pueden ser de acción, de aventura o dramas.
Es una nueva categoría de contenido importante que no hemos identificado en la fase del análisis exploratorio de los datos.

Top 5 países en el grupo 9

	Conteo
Japan	83
India	16
Ireland	5
Netherlands	4
New Zealand	3



Tipos de contenido en el grupo 9

	Conteo
tvSeries	55
movie	51
tvMiniSeries	8
tvEpisode	5
video	2
tvSpecial	2
short	1

Conclusión

El análisis de los datos y el modelo de agrupación nos permitieron identificar categorías de contenido que tienen más popularidad en Netflix. Una característica extremadamente importante es el género. Los contenidos del género drama que pueden ser dramas de crimen y dramas de acción son los más populares. El segunda categoría de género más demanda incluye las comedias que puede ser de aventura o comedias románticas. Finalmente, una última categoría sobresaliente que pudimos identificar usando un algoritmo de agrupación de datos son las películas de animación japonesas, es decir animés.
Luego, se identificó otra característica significativa que es el país de producción de un contenido. Gozan de mayor popularidad los contenidos producidos en los Estados Unidos y en el Reino Unido. Puede ser explicarse por el hecho que es una plataforma estadounidense, por lo tanto, los contenidos locales son los que atraen más público. Dicho esto, se pudieron destacar dos otros países productores con un buen potencial que son Japón, un gran productor de animés, y Suecia.
Finalmente, los tipos de contenidos más representados en la plataforma son las películas y las series y tienen mayor popularidad que los otros tipos. Sin embargo, el algoritmo de agrupación no hizo una separación clara entre estos dos tipos, ya que sólo un grupo de los tres es mayoritariamente compuesto de películas. Así que podemos concluir que el género y el país de producción son criterio de selección más pertinentes para el usuario que el hecho de ser una serie o una película.

Conclusión

Basándonos en el análisis realizado podemos hacer las siguientes sugerencias para la empresa:

- Las películas son un tipo de contenido tan importante como las series, sin embargo tienen una calidad promedia más baja que las series. Al invertir en películas de buena calidad, se podrían atraer nuevos grupos de usuarios en números más grandes.
- Los géneros privilegiados por los usuarios se pueden dividir en dos grupos: los dramas que entre otros pueden ser de crimen o acción, así como las comedias románticas y las comedias de aventura. Proponer más contenidos de este tipo podría aumentar la satisfacción de los usuarios.
- Otra categoría con un buen potencial para la plataforma son los animés japoneses. Una selección más amplia de contenidos de este tipo volvería la plataforma más atractiva para los usuarios interesados en ellos.
- Finalmente, los países de producción más representados en la plataforma son los Estados Unidos y el Reino Unido. Son los contenidos de estos países que llaman más la atención de los usuarios. Sin embargo, dos otros países cuyos contenidos tienen un buen potencial son Japón y Suecia. La agregación de una mayor cantidad de contenidos producidos por ellos permitiría diversificar la oferta de contenidos en la plataforma.

Sugerencias de contenidos

Finalmente, creamos una función para hacer sugerencias de contenidos en base a un contenido gustado.

Primero, modificamos la base de datos, dejando sólo las columnas relativas al tipo, idioma, nota, género y periodo, para eliminar el ruido.

Luego usamos el modelo de Nearest Neighbors para seleccionar los contenidos más parecidos al seleccionado y extrajimos de la base de datos inicial las informaciones sobre la identificación, el título, el tipo, el país y el género de cada contenido sugerido por el modelo.

Al pedir sugerencias para alguien a quien le gustó el Señor de los Anillos y la serie surcoreana Seutateueob, obtuvimos los resultados que se pueden ver en la siguiente diapositiva.

Se puede constatar que para el Señor de los anillos el modelo encontró las otras películas de la saga, mientras que para Seutateueob sugirió otras series producidos en el mismo país.

Basándonos en este testeo, podemos concluir que el modelo logra hacer sugerencias pertinentes.

Ejemplos de sugerencias

Señor de los anillos:

```
Qué película te ha gustado? lord of the rings

Si te gustó "The Lord of the Rings: The Fellowship of the Ring", te recomendamos:

tt0167260 - The Lord of the Rings: The Return of the King - movie - New Zealand - Action,Adventure,Drama
tt0167261 - The Lord of the Rings: The Two Towers - movie - New Zealand - Action,Adventure,Drama
tt1663202 - The Revenant - movie - United States - Action,Adventure,Drama
tt0416449 - 300 - movie - United States - Action,Drama
tt0244365 - Enterprise - tvSeries - United States - Action,Adventure,Drama
```

Seutateueob:

```
Qué película te ha gustado? Seutateueob

Si te gustó "Seutateueob", te recomendamos:

tt12182904 - Once Again - tvSeries - South Korea - Comedy,Drama,Romance
tt12451520 - Saikojiman Gwaenchanha - tvSeries - South Korea - Comedy,Drama,Romance
tt12525622 - Was It Love - tvSeries - South Korea - Comedy,Drama,Romance
tt12401208 - Yashiknamnyeo - tvSeries - South Korea - Comedy,Drama,Romance
tt2782216 - Eung-dab-ha-ra 1997 - tvSeries - South Korea - Comedy,Drama,Romance
```

Futuras líneas

El análisis del conjunto de datos y el desarrollo del modelo de agrupación nos permitieron identificar las categorías de contenido más prometedoras y crear una función para proponer sugerencias de contenido al usuario. Sin embargo, el presente análisis puede ser ampliado de las siguientes forma:

- Se podrían utilizar los datos textuales de la columna “plot”.
- Se podrían extraer datos más completos para otras variables como por ejemplo “certificate”, lo que requeriría el uso de la API de IMDB que no es de acceso libre.
- Se podría actualizar el conjunto de datos agregando contenidos más recentes de Netflix.
- Se podría seguir mejorando la función de sugerencias para, por ejemplo, proponer contenidos basándose en varias apreciaciones del usuario en vez de una sola.
- Se podría desarrollar un modelo para identificar contenidos prometedores en conjunto de datos más grandes que contendrían los datos sobre películas que no están en Netflix