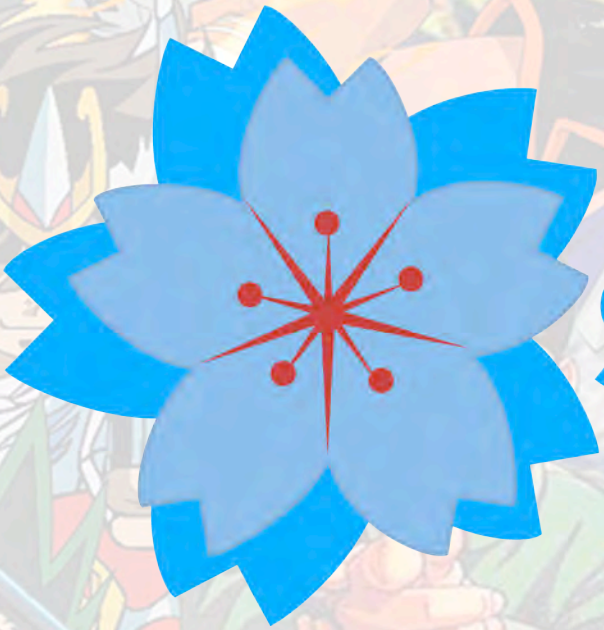


SakuraDat

Sistema de recomendación para animes

ÍNDICE

SakuraDat

Santiago, 20 de mayo de 2020

Señores:

Plataforma Anime

Representantes, Andrea Villaroel
Gonzalo Seguel

EL PROYECTO

1.- Contexto

Los sistemas de recomendación están presentes en todas partes en Internet. Y su propósito de recomendación es sugerir a los usuarios algo basado en su interés o historial de uso.

En la industria del entretenimiento, esto lo hacían muchos empleados supuestamente tienen tanto conocimiento sobre la música que realmente impiden que los clientes compren la música que quieren comprar. Menciono esto porque antes de tener Netflix y Amazon y YouTube, los seres humanos reales en la carne eran los sistemas de recomendación personalizados más cercanos que teníamos.

El dueño de la tienda de discos que sabe lo que le gusta y le recomienda el nuevo álbum, el servidor del restaurante que ha probado todo en el menú y sabe exactamente lo que quiere en función de lo que tenía antes, o el extraño al azar en la calle quien le dice la forma más rápida y fácil de llegar al lugar que está buscando, todos estos son sistemas recomendados en la carne, y muy efectivos.

El problema es que no escalan. Y no podía escalar hasta que llegó Internet con cosas como Google. E incluso entonces no había forma de evaluar de manera efectiva el proceso de recomendación hasta la llegada de la ciencia de datos y su capacidad para manejar una gran cantidad de datos.

No necesita una investigación de mercado para averiguar si un cliente está dispuesto a comprar en una tienda donde obtiene la máxima ayuda para explorar el producto correcto. También es mucho más probable que vuelvan a esa tienda en el futuro. Para tener una idea sobre el valor comercial de los sistemas de recomendación: hace unos meses, estimó Netflix, que su motor de recomendación tiene un valor anual de mil millones de dólares.

Tras esta introducción vemos como los algoritmos es seguro decir que los motores de recomendación de productos mejorarán con el uso del aprendizaje automático. Y cree un proceso mucho mejor para la satisfacción y retención del cliente.

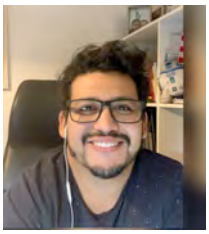
2.- Equipo y organización

Líder del Equipo, Documentador, Producción Web Susana Arce Silva

Primera responsable del proyecto. Su trabajo es reportar el estado de este en la coordinación de actividades de los miembros del equipo, como su seguimiento y el apoyo incondicional a cada uno de ellos. En la labor de producción, el objetivo es realizar una web donde se visualizará el funcionamiento del modelo.

Además, de la formalización de documentos que describen y explican los procesos involucrados en el proyecto.

Ingeniero de Control de Calidad, Ingeniero de Datos y Análisis



Rodrigo Pereira Jofré: como primera labor, es el control de calidad del proyecto y su trabajo consiste en la verificación, validación de los datos extraídos, hacer prueba de los algoritmos de procesamiento de datos y verificación de modelos predictivos.
: como primera labor, es el control de calidad del proyecto y su trabajo consiste en la verificación, validación de los datos extraídos, hacer prueba de los algoritmos de procesamiento de datos y verificación de modelos predictivos.

Además, se enfoca en el desarrollo de algoritmos de transformación y carga de los datos, encontrar tendencias y relaciones en los datos para entrega de un *dataset* para el proceso de modelamiento.

Ingeniero de modelamiento, Documentadora, Producción:

Fabiola Aravena Ortega

Como primera labor, se encarga de la normalización, escalamiento y estandarización de los datos, implementación y calibrado de modelos predictivos, así como la definición de métricas de desempeño.

Además, comparte labor con la líder del equipo en la documentación del proyecto y producción web.

3.- Problema

Nuestro sponsor posee un catálogo de contenido audiovisual de 12.294 títulos de anime, requiere poder efectuar recomendaciones de géneros para los usuarios en base a otros animes que han sido de su gusto.

Actualmente tienen una página web con registros de usuarios, desean integrar algoritmos de recomendación para que mejorar la experiencia y el flujo de usuarios en el servicio que ofrecen y dar un valor agregado frente a la acelerada competencia. Pues lo que desean es hacer un lanzamiento de la web porque debido a estudios de marketing que hicieron saben que han perdido usuarios de animes.

Nuestra motivación es recomendar, sugerir e influir en las decisiones de los usuarios forma parte de la nueva realidad que extrae valor de las informaciones de modo más potente a través de algoritmos y los recursos de Machine Learning. Así como una empresa conoce a sus clientes y sus gustos, el reconocimiento automático de sus perfiles seguido de recomendaciones adecuadas a los mismos permitiría también personalizar la comunicación con los mismos, incluso de forma online, tal y como haría un empleado de forma presencia.

4.-Objetivo

Tenemos objetivo crear un sistema para hacer recomendaciones basado en los géneros de film de animes que solo tiene en cuenta el historial de los ítems elegidos por el mismo, sin usar ningún otro dato adicional de carácter personal. Además, tenemos los siguientes objetivos específicos (OE1):

OE1. Establecer la representación necesaria de los usuarios, género, rating para hacer posible el análisis según los algoritmos utilizados.

OE2. Crear un perfil principal (usuario registrado), usando técnicas de afinidad para descubrir grupos de usuarios similares en cuanto al historial de sus géneros, para, a continuación, generar reglas de asociación en cada grupo por clusterización que serán usadas para determinar las recomendaciones que se hacen a los usuarios.

OE3. Crear un perfil secundario (usuario nuevo), aplicando para aquellos grupos para los que no se hayan generado reglas de asociación.

Tenemos Un sistema de recomendación basado en perfiles generados por agrupamiento y asociaciones

II Análisis de los datos y Modelo de Negocios

1.- Investigación y Desarrollo de Mercado

a.- Hipótesis:

Es posible agrupar los usuarios que han visto o evaluado positivamente algún título de anime, determinando su género más afín, con otros similares para entregar una recomendación y mejorar su experiencia.

b.- Propuesta:

Aplicación web en donde el usuario nuevo seleccione de una lista de Anime propuestos, que han sido de su gusto, y en base a ello se le entregue una lista de otras alternativas que vayan a fin con sus preferencias. En caso de usuario existente, la recomendación se realizará en función de su registro histórico.

2.- Datos

2.1.- Descripción de los Datos Existentes:

El Sponsor ha facilitado 2 data sets en formato ".csv", uno con el listado de los títulos existentes y otro con las evaluaciones de distintos usuarios:

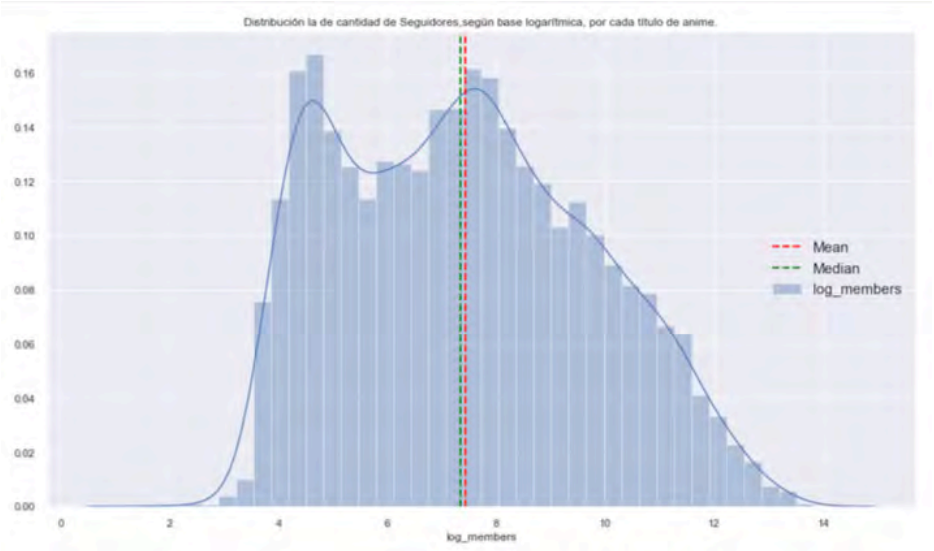
Respecto al data set "anime.csv", contiene 12.294 registros, cada uno de los cuales corresponde a un título de anime diferente, con sus características básicas propias. Esta data set será denominado: **df_titulos**, con los siguientes atributos:

- anime_id: es un código único por cada título de anime
- name: título del anime
- genre: subgéneros a los que pertenece
- type: formato
- episodes: aplica sólo en caso formato "Serie"
- rating: evaluación global del anime
- members: seguidores del título

Se complementarán los atributos de cada título de anime con la *url* de caratula del anime, mediante la conexión con la API https://imdb-api.com para poder utilizarlo en el paso final a producción. Los datos faltantes se completaron, no realizan ingestan destacables en los datos.

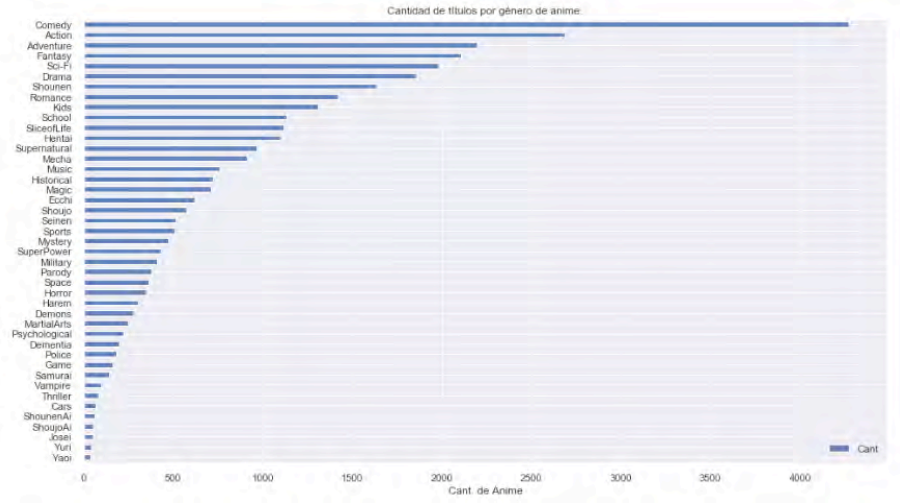
Se observa que existen títulos ampliamente populares, por sobre el resto:

Por tal motivo, este atributo se utilizará mediante su media logarítmica:



Como insight, observamos que existen 2 peaks en la cantidad de títulos, lo cual nos lleva a considerar que existe un número considerable de anime que presentan una popularidad limitada, siendo probablemente producciones dirigidas a un nicho de usuarios específico.

Dado que los títulos pertenecen a diferentes géneros, se procesó dicho atributo y se construyeron variables dummies para cada uno de ellos, obteniéndose la siguiente frecuencia:

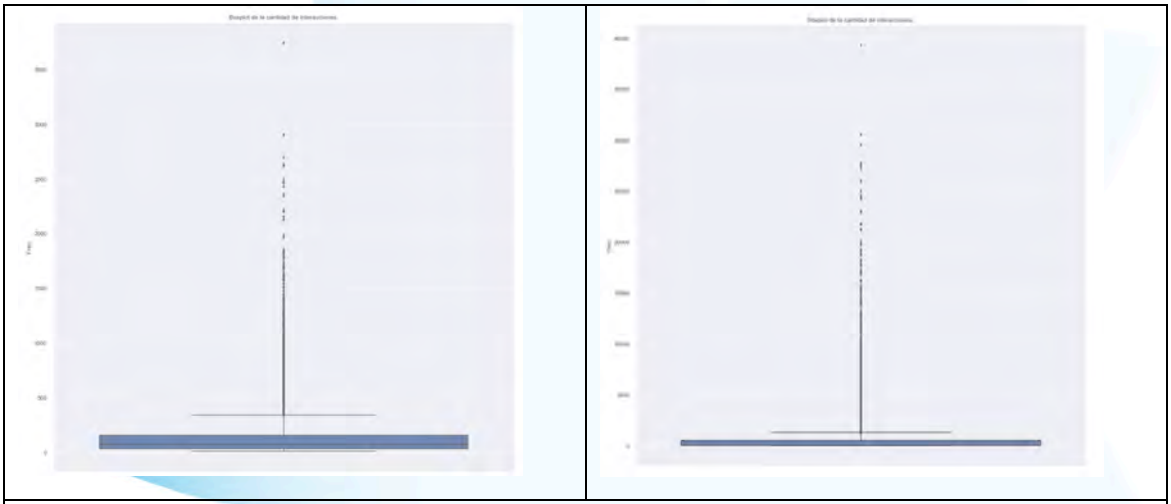


Respecto al dataset "rating.csv", contiene 7.813.737 registros, cada uno de los cuales corresponde a una interacción puntual entre un usuario determinado y un título de anime específico, desde el registro que el usuario vio un anime hasta la evaluación de este.

2.2.- Procesamiento de los datos:

Para facilitar el trabajo del área de modelamiento, se crearon 2 indicadores específico para clasificar la relación entre usuarios y anime.

Se observa varios títulos de anime y usuarios con comportamiento anómalos:



A la izquierda se observa la distribución de la cantidad de interacción por usuario, a la derecha se presenta los mismos datos por título de anime

Por lo que se realiza una selección de los registros a utilizar, establecimiento un mínimo de 10 interacciones para considerar a un usuario como registrado, limitando sus interacciones a 344 y a 1.287 para el caso del análisis de los títulos de anime.

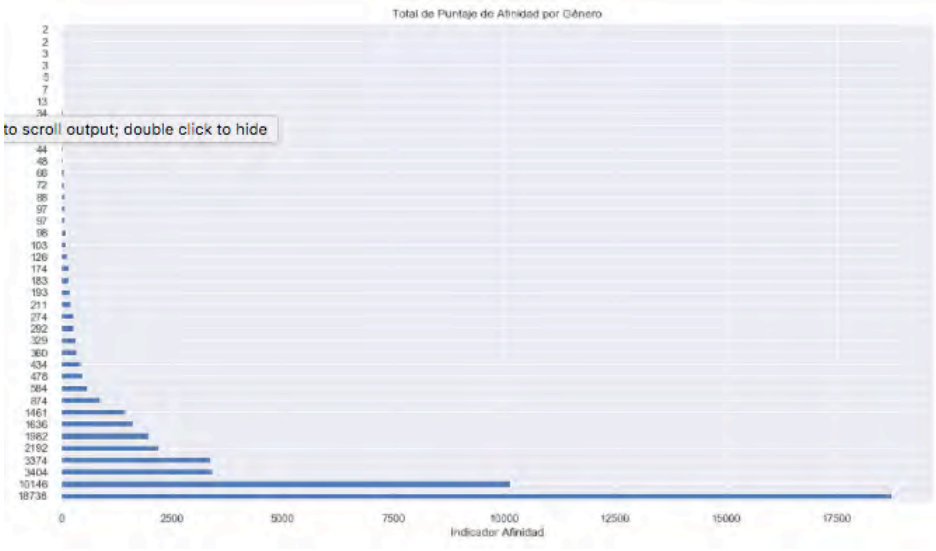
Se obtiene un conjunto de 775.258 registros de interacciones de 48.303 usuarios y 7.492 títulos de anime.

2.3.-Exploración de los datos:

Con los datos ya segmentados, se decide no continuar con la segmentación de data sets pensada originalmente ya que no contribuían a la reducción de dimensionalidad de los atributos a utilizar para el modelamiento (que corresponden a los 43 géneros de anime).

Estos atributos no presentan tampoco una correlación entre ellos que permita descartar inicialmente alguna de las variables.

Al observar las matrices de afinidad construidas de usuarios versus géneros, destacan que 2 géneros presentan una alta afinidad: Comedy (18.738 puntos) y Action (10.146), destacando por sobre el resto.



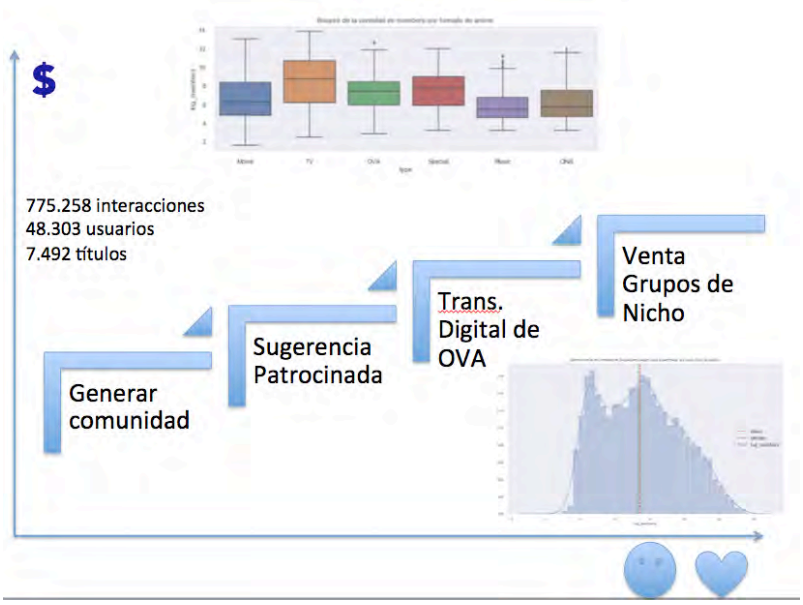
3.-Modelo de Negocio

Nuestra plataforma, es un puente de comunicación entre los usuarios consumidores de anime y las compañías productoras y distribuidoras de anime.

Bajo este paradigma establecimos una segmentación de producto desde una oferta de enganche para fortalecimiento de la comunidad, hasta la

comercialización de servicios Premium orientados a marketing de nicho para el anime.

Nuestros clientes, y quienes esperamos que paguen por nuestros servicios son las empresas distribuidoras de anime, a ellos les ofreceremos una escalera de valor:



La cual hemos organizado en 4 segmentos de negocio:

- **Generación de comunidad:** la base de nuestro servicio, es la gestión de la comunidad de usuarios, acá se encuentran los servicios de menor precio y menor compromiso por parte de los clientes, como la realización de concursos a un público objetivo o campañas de mailing.
- **Sugerencia Patrocinada:** es un servicio orientado a que los clientes posicionen un determinado título entre los demás usuarios, pudiéndose ser relanzamientos de producciones o la introducción de nuevos títulos.
- **Transformación Digital para OVAs:** los Original Version Anime (OVA), son producciones más antiguas, que no cuentan con una cobertura en la web completa, como repositorio de información o imágenes. Con el objeto de que las distribuidoras puedan volver a monetizar contenido antiguo, les ofrecemos el servicio de cerrar la brecha digital de las antiguas posiciones e introducirlas entre los nuevos usuarios.
- **Venta de Producciones a Clientes de Nicho:** Existe un alto número de producciones que no están dirigidas a un público masivo, por lo que dado el conocimiento que poseemos de los usuarios, ofrecemos el servicio de venta Premium, orientada a que los usuarios que frecuentan géneros específicos de anime reciban una oferta dirigida de parte de las productoras, de esta forma estas mejoran la cantidad de ventas de estos títulos. Al corresponder a una venta Premium la experiencia de venta va

más allá de un contenido audiovisual, sino que se acompaña de merchandising de lujo dirigido a público objetivo.

Como el concepto de la escalera de valor indica, los 2 primeros segmentos son productos de precios más competitivos ya que existen variados actores en el mercado que ofrecen servicios similares. Los 2 segmentos superiores corresponden a productos especializados, con baja competencia que presentan mejores oportunidades de obtener mayores márgenes de rentabilidad.

Para lograr comercializar los segmentos superiores se requiere necesariamente de construir las confianzas necesarios entre los usuarios y clientes, con los segmentos inferiores para alimentar el embudo de ventas, y lograr obtener las ventas de los servicios en donde hemos centrado nuestro foco.

Sistema de Recomendación

1.- Modelamiento.

Si nos enfocamos en nuestro objetivo general, éste es crear un modelo que prediga la mejor recomendación de película y/o serie de animé para un usuario, sea éste nuevo o existente en la web. Por esta razón hemos determinado que la forma de abordar esta tarea se divide en dos partes:

- Determinar una recomendación para usuarios nuevos que ingresan por primera vez al sitio web.
- Determinar una recomendación para usuarios existentes que ya cuentan con un registro de vistas y/o votaciones de animé.

En ambos casos nos encontramos frente a un problema de Clasificación y se utilizarán Algoritmos No Supervisados. Para los usuarios nuevos, éstos deberán seleccionar un listado de películas que son de su interés, generando un registro de sus preferencias. Para los usuarios antiguos ya existe un registro de vistas y valoraciones.

Según nuestra hipótesis, es posible agrupar los usuarios que han visto o evaluado positivamente algún título de anime con otros similares para entregar una recomendación y mejorar su experiencia, por lo que específicamente, usamos algoritmos de Búsqueda de patrones y Clustering para identificar tendencias entre ellos.

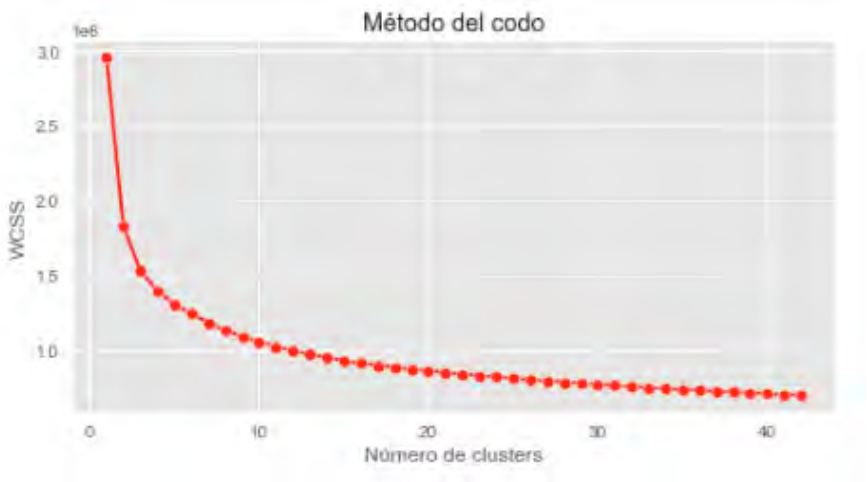
No se normalizaron los datos, porque las 43 variables se encontraban dentro del mismo rango de valores, ya que para K-Means creamos un puntaje de afinidad para cada usuario con cada género. Para el algoritmo Apriori, utilizaremos un set de datos con las preferencias de género de cada usuario.

Se utilizará una estrategia de dividir el set de datos en 70% entrenamiento y un 30% de validación, solo para el algoritmo de Clustering, K-Means. Para el caso del algoritmo de Búsqueda de patrones, Apriori, se utilizará el set de datos completo para permitir una mayor riqueza de datos para las asociaciones.

2.- K-Means:

Este algoritmo agrupa puntos de datos similares y descubre patrones subyacentes. Para esto, K-Means busca un número fijo (k) de clústeres en un conjunto de datos. Una de las dificultades de este algoritmo es la elección del número de clústeres, ya que podríamos obtener grupos muy heterogéneos o datos, que siendo muy similares unos a otros los agrupemos en clústeres diferentes. Escogimos el método del codo para esta tarea. Este utiliza los valores de la inercia obtenidos tras aplicar el K-Means a diferente número de clústeres, siendo la inercia la suma de las distancias al cuadrado de cada objeto de los clústeres a su centroide. Se representa una gráfica lineal de la inercia respecto del número de clústeres. En esta gráfica se debería apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia nos dirá el número óptimo de clústeres a seleccionar.

El set de entrenamiento para K-Means resultó con 33.812 filas y el set de validación con 14.491. Para definir la cantidad de clústeres utilizamos el método del codo con 43 iteraciones, que representan los géneros.



Se observa en la Figura que los valores de inercia dejan de descender más bruscamente cercano a los 20 clusters, por lo que escogimos este número para nuestro modelo.

Para medir la eficiencia del modelo utilizamos el método silhouette. El valor de la silueta mide cuán similar es un punto a su propio grupo (cohesión)

en comparación con otros grupos (separación). Su valor varía entre -1 y 1, siendo valores cercanos a 1 los más óptimos. Si los valores son más cercanos a -1 indica un mal agrupamiento. En la tabla número 1 se observa el promedio del puntaje silhouete para diferentes valores de clústeres.

Número de clústeres	Promedio **silhouette_score**
2	0.637635264525145
3	0.521596064890545
4	0.451621069213378
5	0.435515626447497
6	0.421435827832538
7	0.420861131647063
8	0.355439440725793
9	0.321661170584368
10	0.315565738806485
11	0.322477930779661
12	0.287496038653802
13	0.271217449030937
14	0.275853078195342
15	0.240966712962497
16	0.241026636624395
17	0.238162373266839
18	0.210115026720910
19	0.218546336413907
20	0.225070709262496

Observamos que el mejor valor de silouette lo tenemos para 2 clusters, pudiendo indicar lo observado en el análisis exploratorio, la división de títulos entre *Populares* y de *Nicho*. Pensamos que un valor significativo de clusters serían 7, ya que un valor más cercano a 0 indica que no tiene un gran nivel de significancia.

3.- Apriori:

Este algoritmo utiliza el concepto básico en estadística bayesiana, la probabilidad condicional: Para dos sucesos A y B,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \propto P(A \cap B)$$

Se puede aplicar esta definición también a variables discretas o continuas. Desde el punto de vista bayesiano, en la práctica, todas las probabilidades son condicionales porque casi siempre existe algún conocimiento previo o cierta experiencia acerca de los sucesos.

Los tres componentes principales que utiliza apriori son Soporte, Confianza y Lift. El valor de **soporte** nos indica la popularidad o la frecuencia de un ítem A dentro de un total de transacciones. La **confianza** se define como la probabilidad que un ítem A se encuentre cercano a un ítem B. **Lift** se refiere al aumento del ítem B en presencia del ítem A.

Para Apriori utilizamos el set completo en la búsqueda de patrones. Probamos con diferentes valores para soporte, confianza y lift para obtener un mayor número de asociaciones.

La variación más significativa que observamos respecto a los parámetros es en Soporte. Si disminuimos su valor al ejecutar Apriori, aumenta considerablemente la cantidad de asociaciones.

Min largo	Soporte	Confianza	Lift	Asociaciones
3	0,001	0,01	2	1106
3	0,0005	0,01	2	2444

A priori nos muestra explícitamente qué géneros están mayormente relacionados, con sus respectivas métricas de soporte y confianza. Nuestra recomendación se basa en altos valores de confianza, seguidos de soporte.

	ID1	ID2	support	confidence	lift
0	(Action,)	(Adventure,)	0.030660	0.172604	2.019414
1	(Action,)	(AdventureComedy,)	0.023173	0.130457	2.143079
2	(AdventureComedyEcchi,)	(Action,)	0.001159	0.684211	3.851903
3	(AdventureComedyGame,)	(Action,)	0.000802	1.000000	5.629704
4	(AdventureComedyMystery,)	(Action,)	0.000535	0.857143	4.825461
...	...	...	...	...	...
2439	(Military,)	(Romance, Sci-Fi, Music, Space, nan)	0.000624	0.025271	40.505415
2440	(AdventureComedy,)	(Ecchi, RomanceSchool, nan, Drama, FantasyHare...	0.000891	0.014641	16.427526
2441	(Action,)	(AdventureComedy, nan, MartialArts, Shounen, S...	0.002406	0.013547	5.629704
2442	(HaremMecha,)	(Comedy, Sci-FiShounen, Space, nan, Police, Ac...	0.000624	1.000000	1602.857143
2443	(ActionMecha,)	(Romance, Sci-Fi, Military, Music, Space, nan)	0.000624	0.100000	160.285714

Consideramos que Apriori nos entrega datos más claros para realizar una recomendación personalizada. En base a estas asociaciones, le recomendamos al usuario 10 películas que estén incluidas dentro de un género que tenga una alta probabilidad de asociación con el género preferido del usuario (alto valor de confianza), excluyendo las películas que el usuario ya vió y/o calificó.

Creemos que estos resultados pueden proporcionar una buena experiencia de usuario, ya que no es necesario un registro para optar a las

recomendaciones, simplemente basta con seleccionar títulos de interés y tendrá recomendaciones de acuerdo a sus gustos.

4.- Mejoras:

A nivel de modelamiento se podrían incluir otros algoritmos, para enriquecer la variedad de predicciones, ya sean Supervisados o No Supervisados.

El algoritmo Apriori nos permitió asociar diferentes géneros para recomendar títulos a los usuarios. Hemos considerado realizar predicciones con combinaciones de géneros, ya que un título puede tener más de un género asociado.

De acuerdo al análisis exploratorio pudimos observar que habían tendencias en cuanto a las visualizaciones y calificaciones. Según nuestro criterio podríamos dividir los títulos en populares y de nicho.

IV.- Producción

Este sistema de recomendación de cómo funciona el algoritmo está implementado en un jupyter donde pudimos comprobar su aplicación al colocar los usuarios nuevos (parámetro 'N') y existentes ('parámetro 'E') y que queda en anexos (carpeta recomendación) en la presentación de este proyecto.

El aplicativo de la web está en proceso, logramos hacer un front de la web y hacer una apirest en django integrada con base de datos postgresql. Esto aún está en desarrollo e integración para posteriores estudios UX/UI con los usuarios y se tenga una mejor interacción en la mejora de la plataforma.

V.- Conclusiones

El sistema de recomendación es uno de los algoritmos más usados y somos conscientes del beneficio que traería en la proyección que tienen en su negocio con el lanzamiento que piensan hacer y la integración de nuevos usuarios.

Los usuarios importan y tener un nuevo sistema de recomendación permitirá conocer más el gusto de los usuarios y asociarlos por géneros para conocer su preferencia y tener una visión de en sus gustos en la implementación de catálogos de películas e incluir algunas estrategias de marketing en el negoc