Proyecto Final de Inteligencia Artificial SISTEMA DE RECOMENDACIÓN DE ANIME

Alexander Sánchez, *Estudiante, Icesi,* Juan David Cruz, *Estudiante, Icesi,* Juan Sebastian Perez, *Estudiante, Icesi,* and Kennet Sánchez, *Estudiante, Icesi*

Abstract—Through the application of artificial intelligence models and strategies, this project seeks to create a tool that allows to make anime recommendations, taking the analysis of the ratings of a user.

Index Terms—Clustering, KNN, Recomendacion, Preferencias, Inteligencia artificial, Modelos, Parametros, Clasificacion, KMeans, Dataframe, Analisis, Depuracion de datos, Busqueda de K

1 Introduccion

La industria del entretenimiento ha cambiado mucho con respecto a sus inicios puesto que la inevitable globalización ha generado una mucho más diversa visión del mundo en relación a cómo consumimos contenido proveniente de diferentes lugares del mundo. Dentro de los contenidos extranjeros más relevantes está la industria del anime, la cual es una de las líderes del mercado provocando la creación de plataformas de streaming. Plataformas como Crunchyroll cuentan con más de 20 millones de usuarios al año y siguen un sistema de monetización basada en usuarios suscritos donde el mayor indicador de éxito es la permanencia de estos.

Los encargados de este proyecto como frecuentes consumidores de este tipo de contenido, conocen la importancia que tiene la personalización y la aplicación al gusto del usuario, por ello, surge el interés en la creación de un sistema de recomendación de anime. No solo porque como estudiantes novatos de inteligencia artificial es un tema que genera interés y permite aplicar los conceptos aprendidos en clase sino también porque se identifica una oportunidad de negocio que se base principalmente en que las plataformas de anime estan en este momento lejos de tener sistemas de recomendación tan acertados como lo tienen plataformas occidentales.

2 Marco Teorico

Para el apropiado entendimiento de este reporte, el lector debe estar familiarizado en primer lugar con todo el contexto de la inteligencia artificial. A continuacion se realiza una definicion especifica de los conceptos clave:

- **Anime:** Es la animación dibujada a mano y generada por ordenador originaria de Japón. Fuera de Japón y en inglés, el anime se refiere específicamente a la animación producida en Japón
- **Clustering:** Es la tarea de agrupar un conjunto de objetos de manera que los objetos del mismo grupo (llamado clúster) sean más similares (en algún sentido) entre sí que los de otros grupos (clústeres). Es una de las principales tareas del análisis exploratorio de datos y una técnica habitual del análisis estadístico de datos.
- **Clasificacion:** Accion de usar algoritmos especializados para combinar atributos similares de un conjunto de datos y generar agrupaciones adecuadas.
- **Outlier:** Un outlier o valor atípico es un punto de datos que difiere significativamente de otras observaciones.

3 Antecedentes

Nombre	Objetivo	Metodo	Resultados
Anibrain (2022)	Ayudar a la gente a encontrar nuevos anime y manga. Para que sin importar si una persona tiene o no experiencia previa con el anime, pueda encontrar algo nuevo y emocionante para leer o ver.	A diferencia de nuestro modelo, no se tiene cuenta la popularidad de un anime, sino las características innatas para determinar la similitud y hacer recomendaciones.	Anibrain cuenta con una pagina web en donde al buscar un anime, se devuelve como resultados los animes mas similares a este
My Anime List (2004)	Ofrecer a sus usuarios un sistema de listas para organizar y puntuar anime y manga.	No se usa inteligencia artificial, las recomendaciones son hechas por publicaciones de usuario.	Cuenta con una pagina web activa desde hace ya 18 años, pese a su anticuado metodo, sigue siendo una buena herramienta

"""

4 METODOLOGIA

4.1 Data understanding

Los datos usados provienen de la plataforma kaggle y contiene la información sobre los datos de las preferencias de 73.516 usuarios para 12.294 animes. Estos datos provienen de la pagina MyAnimeList.net.

Las columnas mas interesantes son la columna de genero y la columna de calificacion del anime. Las variables suelen ser de dos tipos, numericos (float y entero) y "categoria".

Se usan herramientas como diagramas de cajas para encontrar outliers e histogramas para encontrar que datos son los mas recurrentes en los dataframes. En esta parte se evidencia que la mayoria de animes se particionan en series de TV u OVAS/Peliculas, esto influye en que la cantidad de episodios y miembros tenga un comportamiento poco descriptivo pues todas las peliculas van a tener solo un episodio. Tambien se descubre que la mayoria de animes tiene una calificacion entre 6 y 7 aun cuando los usuarios del dataframe suelen calificar con una nota de 8.

4.2 Data preparation

Se realiza una limpieza de datos en donde se eliminan los datos nulos, repetidos e irrelevantes de los dataframes. Un ejemplo de dato inutil son los '-1' que aparecian en el dataframe de calificacion de los usuarios puesto que simbolizaba que el usuario no habia calificado un anime que habia visto, lo cual es irrelevante pues necesitamos conocer la opinion de los usuarios.

Despues de esto, se toma la decision de separar el dataframe de anime en 2 partes, uno para toda serie de TV y otro para las demas producciones. De aqui se encuentra que para una primera aproximacion, trabajar con 2 particionamientos es complejo, por ello se decide solo utilizar la particion que contiene las series de TV.

Tambien se observa que la columna de genero contiene a los generos como un texto separado por comas, lo que no permite identificarlos de una manera optima, por tanto, se realiza un procedimiento para generar nuevas columnas a partir de los 15 generos mas populares, en donde se tiene un 1 o un 0 dependiendo de si el anime pertenece o no al genero respectivamente.

4.3 Modeling

Para abordar este problema se tienen en cuenta dos fases importantes, una fase de clustering y una fase de clasificacion.

El objetivo principal de la fase de clustering es crear clusteres que permitan a los animes pertenecer a un grupo que los represente en la mayor medida de lo posible, para ello, deben seleccionarse del dataframe de series de tv, las variables mas adecuadas. Realizando el analisis notamos que las variables mas importantes son los generos, el numero de miembros y el rating de los animes, se descartan las variables inutiles como el nombre, el id y el numero de episodios. Lo siguiente es analizar las variables obtenidas y evaluar la correlacion de estas con otras para determinar cuales de ellas pueden ser descritas por otras, en este caso las correlaciones encontradas no fueron muy altas pero a partir de estas y de evaluar las caracteristicas de los generos, se eliminan 3 de estos: 'Shoujo', 'Mecha' y 'School'. Despues de esto se busca una K adecuada, es decir, que numero de grupos es el mas adecuado, para ello se tienen 3 metodos, el metodo del codo, el metodo de la silueta y el metodo de calinski-harabasz. Sin embargo, los metodos utilizados no llegan a un consenso sobre cual es el mejor valor de K, para el metodo del codo el mejor valor es 6, para el metodo de las siluetas entre mayor sea la k es mejor y para calinski-harabasz es totalmente lo contrario puesto que para este entre menos grupos mejor. Es por esto que se decide usar como valor de K el numero de generos que poseiamos, es decir 12. Finalmente se entrena el modelo y se aplica al dataframe creando la columna 'classification'.

Ahora se puede proceder a la fase de clasificacion, cuyo objetivo principal es que dado uno o varios animes se permita clasificarlos en uno de los grupos obtenidos en el clustering, esto con el fin de simular la forma en que un usuario puede interactuar en cualquier plataforma de streaming de anime. Para el procedimiento de esta fase, se utilizan tanto los atributos como el K que se obtuvieron en la fase de clustering puesto que es logico querer clasificar en los grupos que creamos previamente. Una vez que esto esta definido, se usa el metodo Train/Test/Validation para obtener las precisiones respectivas.

5 RESULTADOS

Una vez realizado el clustering se obtienen efectivamente 12 clasificaciones que corresponden a la mejor representacion de los generos del dataframe. En el caso de la clasificacion, se obtienen precisiones por encima del 95 por ciento en los conjuntos de entrenamiento, prueba y validacion.

6 ANALISIS DE RESULTADOS

Se dennota que las precisiones de la clasificacion poseen valores tan elevados puesto que tanto para el metodo de clasificacion como el de clustering se usan los mismos parametros. Esto genera que no se pueda realizar una comparacion adecuada con la informacion contenida y nos obliga a usar otro tipo de estrategias. Una estrategia cualitativa y que se acerca al despliegue es crear datos que unifiquen la informacion de distintos animes para que de esta manera se simule la manera en que un usuario interactua en las diversas plataformas de streaming que existen. De esta manera, se mostrarian los animes con mejor calificacion y mayor cantidad de miembros para que el usuario tenga una aproximacion al tipo de contenido que le gustaria ver. Sin embargo, es importante revisar este aspecto para encontrar una mejor manera de analizar la informacion de manera cuantitativa. Este aspecto sera consultado con el profesor y queda pendiente para una proxima entrega.