Exploración de Anime & Sugerencias de títulos

Sabrina Cabrera

Data Science

Comisión 42410 CoderHouse

Introducción

Este proyecto se centra en el tan interesante y amplio mundo del anime, un fenómeno cultural que ha capturado la imaginación de audiencias a nivel mundial.

El objetivo principal es desarrollar un sistema de recomendación que proporcione sugerencias personalizadas de títulos de anime a los usuarios. La necesidad de este sistema surge de la enorme y siempre en aumento biblioteca de animes disponibles, lo que puede ser abrumador para los espectadores al intentar seleccionar qué ver a continuación. Utilizando un conjunto de datos exhaustivo de MyAnimeList, uno de los recursos más completos sobre anime, este proyecto emplea técnicas de análisis de datos y aprendizaje automático para ofrecer recomendaciones precisas y significativas. A través de este sistema, buscamos mejorar la experiencia de los aficionados al anime, facilitando el descubrimiento de títulos que se alineen con sus gustos y preferencias individuales.

Metodología

La metodología adoptada para este proyecto abarca varias etapas clave, comenzando con la recopilación de datos. Se extrajo un conjunto de datos exhaustivo de MyAnimeList utilizando la API de Jikan, incluyendo información detallada sobre animes y preferencias de usuarios. Posteriormente, se realizó un proceso meticuloso de limpieza y preprocesamiento para asegurar la calidad y consistencia de los datos.

En la fase de análisis, se implementaron técnicas de análisis exploratorio para identificar patrones y tendencias clave. Este análisis fue fundamental para entender las variables críticas que influyen en nuestro modelo de recomendación.

Para el desarrollo del modelo, se adoptó un enfoque híbrido, combinando elementos de filtrado colaborativo y técnicas basadas en contenido. Esto permitió generar recomendaciones que no solo se basan en la popularidad general de los títulos, sino también en las preferencias específicas de los usuarios.

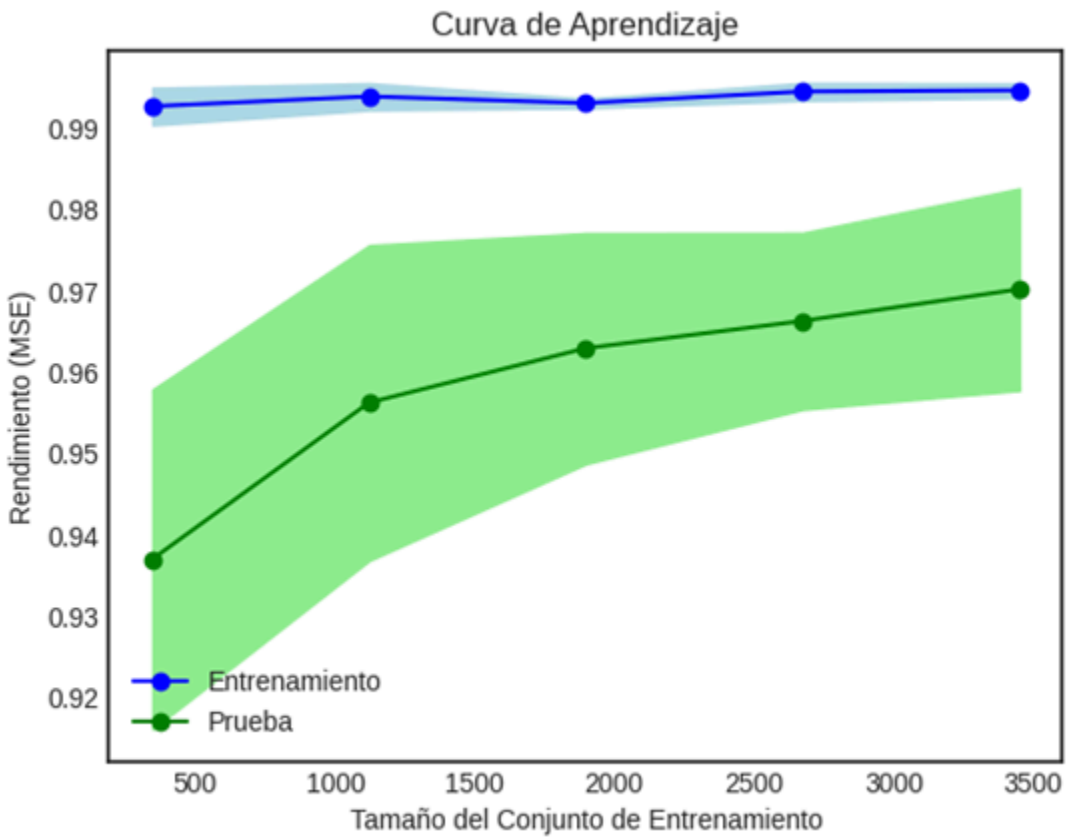
Finalmente, el modelo fue sometido a un riguroso proceso de validación y prueba, utilizando métricas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R2) para evaluar su rendimiento y precisión.

Desarrollo del Modelo de Recomendación

El núcleo de este proyecto es el desarrollo de un modelo de recomendación de anime preciso. Para lograr esto, se implementó un enfoque híbrido con combinación de técnicas de filtrado colaborativo y basado en contenido. Esta metodología permite que el sistema no solo considere las tendencias generales de popularidad y clasificación de los animes, sino también las preferencias y comportamientos individuales de los usuarios.

Se emplearon algoritmos avanzados de aprendizaje automático para procesar y analizar el conjunto de datos. Estos algoritmos fueron cuidadosamente seleccionados y ajustados para optimizar la capacidad del modelo para generar recomendaciones relevantes y personalizadas. La integración de múltiples factores, como la popularidad del anime, las clasificaciones, y las preferencias del usuario, fue clave para construir un sistema robusto y eficaz.

El modelo fue entrenado y validado utilizando un conjunto de datos significativo, asegurando su fiabilidad y precisión. Se utilizaron métricas de evaluación como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R2) para medir su rendimiento y ajustar los parámetros según fuera necesario.



Sobre Ajuste

El modelo tiene un rendimiento notablemente mejor en el conjunto de entrenamiento en comparación con el conjunto de prueba. El MSE para el entrenamiento es significativamente más bajo (0.0041) que para la prueba (0.0157), lo que indica que el modelo puede estar sobre ajustado a los datos de entrenamiento.

Curva de Aprendizaje

La línea azul (entrenamiento) se mantiene relativamente plana y alta, lo que indica un buen rendimiento a medida que se incrementa el tamaño del conjunto de entrenamiento.

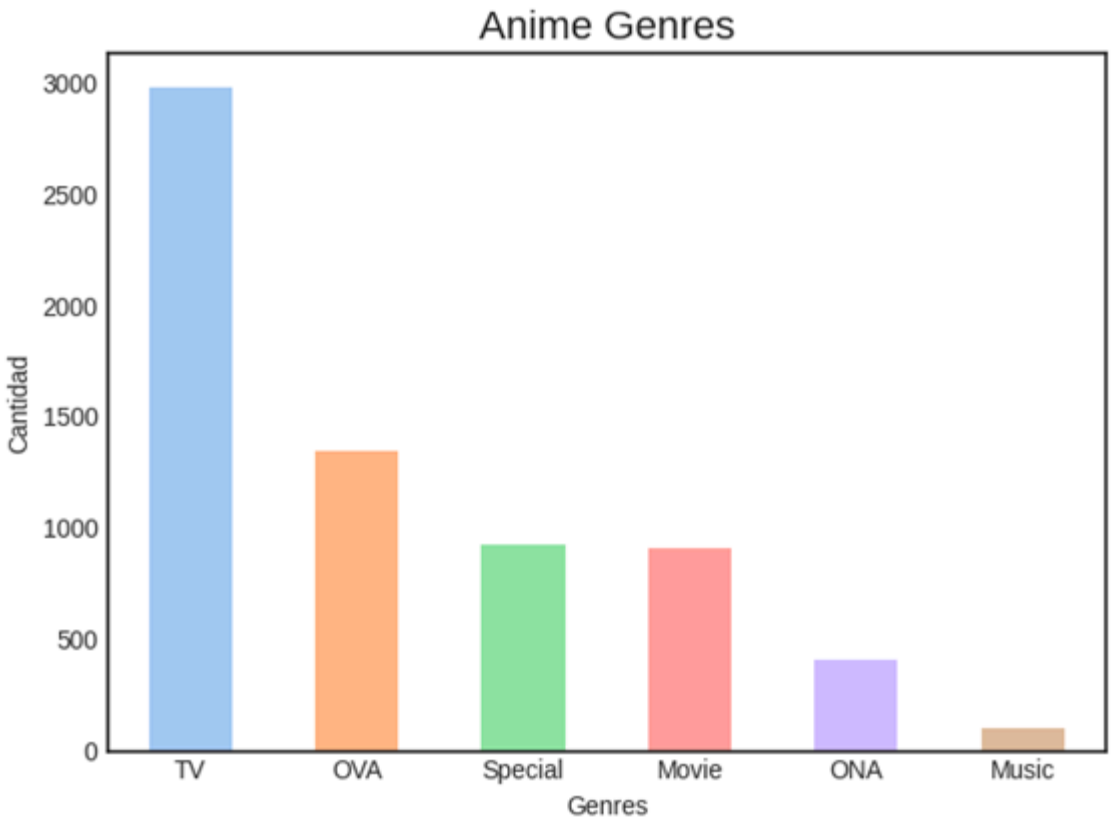
La línea verde muestra una mejora en el rendimiento a medida que más datos de entrenamiento están disponibles, lo que es una señal positiva.

Brecha entre las Curvas

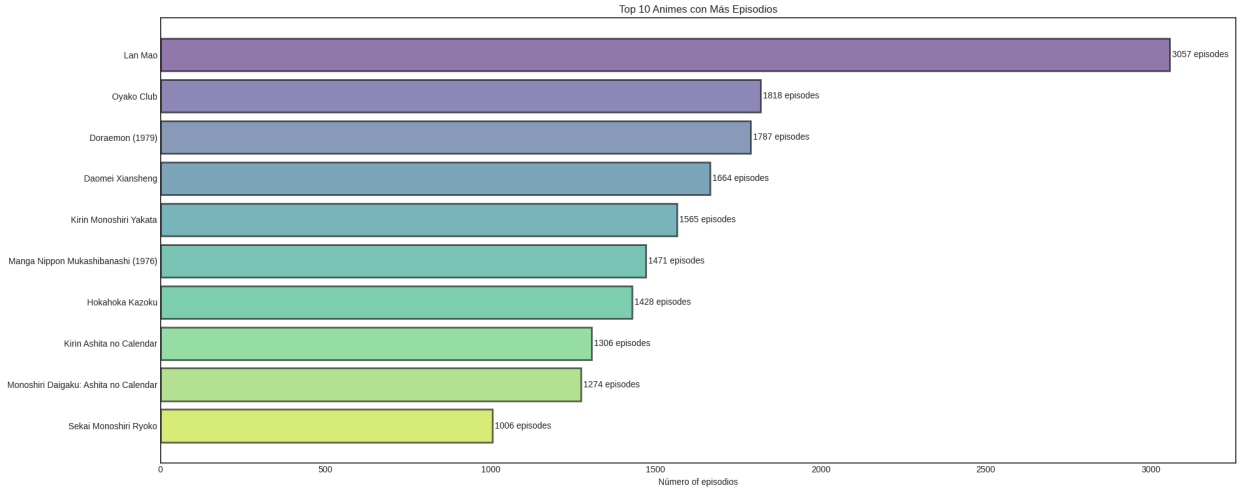
La brecha constante entre las líneas de entrenamiento y prueba sugiere que agregar más datos de entrenamiento no estaría reduciendo el sobreajuste. Entonces podríamos decir que se necesita de una regularización más fuerte o que el modelo es demasiado complejo para los patrones presentes en los datos.

Algunos insights obtenidos

La mayoría de los títulos son producciones para TV, el resto son OVA ('Original video animation', producciones generalmente especiales de pocos o un solo episodio), luego especiales y películas, casi la misma cantidad. ONA ('Original net animation', producciones directamente creadas para internet) y en mucho menor medida animes musicales.

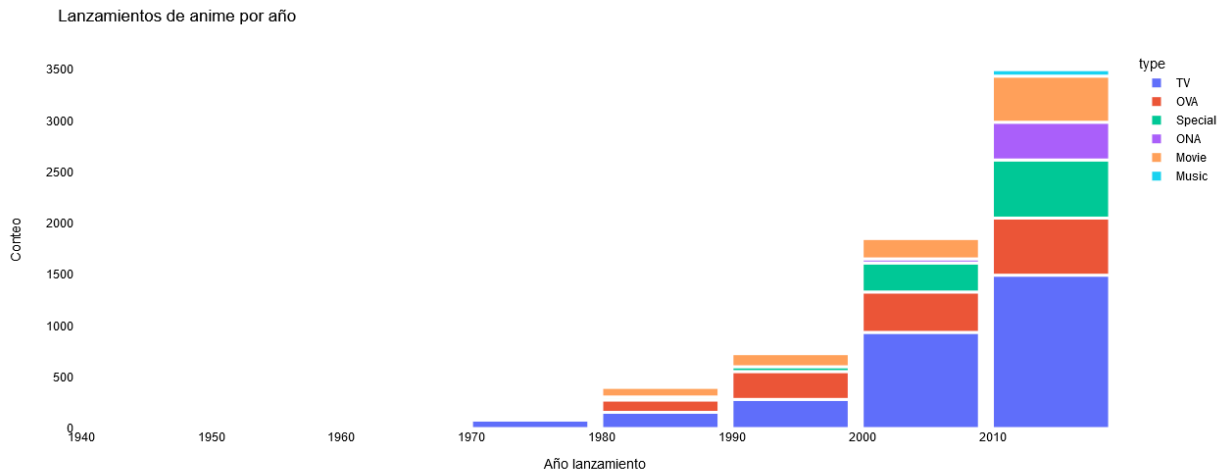


Veamos la popularidad de los top 10 animes de todos los tiempos, con mayor cantidad de episodios:



No hay una relación aparente entre la cantidad de episodios, al menos positiva, y los favoritos (popularidad) de los animes. El único a destacar es Doraemon, un anime muy popular, orientado a niños.

¿Cómo ha sido la evolución de lanzamientos de títulos a través de las décadas?



El crecimiento de títulos y variedad de géneros es notable, y probablemente influye, retroactivamente, en la popularidad y creación de nuevos títulos.

Análisis de Popularidad y Tendencias

¿Cuáles son los géneros de anime más populares?

Y una de las preguntas más debatidas, ¿cuáles son los MEJORES animes de todos los tiempos?

Fullmetal Alchemist: Brotherhood 9.10

Steins;Gate 9.07

Gintama° 9.06

Sousou no Frieren 9.05

Gintama: The Final 9.04

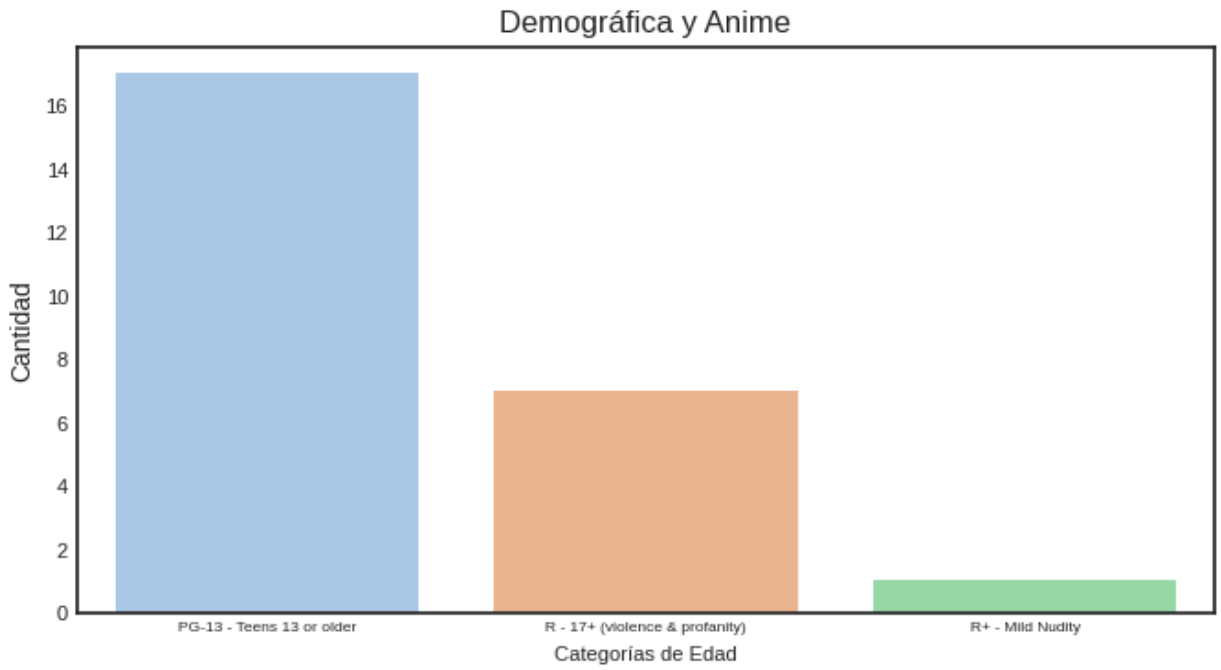
Bleach: Sennen Kessen-hen 9.04

Hunter x Hunter (2011) 9.04

Gintama' 9.03

Gintama': Enchousen 9.03

Análisis Demográfico

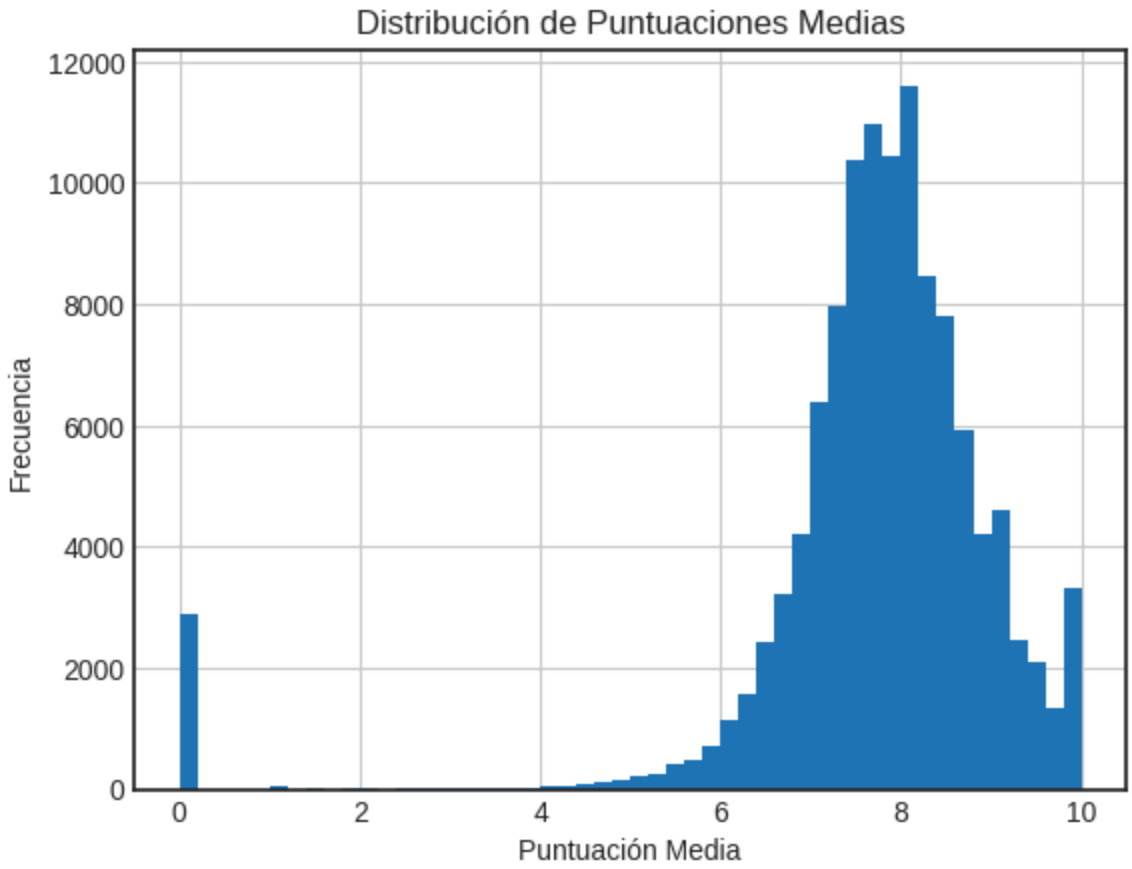


Mayoritariamente para un público adolescente, con un considerable contenido Restringido +17 y luego un menor, pero igualmente significamente, R + por mínima desnudez.

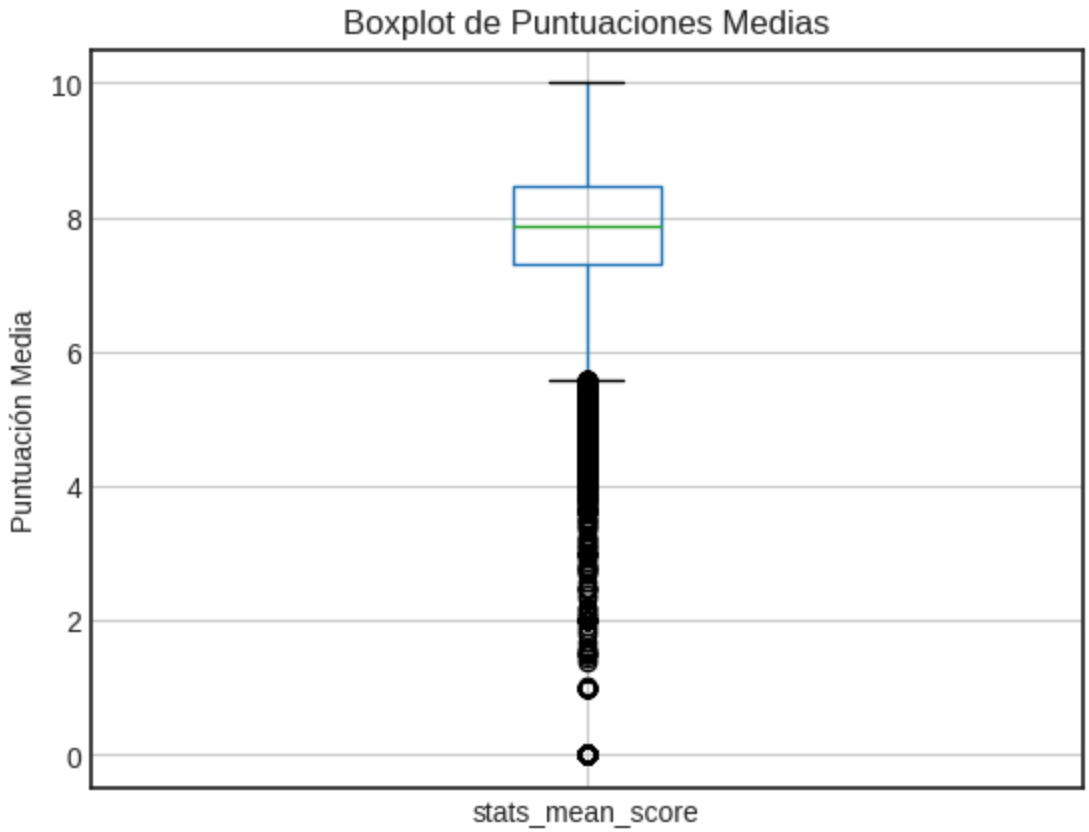
Tengamos en cuenta que no se está contando acá el contenido que es explícito para adultos.

Vamos a observar ahora el df de usuarios de MAL

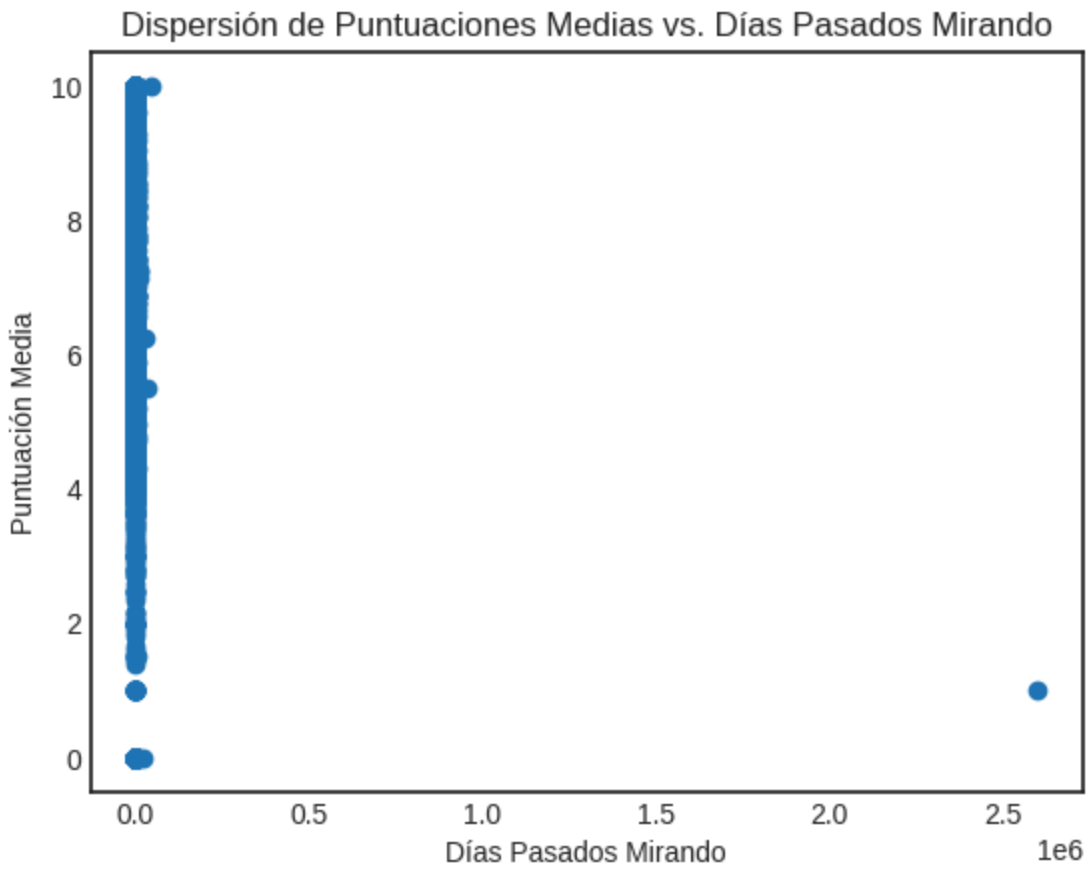
Démosle una mirada la distribución de las puntuaciones que los usuarios de MAL le dieron a los animes que vieron



Las puntuaciones se concentran significativamente entre el 7 a 8, lo que sugiere que la mayoría de los títulos han recibido puntuaciones dentro de este rango. Hay muy pocas puntuaciones en los extremos bajos (cerca de 0) y altos (cerca de 10), lo que nos dice que existen pocos títulos con puntuaciones muy bajas o perfectas. La forma de la distribución es ligeramente sesgada hacia la izquierda, lo que significa que hay una tendencia a que las puntuaciones sean más altas en lugar de más bajas.



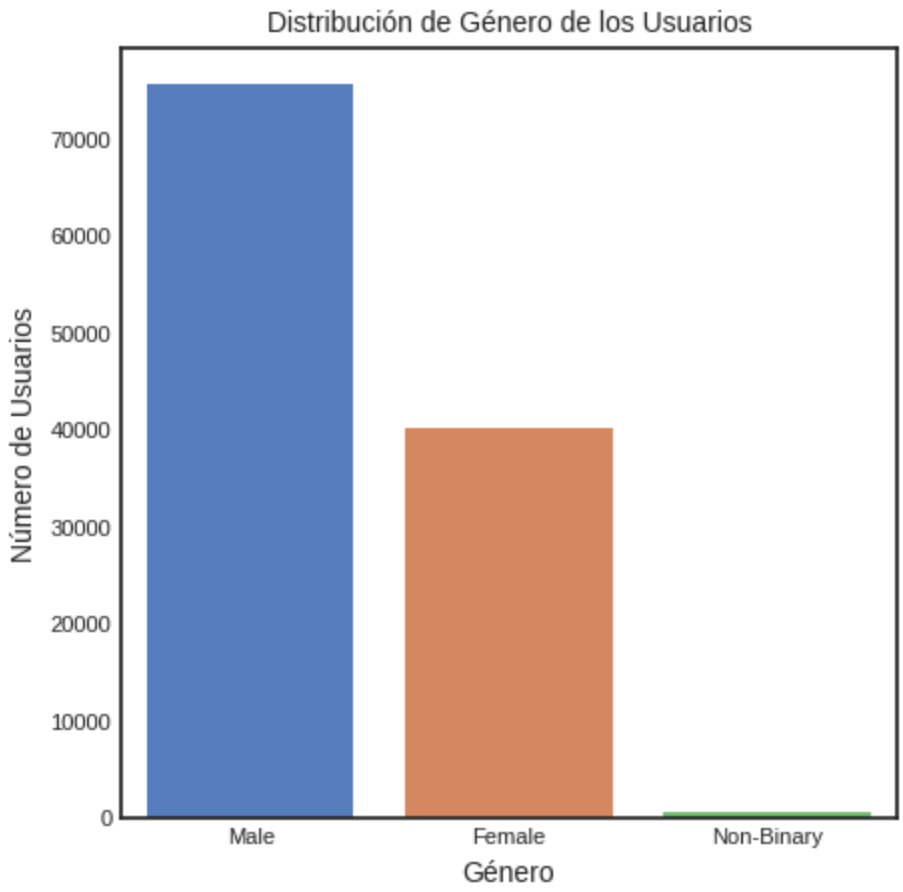
Este boxplot confirma lo que las estadísticas descriptivas sugerían: aunque la mayoría de las puntuaciones están agrupadas alrededor de 7 a 8.5, hay una dispersión considerable y una cantidad significativa de valores atípicos bajos. Estos valores atípicos podrían estar afectando la media y podrían merecer una investigación más detallada para entender por qué algunos usuarios tienen puntuaciones medias tan bajas (entiendo que, es común en la web de MAL, no tener puntuaciones, la web lo permite. Esto afecta las medidas). Además, el hecho de que la caja no esté centrada perfectamente alrededor de la línea de la mediana indica que hay una ligera asimetría en los datos, con una cola de distribución hacia las puntuaciones más bajas.



El scatter plot muestra que no hay una correlación evidente entre el tiempo dedicado a mirar y las puntuaciones medias otorgadas por los usuarios. Además, algunos valores atípicos extremos indican comportamientos o datos inusuales.

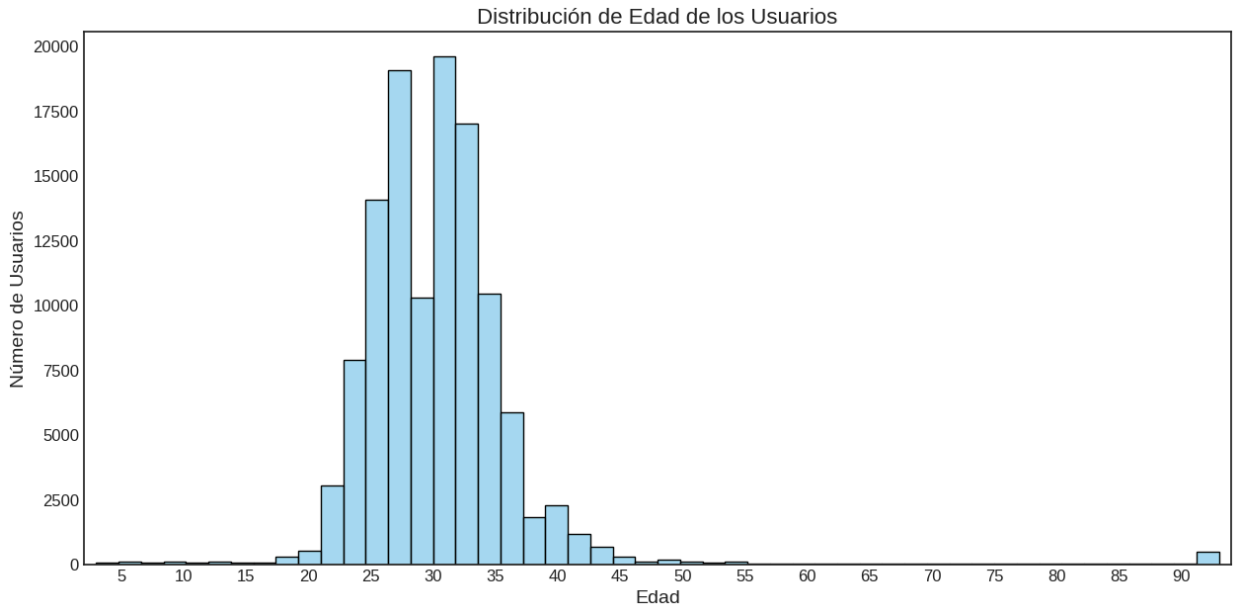
Al comprobar si hay datos faltantes o nulos que podrían afectar el análisis no encontramos.

Observemos la Distribución de Género de los Usuarios de MAL



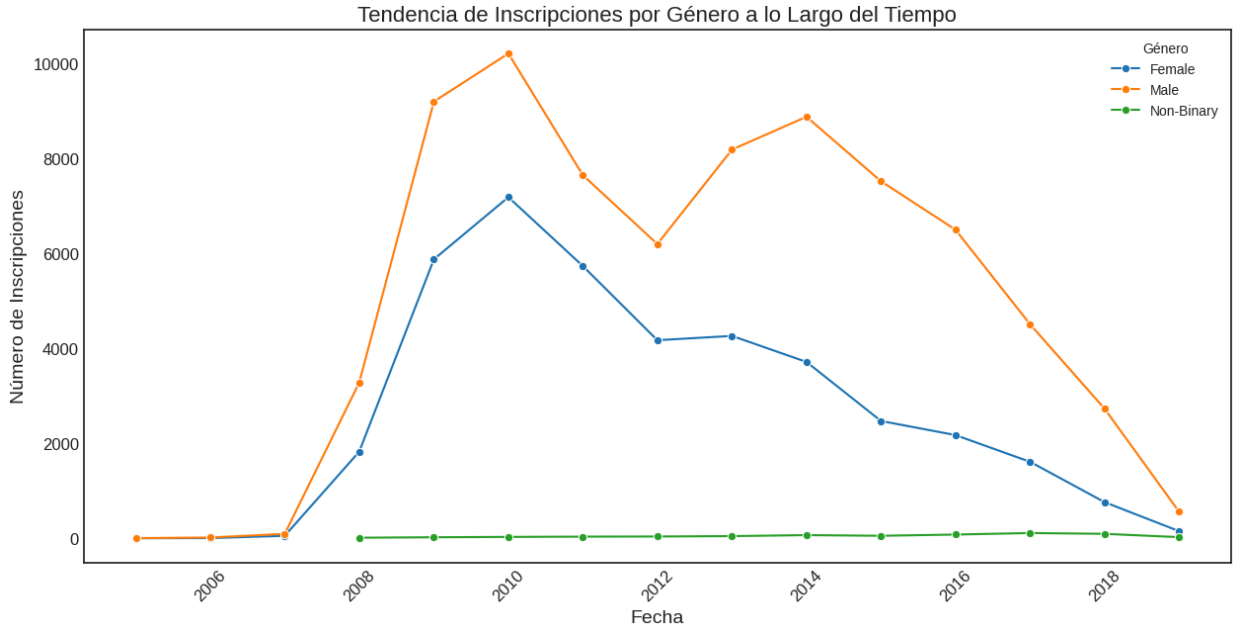
La mayoría de los usuarios son hombres y la mitad de esta cantidad, son mujeres. Con una muy breve cantidad de usuarios no binarios. Son datos esperados, pero que con el tiempo van cambiando y en tendencia de aumento, para los géneros femenino y otros.

Veamos entonces la Distribución de Edad de los usuarios



Esto es normal para lo que ya hemos observado antes, el estereotipo del anime para infantes no es más que eso. Y las edades se distribuyen en una gran mayoría en jóvenes adultos.

¿Cómo es la tendencia de inscripciones por género a lo largo del tiempo para MAL?



Cómo mencionaba anteriormente, la tendencia ha ido en aumento para los generos mujer y no binario, aunque mínimamente para este último. Luego hay una decaída en las inscripciones

hasta ahora, pero eso puede ser debido a una caída en la popularidad de la website o un cambio en la data que se registra, pudiendo ser un dato, género.

Al realizar una prueba t para comparar media entre grupos:

El resultado de la prueba t es bastante significativo. Aprox. -14 en Estadístico t y Valor p 3.4742875388245588e-47 sugiere que hay una marcada diferencia estadística entre las puntuaciones medias de los dos grupos que analicé: hombres y mujeres.

Dado que el valor p es extremadamente pequeño, mucho menor que el umbral típico de 0,05 o umbrales incluso más estrictos, se puede rechazar la hipótesis nula de que los dos grupos tengan la misma puntuación media con alta confianza. El signo negativo del estadístico indica que la puntuación media del primer grupo que pasó (usuarios hombres) es menor que la puntuación media del segundo grupo (mujeres).

Evaluación del Modelo

La efectividad del modelo de recomendación se evaluó utilizando varias métricas estándar en el campo del aprendizaje automático. El Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE) se calcularon para cuantificar el grado de error entre las predicciones del modelo y los valores reales. Estas métricas proporcionaron una valoración clara de la precisión del modelo en términos numéricos.

Además, se utilizó el coeficiente de determinación (R2) para medir qué tan bien las predicciones del modelo se corresponden con los datos observados. Un valor R2 cercano a 1 indica que el modelo puede explicar una gran proporción de la variabilidad en los datos.

El análisis de estas métricas reveló aspectos clave sobre la precisión y la eficiencia del modelo, así como áreas para mejoras futuras. Esta evaluación profunda aseguró que las recomendaciones generadas sean tanto relevantes como confiables para los usuarios.

Testing

Se realizaron pruebas para observar qué títulos le recomendaba nuestro modelo a un usuario de MAL.

Acá tenemos un caso de análisis:

Para el usuario **soccerscot15**, las recomendaciones fueron:

lo cual tiene sentido a simple vista, ya que si observamos algunos de los animes favoritos del usuario, estos son Rekka no Honoo (acción, aventura), Ouran Koukou Host Club (comedia, romance) y Karin (comedia, romance, sobrenatural).

Estos tres animes del usuario tienen características en común con nuestras recomendaciones:

- La mayoría (excepto 1, son animes previos al 2010)
- El estilo de animación es conservadoramente tradicional, incluso nuestra recomendación del remake de Sailor Moon es una versión que respeta lo tradicional de la publicación original.
- Los géneros son consistentes.

Conclusiones

Este proyecto demostró la viabilidad de aplicar técnicas avanzadas de análisis de datos y aprendizaje automático para crear un sistema efectivo de recomendación de anime. La implementación de un modelo híbrido que combina filtrado colaborativo y basado en contenido resultó ser una estrategia exitosa, equilibrando las preferencias individuales con las tendencias generales.

En resumen, el algoritmo tiene un comportamiento bueno e interesante de analizar, destacando como lo más importante que:

- El modelo muestra un nivel alto de precisión y una buena capacidad para predecir, dados el bajo MSE y el alto valor de R2.
- La importancia asignada a ciertas características, como el rank (rango), es crucial para las predicciones del modelo.
- Se observaron estas conclusiones en las pruebas, con una buena consistencia, incluso a “la vista” se pueden observar algunas características interesantes, como que el algoritmo elija un estilo de animación determinado, o que la temática de los títulos sean similares en algunos aspectos, al analizar la trama. Por esto puedo decir que puede que

incluso tengamos algunos puntos de exactitud más allá de los esperados, que nos dan unos resultados muy interesantes.

Este proyecto tiene una base interesante para futuras investigaciones y desarrollos en el campo de la recomendación personalizada.