Exploración de Anime & Sugerencias de títulos

Sabrina Cabrera
Data Science - Comisión 42410
CoderHouse

Agenda

- Introducción
- Metodología
- Resumen de metadata
- Preguntas/hipótesis
- Análisis Exploratorio
  - Top 10 Animes con Más Episodios
  - Top 10 Animes Mejor Puntuados de Todos los Tiempos
  - Top 10s - Cuestión de Género
  - Top 10s - Demográfica de Géneros y Edades
- Insights
- Modelo: Funcionamiento del Sistema de Recomendaciones
- Evaluación del Modelo
- Testing
- Conclusiones
- Agradecimiento

Links

Github
API Jikan
Entrega final - Google Colab

Introducción

Motivación y audiencia

El anime ha pasado de ser un nicho cultural en Japón a convertirse en un fenómeno global que influye en el entretenimiento, la moda y el arte de todo el mundo. Mediante el análisis de los datos recogidos en MyAnimeList, busco comprender las tendencias, preferencias y comportamientos de los usuarios del anime. Este análisis no sólo permitirá descubrir qué series son más populares entre los aficionados, sino también identificar patrones demográficos y temporales (entre otros) que podrían ayudar en recomendaciones para personas en búsqueda de nuevos títulos.

El **objetivo principal** es generar un algoritmo de recomendación de anime, en base a los datos obtenidos de MyAnimeList.

Otros **objetivos secundarios** son realizar un análisis acerca del fenómeno del anime, explorar datos a través de los años, los títulos más populares, demográfica, etc.

Audiencia

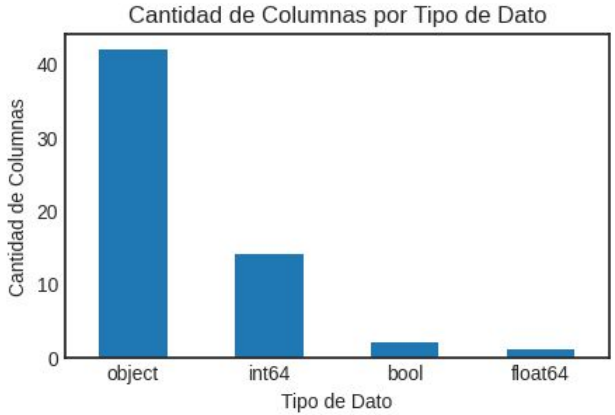
Los fans del anime, que podrán descubrir títulos que podrían desconocer y obtener una comprensión más profunda de cómo sus propias preferencias se alinean o difieren de las tendencias generales.

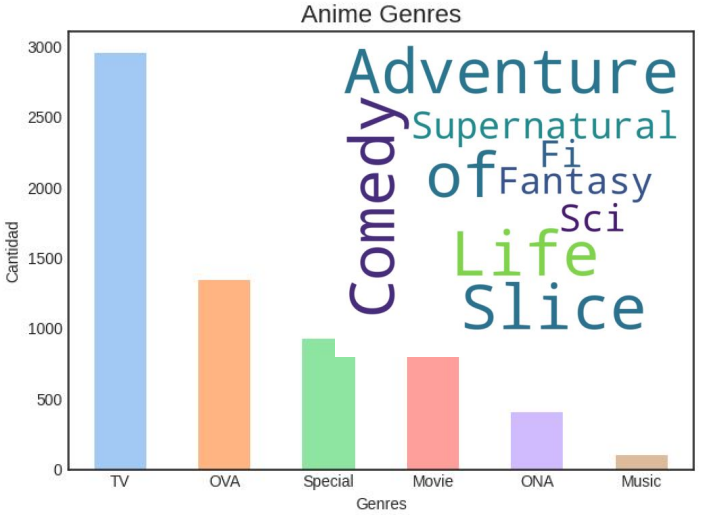
Metodología

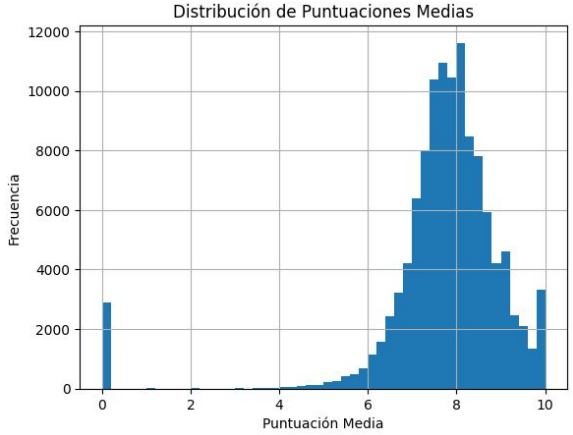
1. **Recopilación de Datos:** la fuente de los datos es MyAnimeList, extraídos mediante la API Jikan incluyendo información detallada sobre animes y usuarios.
2. **Limpieza y Preprocesamiento:** se realizó un proceso de limpieza de datos para manejar valores faltantes y corregir inconsistencias. Esto incluyó la conversión de fechas a un formato uniforme y el manejo adecuado de listas en ciertas columnas.
3. **Análisis Exploratorio:** se llevó a cabo un análisis exploratorio para comprender la naturaleza de los datos, verificando tipos de datos y explorando distribuciones y relaciones clave.
4. **Transformación de Datos:** se aplicaron técnicas específicas de transformación de datos, como la normalización y la creación de variables derivadas, para preparar los datos para análisis posteriores.
5. **Herramientas:** para este proceso, se empleó el lenguaje Python y las librerias pandas, NumPy, matplotlib, seaborn y sklearn.

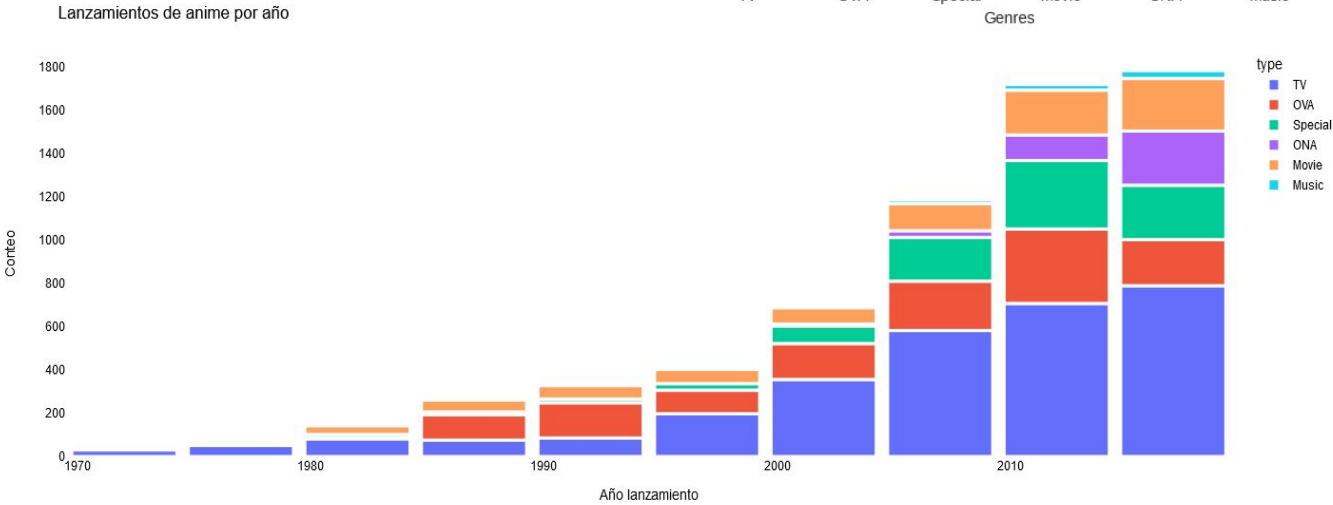
pandas NumPy seaborn scikit learn

Resumen de metadata









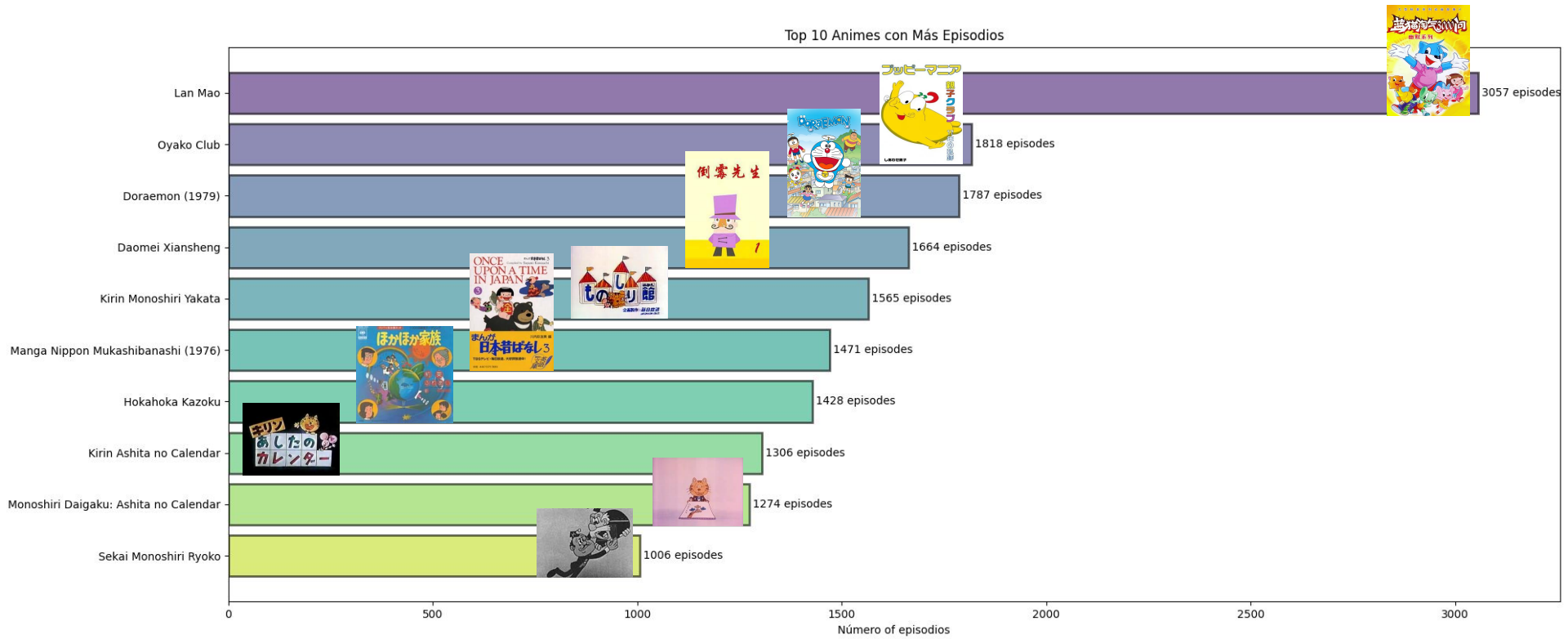
Preguntas/hipótesis

- ¿Incide la demográfica en la elección de los animes?
- ¿Suele un género realmente mirar el anime demográficamente denominado como tal?
- ¿Pasa similar a ésto con la demográfica de edades?

Análisis Exploratorio

Top 10s

MyAnimeList



Análisis Exploratorio
Top 10s
MyAnimeList

Top 10 animes mejor puntuados 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

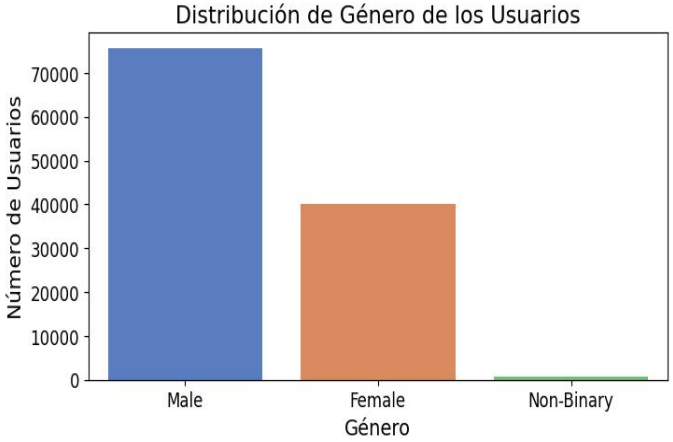
Bleach: Sennen Kessen-hen 9.04

Hunter x Hunter (2011) 9.04

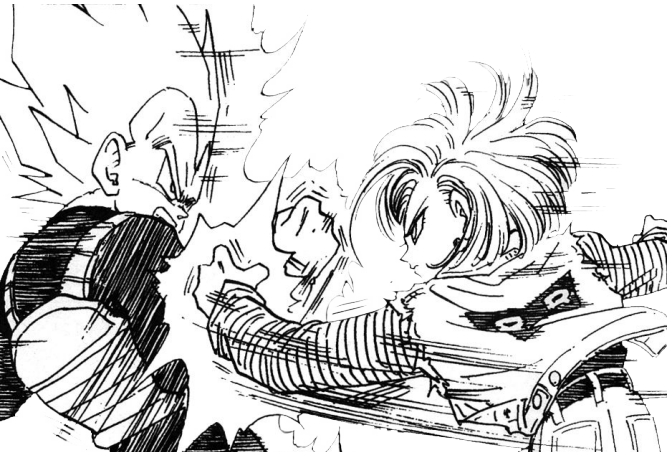
Gintama' 9.03

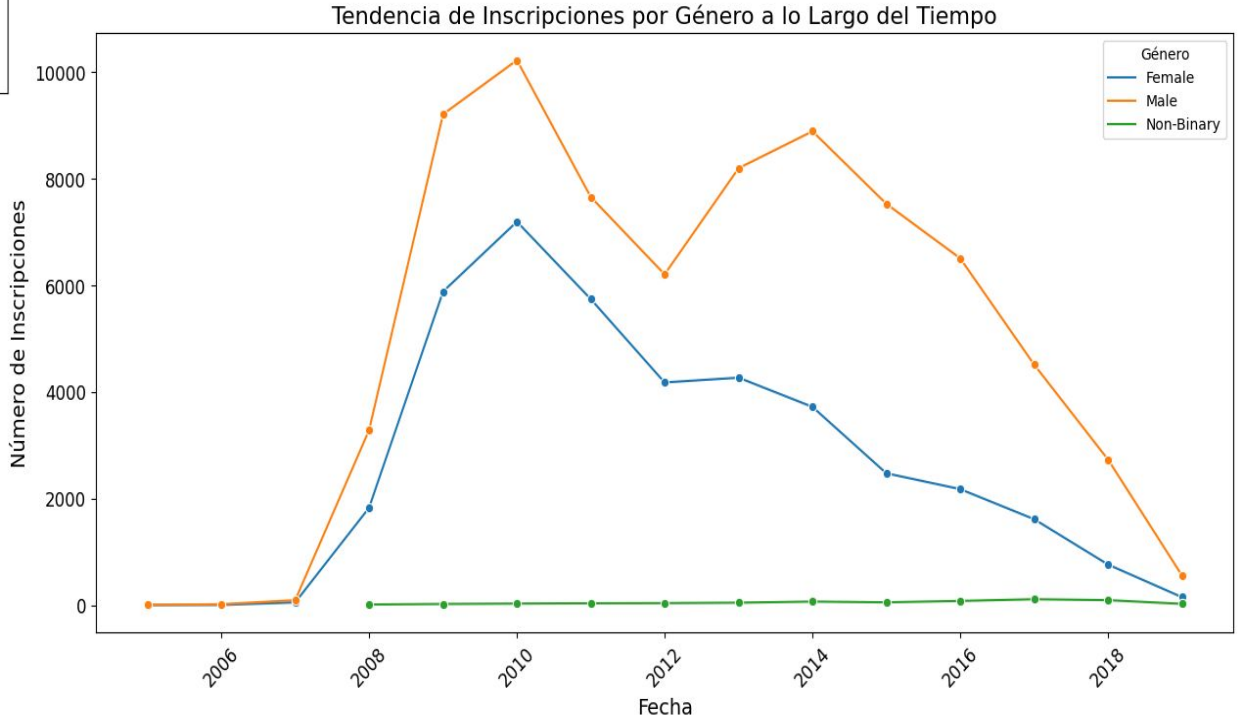
Gintama': Enchousen 9.03

Análisis Exploratorio
Top 10s
MyAnimeList



Cuestión de género

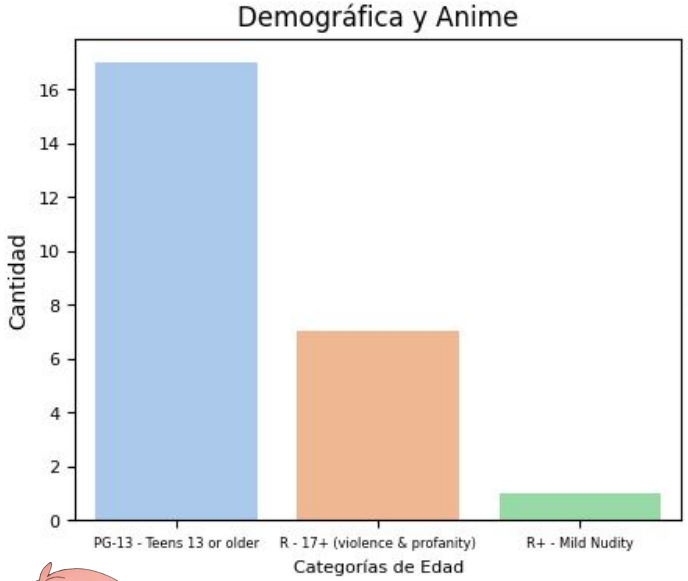


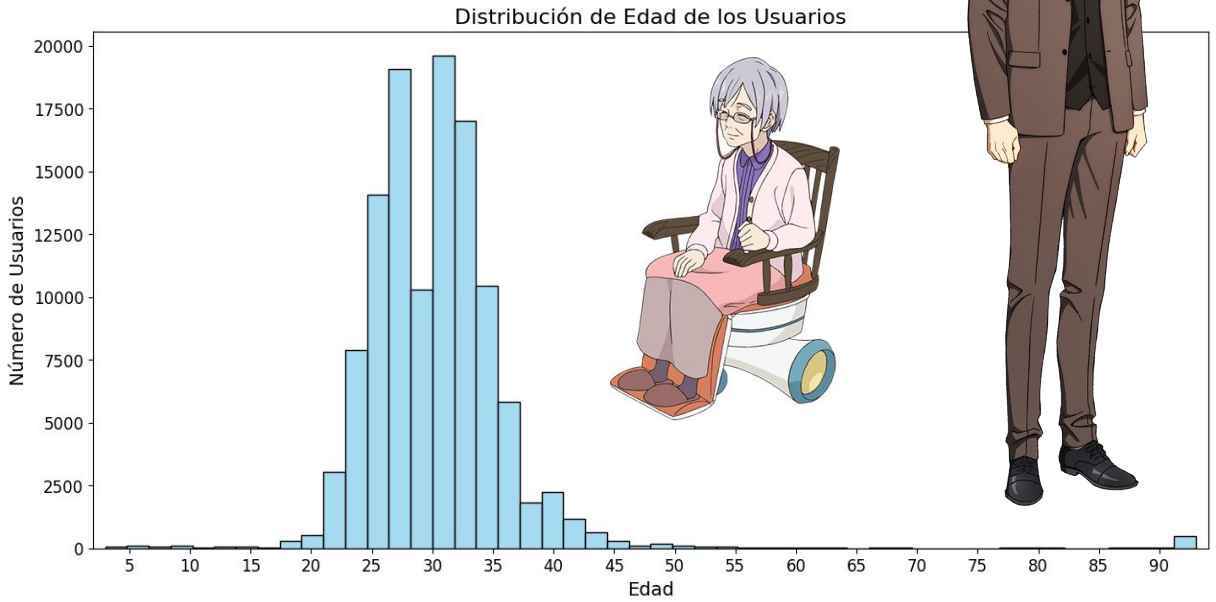


Análisis Exploratorio

Top 10s MyAnimeList

Demográfica de Géneros y Edades





Insights

Principales

- 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.

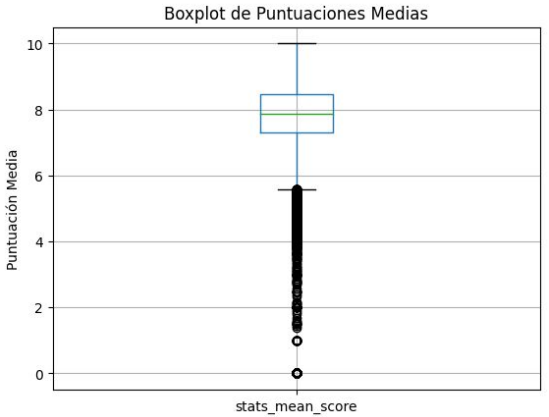
- La tendencia ha ido en aumento para los generos mujer y no binario, aunque mínimamente para este último. Luego hay una decaída en las inscripciones hasta ahora, pero eso puede ser debido a una caída en la popularidad de la website o un cambio en la data que se registra, pudiendo ser un dato, género.

- El estereotipo de que el anime es para infantes no es más que eso, un estereotipo. Y las edades se distribuyen en una gran mayoría en jóvenes adultos.

Secundarios

- No hay una correlación evidente entre el tiempo dedicado a mirar y las puntuaciones medias otorgadas por los usuarios.

- 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. (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.



Modelo: Funcionamiento del Sistema de Recomendaciones

Se llevó a cabo el desarrollo de un sistema de recomendación utilizando técnicas de análisis de datos y aprendizaje automático.

- **Datos Utilizados:** se utilizó un conjunto de datos grande y con mucha información; incluído valoraciones, preferencias de usuarios y características de los animes.

- **Metodología del Modelo:** se empleó un enfoque híbrido, combinando técnicas de filtrado colaborativo y basado en contenido, para generar recomendaciones personalizadas.

- **Proceso de Desarrollo:** se realizó un análisis exploratorio para identificar patrones y tendencias, seguido por la selección y ajuste de algoritmos apropiados.

- **Validación y Pruebas:** se llevaron a cabo pruebas para validar la efectividad del modelo, asegurando que las recomendaciones sean relevantes y precisas.

Evaluación del Modelo

- El modelo de Random Forest, después de un ajuste de hiper parámetros, muestra un buen rendimiento en el conjunto de entrenamiento pero tiene algunas señales de sobreajuste, dada la diferencia significativa entre los errores del conjunto de entrenamiento y de prueba.

- La correlación entre el 'rank' y el 'score' es muy alta, lo que sugiere que 'rank' es un predictor clave del 'score'.

- La reducción de dimensionalidad mediante PCA y la selección de características basada en modelos ayudaron a simplificar el modelo, aunque con un ligero decremento en el rendimiento comparado con el modelo más complejo.

Testing
Ejemplos de recomendaciones generadas

S

Conclusiones

- 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.

ありがとう

¡GRACIAS!