Resumen Ejecutivo

Introducción

La era digital ha revolucionado la forma en que consumimos contenido de entretenimiento, proporcionando acceso inmediato a una gran diversidad de series y películas de todo el mundo. En particular, el Animé, un género originario de Japón ha captado la atención de millones de personas y se ha consolidado como una fuerza significativa en la industria del entretenimiento global.

Somos una plataforma líder para la exploración y valoración de series y películas de Animé, facilitando a los usuarios la elección de contenidos que se adapten a sus gustos y preferencias, desempeñando un papel clave en la toma de decisiones de los fans de este género.

Sin embargo, hemos identificado un reto importante, un porcentaje considerable de nuestra muestra de contenido no tiene valoraciones. Esta falta de información puede afectar la percepción de nuestros usuarios sobre la calidad y el atractivo de nuestro contenido, y potencialmente, su decisión de invertir tiempo en ellas.

Pregunta de investigación

El presente documento, trata de responder la siguiente pregunta de investigación:

¿Se puede desarrollar un sistema predictivo para estimar el puntaje de series y películas de Animé que carecen de valoraciones utilizando técnicas de aprendizaje automático, con el objetivo de implementar un modelo de negocios para comercializar a otros usuarios interesados?

Planteamiento de la hipótesis de trabajo

La pregunta planteada anteriormente, la analizaremos y trabajaremos siguiendo la siguiente hipótesis:

"Si se utiliza un algoritmo de análisis de las características claves de series y películas de Animé, junto con técnicas de aprendizaje automático, es posible desarrollar un sistema de

estimación de puntajes preciso que pueda predecir de manera efectiva el puntaje que una serie/película debería tener, incluso en ausencia de valoraciones, lo cual permitiría a la empresa crear una herramienta de valoración confiable y comercializable para otros usuarios interesados."

Tratamiento de datos

En el dataset que estamos trabajando hay presencia de valores nulos en las columnas ‘genre’, ‘type’, ‘episodes’ y ‘rating’. Para los atributos (que son todas las variables antes mencionadas con la excepción de rating, que es nuestro vector objetivo) imputaremos los valores nulos por medio de la moda de cada variable, ya que en ninguno de los casos los valores nulos superan el 5% del total En caso de que el porcentaje de nulos hubiese sido mayor al 5% se procede a analizar e identificar una posible solución para cada caso, como una recodificación, por ejemplo.

También se estandarizaron las variables con standard y robust scaler, además se transformaron con logaritmo y raíz cuadrada. De igual manera, se observó que aproximadamente el 99,7% de los datos se encuentran dentro de tres desviaciones estándar, por lo que se utilizaron tres desviaciones estándar como criterio de selección para el tratamiento de outliers o valores atípicos, estos son valores que están muy alejados de los demás datos. Con todas estas combinaciones se llegó a trabajar con 16 datasets que se listarán a continuación, cabe destacar que de los dataset 1 al 14 se trabajó con la variable type aplicándole un método llamado get_dummies, este método lo que hace es transformar variables categóricas (son variables que representan diferentes grupos y estos grupos no se pueden medir) a variables numéricas de tal forma que quedan finalmente como variables binarias o variables que tienen solo dos valores unos y ceros. Y los datasets 15 y 16, esta variable type se trabajo de manera jerarquizada, esto quiere decir que a cada grupo que pertenece a esta variable categórica se le asigno un valor único, por ejemplo, al grupo 1 se le asignó el valor de 1, al grupo 2 el valor asignado fue el 2, y así sucesivamente, pero esta asignación no es necesario que siga ningún tipo de orden. De todas maneras, en la tabla 1 se muestra la jerarquización que se utilizó, en este caso se hizo de acuerdo con el rating.

1. df_clean: Dataframe que no considera ningún tipo de estandarización, solo limpieza de datos.
2. df_std: Dataframe en que sus variables continuas están estandarizadas con StandardScaler.
3. df_robust: Dataframe en que sus variables continuas están estandarizadas con RobustScaler.
4. df_log: Dataframe en que sus variables continuas están estandarizadas con logaritmo.

5. df_sqrt: Dataframe en que sus variables continuas están estandarizadas con la raíz cuadrada.
6. df_clean_sin_out: Dataframe con un tratamiento para disminuir los outliers y en que no considera ningún tipo de estandarización, solo limpieza de datos.
7. df_std_sin_out: Dataframe con un tratamiento para disminuir los outliers y que sus variables continuas están estandarizadas con StandardScaler.
8. df_robust_sin_out: Dataframe con un tratamiento para disminuir los outliers y que sus variables continuas están estandarizadas con RobustScaler.
9. df_log_sin_out: Dataframe con un tratamiento para disminuir los outliers y que sus variables continuas están transformadas con logaritmo.
10. df_sqrt_sin_out: Dataframe con un tratamiento para disminuir los outliers y que sus variables continuas están transformadas con la raíz cuadrada.
11. df_rating_log: Dataframe en que su vector y variables continuas están transformadas con logaritmo.
12. df_rating_sqrt: Dataframe en que su vector y variables continuas están transformadas con la raíz cuadrada.
13. df_rating_log_sin_out: Dataframe con un tratamiento para disminuir los outliers en que su vector y variables continuas están transformadas con logaritmo.
14. df_rating_sqrt_sin_out: Dataframe con un tratamiento para disminuir los outliers en que su vector y variables continuas están transformadas con la raíz cuadrada.
15. df_log_jerarquia: Dataframe en que su vector y variables continuas están transformadas con logaritmo, la variable type esta jerarquizada.
16. df_sqrt_jerarquia: Dataframe en que su vector y variables continuas están transformadas con la raíz cuadrada, la variable type esta jerarquizada.

Tabla 1: Nivel de jerarquización para type

type	rating	jerarquización
TV	6,9023	1
Special	6,5235	2
OVA	6,3752	3
Movie	6,3181	4
ONA	5,6433	5
Music	5,589	6

Fuente: Microsoft Excel, elaboración propia

Análisis Descriptivo

Se hizo un análisis inicial de las variables quebradas por tipo, como se aprecian en la figura 1, podemos observar que el rating promedio es de 6,5 aproximadamente, donde el tipo TV es el mejor valorado, seguido de OVA y movie, cabe mencionar que los 12,2k de observaciones están bastante distribuidas entre TV, OVA y movie.

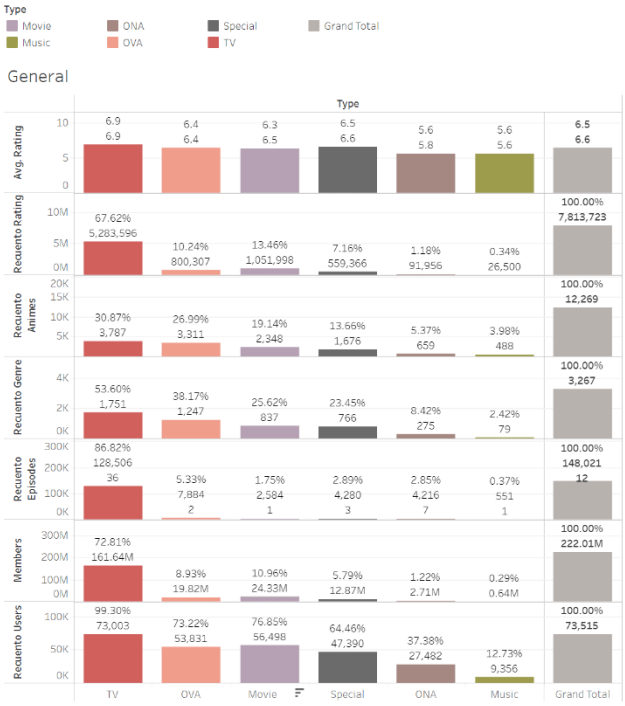

Figura 1: EDA
Fuente: public.tableau, elaboración propia

Además, del gráfico de la figura 1 podemos concluir lo siguiente:

- Rating cantidad: La mayor cantidad de rating (67,6%) es para TV, seguido de Movie. Dejando en claro que para el animé el tipo más fuerte es TV.
- Recuento animés: De un total de 12.2k de animés, la distribución es bastante pareja para TV, OVA y Movie.
- Recuento Genre: Podemos observar que hay una gran cantidad de géneros, debido a que un animé puede tener muchos tipos de géneros en un mismo animé, por lo que el número enorme que observamos es debido a la permutación de combinaciones, se verá este tema más adelante.
- Recuento episodes: Podemos ver que en total promedio hay 12 episodios por animé, y que para TV es 36, esto es un indicador de que su varianza es bastante grande, a diferencia de rating.
- Members: Podemos ver una enorme varianza, en donde la distribución de TV es abismalmente mayor que cualquier otro, confirmando los datos de Recuento Rating.
- Recuento users: Podemos observar que, de un total de 73.515 usuarios únicos, el 99,3% de ellos corresponden a TV, luego como segundo lugar tenemos que el 76,85% de ellos son de Movies.

Además de boxplots, figura 2 y 3, en el que comparamos un dataset sin y con tratamiento de outliers, respectivamente. Se observa una clara disminución de los outliers en la figura 3, cabe destacar que el dataset utilizado en estas figuras corresponde a un conjunto de datos en el que las variables fueron transformadas mediante raíz cuadrada.

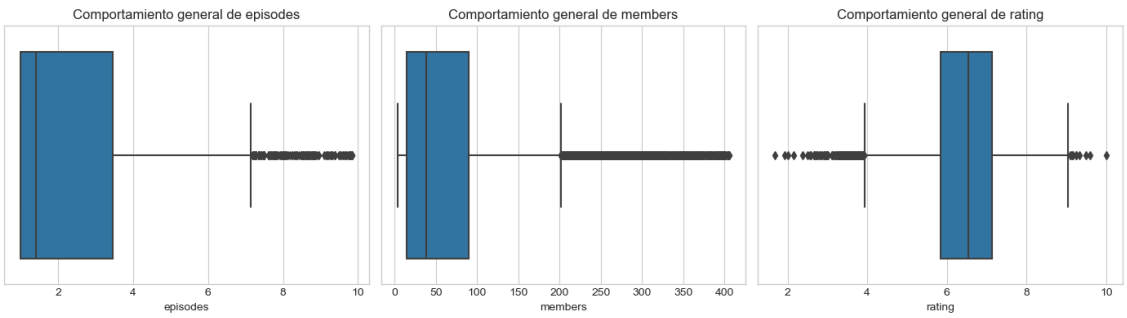

Figura 2: Boxplot del dataset sin tratamiento de outliers
Fuente: Matplotlib, elaboración propia

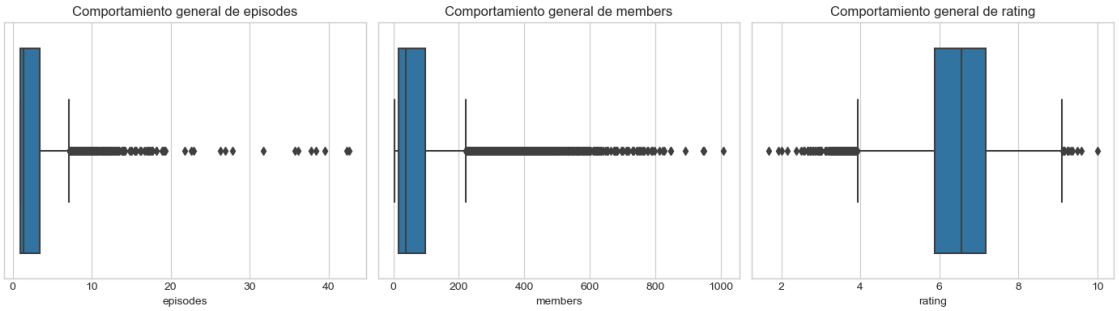

Figura 3: Boxplot del dataset con tratamiento de outliers
Fuente: Matplotlib, elaboración propia

A continuación, se muestra en la figura 4 una matriz de correlaciones entre las variables rating, members y episodes. Esta matriz considera un dataset con un conjunto de datos en el que las variables fueron transformadas mediante logaritmo. Se puede apreciar una fuerte correlación entre las variables miembros y rating con un coeficiente de 0.65, caso contrario sucede para las variables miembros y episodios con un coeficiente 0.27. No se observa una mayor diferencia de los coeficientes de correlación con respecto al dataset que tiene tratamiento de outliers (derecha) con el que no (izquierda). De todos los coeficientes de correlación, particularmente de este dataset, es el que las variables miembros y rating están más asociados.

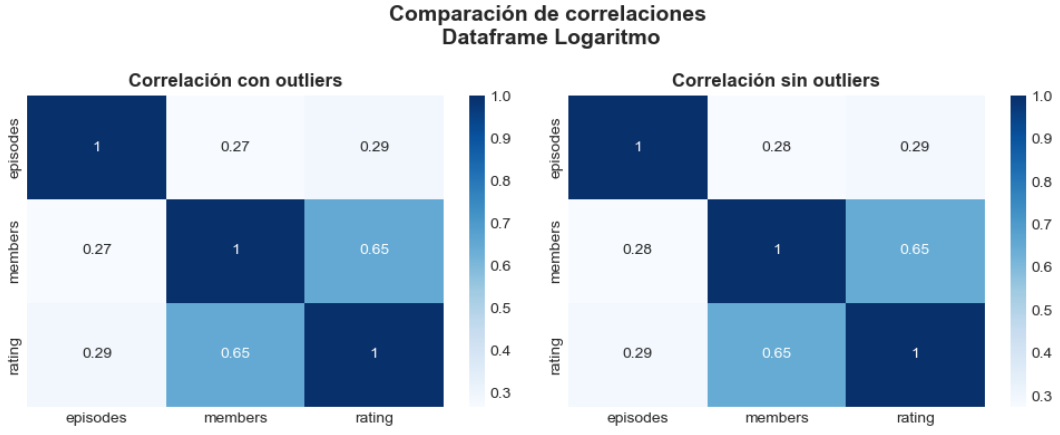

Figura 4: Matriz de correlaciones entre rating, episodes y members
Fuente: Matplotlib, elaboración propia

Criterios de selección e implementación de modelos

Al ser este un problema de regresión, se utilizarán tres criterios de selección o métricas para seleccionar el mejor modelo, estas son MAE, MAPE y R^2.

El MAE es, del inglés Mean Absolute Error, error absoluto medio, es una medida de diferencia promedio entre los valores reales y las predicciones, se utiliza para evaluar la precisión de un modelo de regresión, de manera simple, el MAE calcula cuánto se desvían, en promedio, las predicciones de un modelo de los valores reales. Cuanto menor sea el valor del MAE, mejor será la precisión del modelo. En términos prácticos, si tenemos un MAE de 0,6; ello significa que en nuestra predicción del rating puede que haya un error de 0,6 hacia arriba o hacia abajo, es decir si nuestra predicción fue de 6,5; el valor real puede oscilar entre 5,9 y 7,1.

El MAPE es, del inglés Mean Absolute Percentage Error, error porcentual absoluto medio, es una medida para evaluar el rendimiento de un modelo en términos de porcentaje de error promedio, de manera simple, el MAPE calcula el promedio de los errores porcentuales de las predicciones en relación con los valores reales. Cuanto menor sea el valor del MAPE, mejor será la precisión del modelo. En términos prácticos, si tenemos un MAPE del10%, ello significa que en nuestra predicción del rating puede haber una variación del 10% en promedio en relación con el valor real, es decir si nuestra predicción fue de 6, el valor real puede oscilar entre 5,4 y 6,6.

El R^2 o coeficiente de determinación, es una medida para evaluar qué tan bien se ajusta un modelo a los datos observados, en palabras sencillas el R^2 representa la proporción de la variabilidad de la variable dependiente que puede explicarse por el vector objetivo en el modelo, es un número entre 0 y 1, donde 0 significa que el modelo no explica nada de la variabilidad y 1 significa que el modelo explica toda la variabilidad. Cuanto más cercano sea al valor de 1, mejor es el ajuste del modelo y mayor es la capacidad para explicar la variabilidad de la variable dependiente.

De manera práctica se utilizarán de manera inicial 7 modelos con 16 datasets distintos, de todos estos modelos ejecutados se seleccionarán los mejores 5 modelos y el mejor dataset, con respecto a esto último, en la tabla 2 se muestra de manera ordenada según el MAPE los 16 datasets. Observamos que la mejor métrica, es decir un MAPE más bajo, es para el dataset en que su vector y variables continuas están transformadas con la raíz cuadrada. Sin embargo, se utilizará el2do dataset, esto debido a que, si bien hay una diferencia en el MAPE, esta no es significativa ni importante, no así el R^2 explica más en el 2do dataset. En definitiva, se utilizará el dataset en que su vector y variables continuas están transformadas con la raíz cuadrada, la variable type esta jerarquizada.

En cuanto a los modelos que utilizaremos son:

1. SVR
2. Gradient Boosting

3. Random Forest
4. Elastic Net
5. Voting Regressor

Resultados

El mejor modelo se muestra, gráficamente en la figura 5, y visualmente en la tabla 2. Las métricas, principalmente el MAPE muestra un comportamiento estable en todos los modelos. Además, se puede observar que los modelos que utilizan árboles de decisión como base, muestran un mejor desempeño que los demás modelos. Como observación final, el mejor modelo se logra mediante Gradient Boosting Regressor que claramente tiene el R^2 más alto y además el MAE y MAPE más bajos, tanto en el conjunto de entrenamiento como en el de prueba.

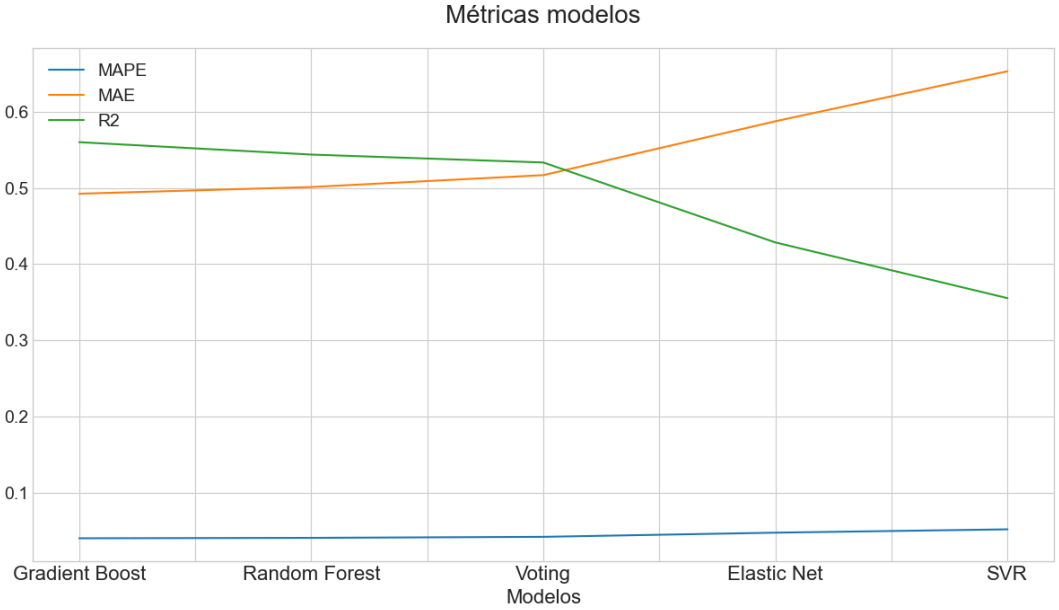

Figura 5: Métricas de los 5 mejores modelos
Fuente: Matplotlib, elaboración propia

Tabla 2: Resumen de las métricas de los 5 mejores modelos

Modelos	Métricas					
	MAPE		MAE		R^2	
	Train	Test	Train	Test	Train	Test
SVR	0,0497	0,0525	0,615	0,6528	0,4598	0,3555
Gradient Boosting	0,0382	0,0408	0,4626	0,423	0,6407	0,5598
Random Forest	0,0395	0,0414	0,477	0,5011	0,6096	0,5436
Elastic Net	0,0431	0,0482	0,5204	0,5872	0,5446	0,4284
Voting Regressor	0,0424	0,0427	0,5109	0,5167	0,5609	0,5332

Fuente: Matplotlib, elaboración propia

Como se dijo con anterioridad, hay una pequeña muestra de aproximadamente 230 observaciones que no tienen rating, por lo que se utilizará esa muestra para probar nuestro

modelo, en la tabla 3 se muestra un extracto de las predicciones realizadas por nuestra herramienta.

Tabla 3: Extracto del dataframe con predicciones del rating

Nombre	Género	Tipo	Episodios	Rating	Predicción Rating
Inazma Delivery	acción, comedia	TV	10	Nulo	6,1816
Nananin no Ayakashi	comedia, super natural	TV	1	Nulo	5,6424
Gintama	acción, comedia	TV	1	Nulo	7,4607
One punch man 2	acción, comedia	TV	1	Nulo	7,3674
Steins gate	Sci-fi, thriller	TV	1	Nulo	7,2639
Nuki doki	Hentai	OVA	1	Nulo	5,7997
Sagurare otome	Hentai	OVA	1	Nulo	5,0001
Saimin Class	Hentai	OVA	1	Nulo	5,7269
Shikkoku no shaga	Hentai	OVA	1	Nulo	5,7269
Taimanin asagi 3	Demon, Hentai	OVA	1	Nulo	5,9649

Fuente: Microsoft Excel, elaboración propia

Análisis

En la figura 6 se está representando una comparación de un dataframe sin ninguna transformación de logaritmo, raíz cuadrada ni de jerarquización del atributo type aplicándole un modelo gradient boosting sin hiperparámetros con el modelo que mencionamos anteriormente.

Vemos que se logra reducir el error porcentual en un 4% y a la vez el MAE se logra disminuir en un 1%, lo cual indica una clara mejora. El R2 y el MAE tienen una mejora muy leve del 0,4% y 0,7% respectivamente.

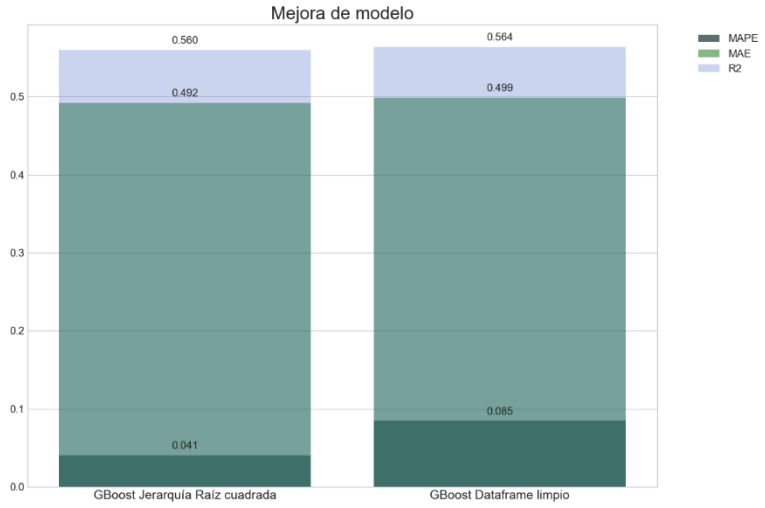

Figura 6: Comparación de modelos GB
Fuente: Matplotlib, elaboración propia

Lo que se muestra en la figura 7 son los 10 animes con peor valoración y los 10 animes con mejor valoración, si observamos los 10 peores se ve que de todas formas tienen buena valoración entre 4 y 5 puntos. Llama la atención que dentro de los mejores aparecen animes bastante conocidos como shingeki, evangelion que es un anime bastante antiguo, entre otros y en los peores valorados aparecen animes no tan conocidos.

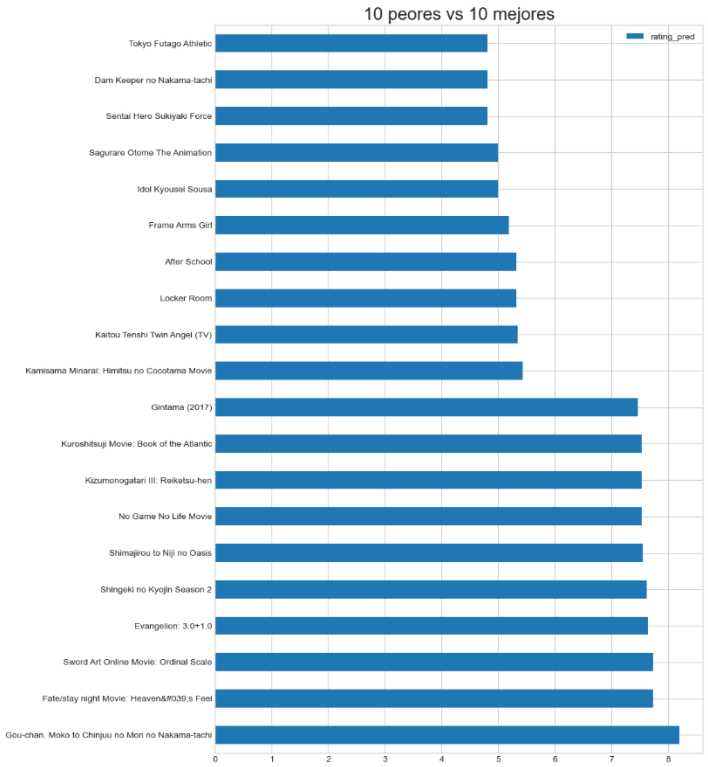

Figura 7: 10 peores vs 10 mejores animés
Fuente: Matplotlib, elaboración propia

Para el proyecto se realizaron dos árboles de decisiones, uno considerando todos los atributos, figura 8, y un 2do árbol que no considera los atributos ni episodios ni miembros. Como observamos en la figura 8, en los recuadros marcados en rojo, se trata de una muestra que corresponde al 22,5% del total, logra un error de 0,56 y un rating promedio de 6,19. Este grupo marcado tiene la característica de estar con un número de miembros mayor a 258 y menor o igual a 2548 miembros, además de contar con una cantidad menor o igual a 13 episodios. El funcionamiento del árbol de decisión de la figura 38, es que con todas las variables del dataset se busca minimizar el error con todos los puntos de corte, en este caso el MAE, entonces se va probando la cantidad de miembros desde la cantidad mínima hasta la máxima y va calculando el MAE por dentro, hace lo mismo para episodios y todas las demás variables; y se elige la variable con el punto de corte que tenga el menor MAE, en el caso que se muestra se eligió la variable miembros justo en el corte 2548, que es donde hace la mejor separación, en este punto si cumple o no esta condición se va hacia la izquierda o si no hacia la derecha; este proceso se repite con la misma u otra variable; el número de separaciones se definió en 3 (max depth).

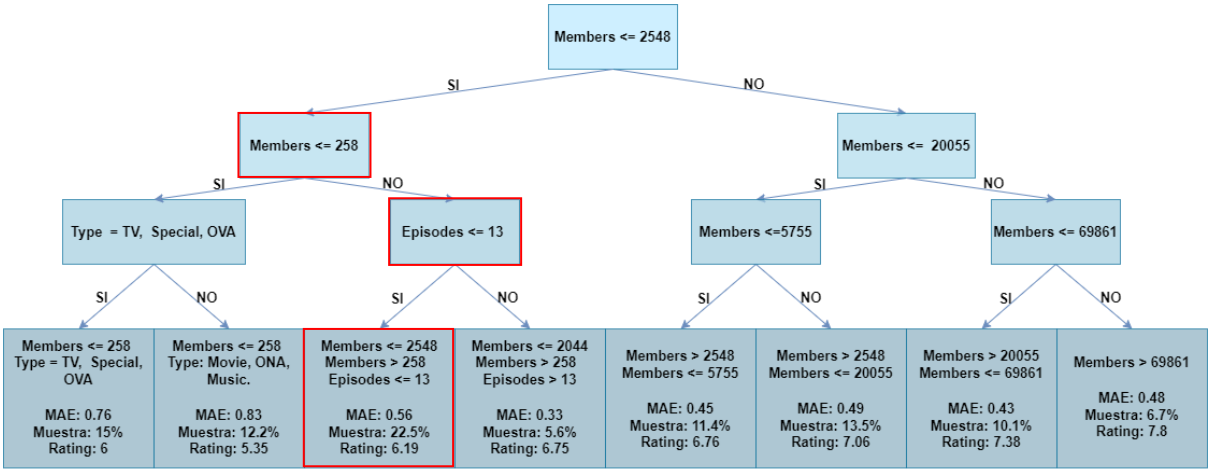

Figura 8: Árbol de decisión
Fuente: Matplotlib, elaboración propia

Mejoras

Una de las formas de mejorar el modelo, es la constante actualización de los datos, se recomienda que después de cada temporada de animés, se agreguen a la base de datos estos animés que salieron junto con su respectiva valoración real para ir aumentando la cantidad de observaciones y así cuando se modele nuevamente para predecir el rating de una nueva temporada de animés, contar con más datos para el entrenamiento del modelo.

Modelo de negocios

Para aprovechar al máximo esta herramienta es recomendable agrupar los animés que obtuvieron una buena predicción, es decir, los animés que nuestra herramienta predijo iban a tener valoraciones altas y así aprovechar esa información para colocar a ese animé en una franja horaria adecuada para que tenga una mayor audiencia y así optimizar las ganancias en publicidad.

Conclusiones

El documento de investigación presenta un análisis detallado sobre la posibilidad de desarrollar un sistema predictivo para estimar puntajes de series y películas de Anime sin valoraciones utilizando técnicas de aprendizaje automático. El estudio abordó una pregunta de investigación y una hipótesis de trabajo, demostrando que es factible crear un modelo de estimación preciso y confiable.

A través del tratamiento de los datos, la imputación de valores nulos y la transformación de variables, se crearon diferentes conjuntos de datos para evaluar el desempeño de varios modelos de regresión. El modelo de Gradient Boosting Regressor se destacó como el mejor, mostrando un alto coeficiente de determinación (R^2) y bajos errores absolutos medios (MAE) y errores porcentuales absolutos medios (MAPE).

El análisis descriptivo reveló una correlación significativa entre el número de miembros y el rating de los Anime, lo cual proporciona información valiosa sobre la influencia de la popularidad en la valoración. Se sugirieron posibles mejoras, como la actualización constante de los datos y la incorporación de nuevas observaciones, así como la propuesta de un modelo de negocios basado en la herramienta desarrollada.

En resumen, el estudio demuestra que es posible estimar puntajes de Anime sin valoraciones con precisión utilizando técnicas de aprendizaje automático. Esto ofrece oportunidades prometedoras para la comercialización de Anime y la toma de decisiones basada en las predicciones de valoración. El documento proporciona una base sólida para futuras investigaciones y desarrollos en este campo.