PROPUESTAS DE PROYECTO - BIG DATA

Estudiante: Yoksan Varela Cambronero

3 de Julio, 2024

1. Investigación preliminar

1.1. Primera propuesta: Un nuevo desarrollador de video juegos

1.1.1. Problema a resolver

De acuerdo con el artículo publicado por *Fortune Business Insights* (https://www.fortunebusinessinsights.com/video-game-market-102548), el valor de mercado de los video juegos era de aproximadamente de $250 mil millones para 2022, y se espera un crecimiento a una tasa compuesta anual (CAGR por sus siglas en inglés) del 13.1 %, lo que implica que alcanzará los $665 mil millones para 2030.

Una nueva desarrolla costarricense de video juegos quiere formar parte de este crecimiento de mercado y obtener una parte de ese dinero. Dicha desarrolladora hizo una inversión importante en todo lo referente a infraestructura para el desarrollo de un nuevo juego, pero pero esto produjo un problema: Para poder recuperar su inversión, los cálculos realizados indican que este nuevo viejo juego tiene que vender más de 100 mil copias, por lo tanto, esta empresa necesita entender cuál es la receta que les permita llegar a esa meta tan agresiva.

1.1.2. Solución propuesta

Se le plantea a dicha empresa crea una sistema de aprendizaje automático que pueda predecir si el nuevo video juego va a llegar a vender la cantidad deseada o, en su defecto, no va a alcanzar las 100 mil copias en ventas.

1.1.3. Fuentes de datos

Para hacer este estudio de mercado, se van a utilizar dos conjuntos de datos tomados de Kaggle:

1.1.4. Primer set: Video Games Dataset

URL: https://www.kaggle.com/datasets/beridzeg45/video-games
Este conjunto de datos cuenta con información de 14035 video juegos. Los atributos que contiene dicho documento es:

1. *Title:* Nombre del video juego.
2. *Release Date:* Fecha del lanzamiento del video juego.
3. *Developer:* Desarrollador del video juego.
4. *Publisher:* Editor del video juego (quien provee el financiamiento del mismo, incluyendo mercadeo y promoción).
5. *Genres:* Tipo a cual pertenece el video juego, por ejemplo: acción, aventura, roles, deportes, etc.
6. *Product Rating:* Es la calificación del producto, en este caso, quienes deberín jugar el juego: apto para todo publico, mayores de cierta edad, adultos, etc.
7. *User Score:* Puntaje de evaluación impuesto por el jugador final.
8. *User Rating Count:* Conteo de veces que el juego recibió un puntaje por parte de un jugador.
9. *Platforms info:* Información de las plataformas donde se juega el juego, pero en forma de diccionario.

1.1.5. Calidad de datos:

Este análisis empieza con entender la cantidad de valores NaN presentes en el conjunto de datos. Los resultados se tabulan a continuación:

Atributo	**Conteo de NaN**
Title	21
Release Date	64
Developer	138
Publisher	138
Genres	21
Product Rating	3050
User Score	2341
User Rating Count	2756
Platforms info	0

Cuadro 1: Conteo de valores NaN en el set de datos Video Game Dataset

Para lidiar con los valores NaN, se decide reemplazar los NaN en con "Not Specified.[en] los siguientes atributos: Genres, Publisher y Product Rating, dado que se consideran atributos de interés para el estudio debido a su conocido impacto en las ventas de un video juego. Además, para el atributo Product Score se decide reemplazar los NaN por el valor medio del atributo, por la misma razón explicada anteriormente.

Los siguientes atributos fueron descartados por lo siguiente:

1. User Ratings Count: Este conteo carece de información importante, que la vez es redundante con User Score.

2. Platforms Info: Este atributo se puede obtener del otro set de datos que se explicará posteriormente de una forma más fácil.

3. Developer: Dado que el cliente es un desarrollador en sí, no esta interesado en saber quién desarrolló el juego para evitar comparaciones con desarrolladores más establecidos y con mayor capital de operación.

Ya con estos cambios hecho, se procede a descartar las instancias que tienen NaN en Title y Release Date para obtener el set de datos limpios y proceder a analizar de forma estadística. Al final de estos pasos, quedaron 13991 instancias.

1.1.6. Análisis del primer set de datos:

Este conjunto de datos presenta 4752 fechas de lanzamiento distintas, 2302 editores y 119 diferentes tipos de juego. Usar estadística descriptiva en estas categorías representa visuales sumamente difíciles de interpretar, por lo tanto, solo se van a analizar de forma visual el Product Rating y el User Score:

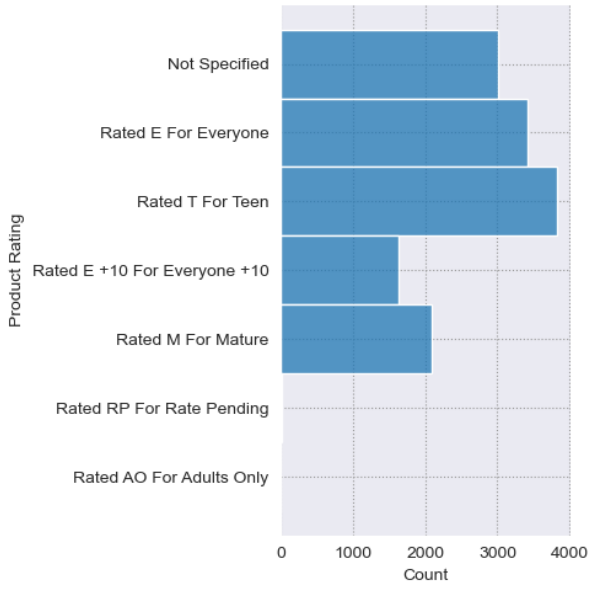

Figura 1: Histograma Product Rating

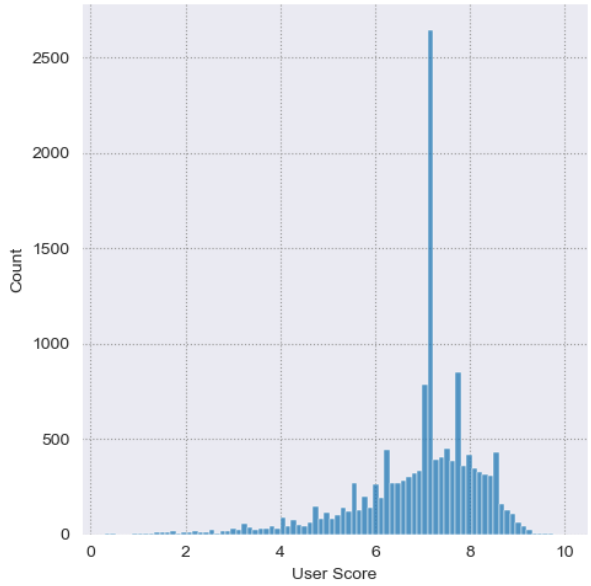

Figura 2: Histograma User Score

Aunque se puede notar que en 1 hay dos categorías que no tiene datos, se ve una distribución saludable entre el resto. Por otro lado, se puede apreciar un desbalance con uno de los ratings en 2. Este desbalanceo se debe al reemplazo que se hizo anteriormente de los NaN, ya que este atributo contaba con muchos de ellos. A este punto, como este set de datos se piensa utilizar con información adicional del próximo conjunto de datos, no hay razón para lidiar con este problema en este momento.

1.1.7. Segundo set: Video Game Sales 2024

URL: https://www.kaggle.com/datasets/asaniczka/video-game-sales-2024
Este es el conjunto de datos principal, ya que cuenta con la información de las ventas actualizada hasta

lo que se lleva de este año 2024. Cuenta con información de más de 64 mil video juegos. Los atributos en este set de datos son los siguientes:

1. *img:* Imagen de la portada del video juego.
2. *title:* Nombre del video juego. **Esta va a ser la llave a utilizar para poder unir los dos set de datos.**
3. *console:* La plataforma donde se corre el video juego.
4. *genre:* El tipo de juego, ejemplos: acción, deportes, estrategia, otros.
5. *publisher:* Editor del video juego.
6. *developer:* Desarrollador del video juego.
7. *critic_score:* Puntaje de la crítica.
8. *total_sales:* Ventas totales a nivel mundial. **De los datos de este atributo se va a generar el objetivo predictivo.**
9. *na_sales:* Ventas totales en Japón.
10. *jp_sales:* Ventas totales en Estados Unidos.
11. *pal_sales:* Ventas totales en Europa y África.
12. *other_sales:* Ventas totales en otras regiones de menor volumen.
13. *release_date:* Fecha del lanzamiento del video juego.
14. *last_update:* Fecha de la última actualización de los datos.

1.1.8. Calidad de datos:

De forma similar con el set de datos anterior, empieza con entender la cantidad de valores NaN presentes en el conjunto de datos. Los resultados se tabulan a continuación:

Atributo	**Conteo de NaN**
img	0
title	0
console	0
genre	0
publisher	0
developer	17
critic_score	57338
total_sales	45094
na_sales	51379
jp_sales	57290
pal_sales	51192
other_sales	48888
release_date	7051
last_update	46137

Cuadro 2: Conteo de valores NaN en el set de datos Video Game Sales 2024

Aunque inicialmente se tenían mas de 64 mil instancias, el hecho que 45094 de ellas no tengan la información relacionada con las ventas totales indica que hay que descartar la mayoría de las instancias. No obstante, la cantidad restante es suficiente para el desarrollo de este proyecto.

Con base a esto, se toman las siguientes decisiones para limpiar los datos: Primero, se eliminan los siguientes atributos:

1. img: No se necesita la imagen de la portada para este análisis.
2. Todos los relacionados con ventas que no sea total_sales: Dado que se quiere vender a nivel mundial, la información regional se descarta dando como prioridad la global.
3. critic_score: Es una práctica muy común que se pague por las buenas revisiones de los medios de la crítica, por lo tanto, este atributo podría introducir este ruido en el estudio.
4. release_date: Información presente en el set de datos anterior.
5. last_update: Información que no es relevante en este estudio.
6. developer: Se descarta por la misma razón mencionada anteriormente.

Luego de esto, de desechan todas las instancias con NaN, lo que deja un total de 18922 instancias con 5 atributos. Es importante mencionar en este punto que, aunque este set de datos se queda con 5 atributos y el anterior con 6, algunos de estos atributos categóricos como console (39), publisher (733 en este último set de datos) y genres (19 en este último set de datos) contienen mucho valores distintos, lo que da pie a un estudio con más de 800 atributos.

1.1.9. Análisis del segundo set de datos:

Haciendo un enfoque en las ventas, veamos como se ven para generar la clase objetivo:

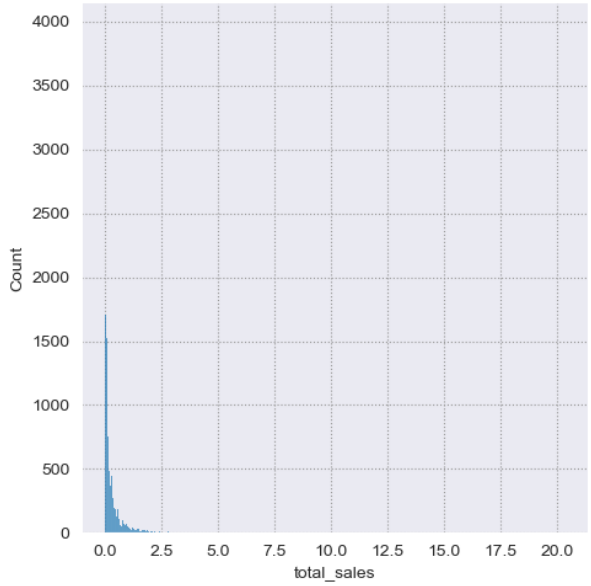

Figura 3: Total de Ventas Globales

En el histograma 3 se aprecia que muy pocos video juegos que han vendido cantidades exorbitantes de unidades (por ejemplo, GTA 5 ha vendido mas de 20 millones de unidades). Estos outliers causan que se pierda visibilidad del punto de interés, las cuales son 100 mil copias. Por lo tanto, si se ignoran los video juegos que vendan más de 1 millón de copias, se observa lo siguiente:

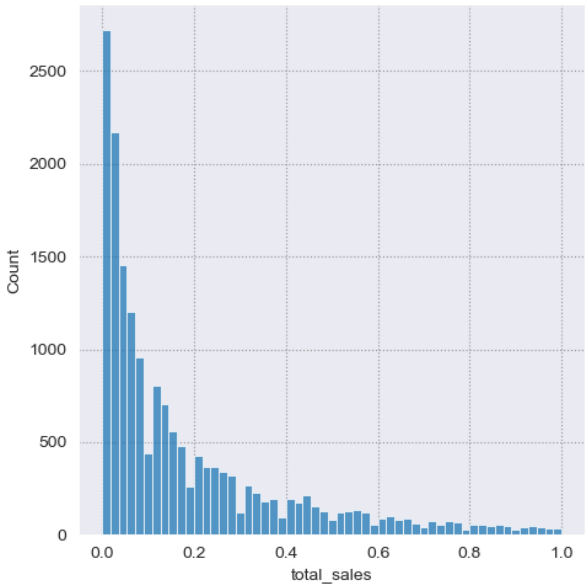

Figura 4: Ventas globales menores o iguales a 1 millón de copias

Ahora los datos siguen, de forma muy fiel, una función de densidad exponencial. Se podría pensar que hay un problema de balanceo de datos, pero si se hace un corte entre todos los juegos que no han vendido mas de 100 mil copias contra los que sí lograron esa meta, se aprecia lo siguiente:

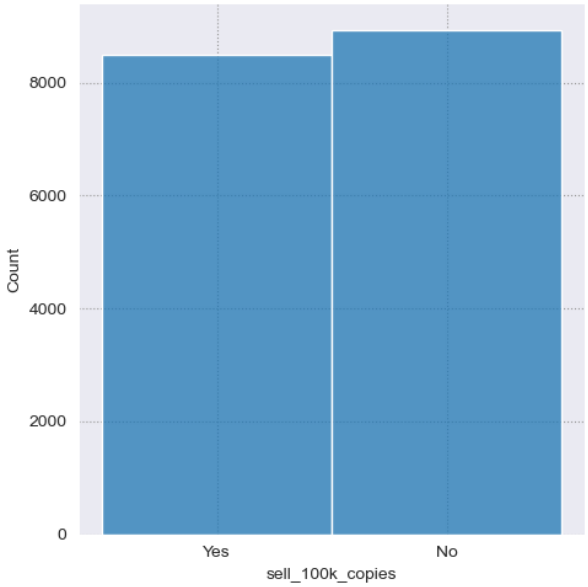

Figura 5: Total de ventas separado por juegos que vendieron más de 100 mil copias o no

De esta forma, podemos ver en 5 que la clase objetivo esta, prácticamente, perfectamente balanceada en este set de datos. **Este es el objetivo predictivo de esta propuesta.**

Para concluir este análisis de datos, verificamos que no hayan problemas de balanceo entre otros atributos de interés:

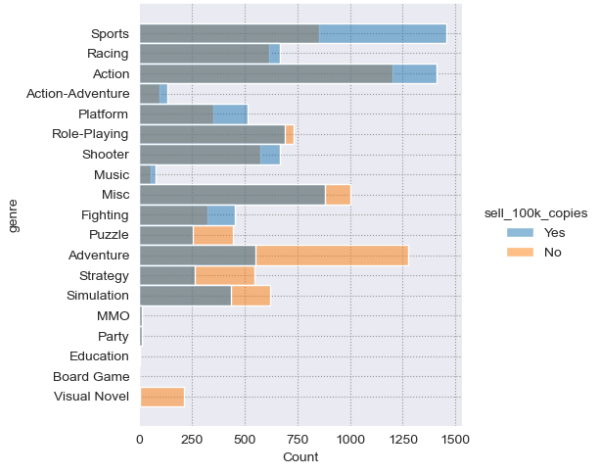

Figura 6: Histograma de los tipos de juegos, tomando en cuenta si vendieron 100 mil copias o no

Aunque en 6 se pueden ver 4 categorías predominantes, hay una buena distribución de datos entre varios tipos de video juegos.

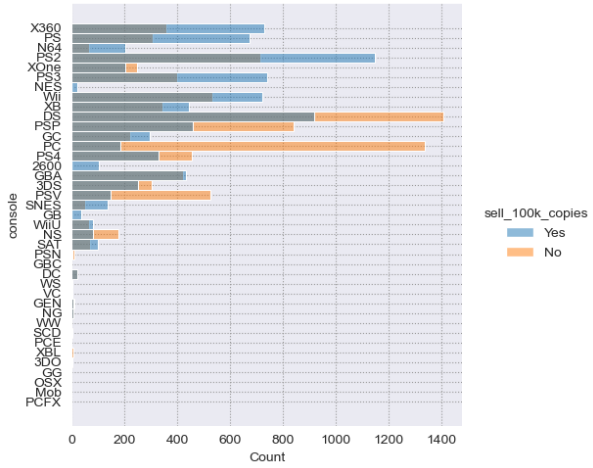

Figura 7: Histograma de las consolas, tomando en cuenta si vendieron 100 mil copias o no

Hay muchas consolas en 7 en las cuales se observa que los juegos no son jugados, pero en las más populares se observa una distribución que no sugiere un problema de balanceo.

1.1.10. Conclusión de la primera propuesta:

Este es un problema de categorización binaria que presenta un reto interesante, dado que se tiene un objetivo predictivo balanceado, pero a la vez se cuenta con muchos datos categóricos que presentar retos importantes a la hora de modelar.

1.2. Segunda propuesta: Desarrollar un animé que la gente vea por completo

1.2.1. Problema a resolver

El animé tiene su origen en Japón a inicios del siglo 20. Desde ese momento, la popularidad de dicho arte ha ido en aumento, y gracias a una exposición más global a partir del acceso a plataformas de streaming que cuenta con este contenido. En 2023, el animé reportó un mercado de mas de $32 mil millones de dolares, y por su tendencia en popularidad, se calcula que superará los $70 mil millones para 2032

(https://www.grandviewresearch.com/industry-analysis/anime-market).

Un estudio de animación en Japón esta interesado en entrar en el mercado, pero en un ambiente tan competitivo, la cantidad de animés es tal que los servicios de streaming están tomando la iniciativa de dar un bono económico al estudio que cree animés que sean completados (ya sea en su totalidad o cada temporada, es decir, que lo lleven al día) dado que quieren mantener una concurrencia de usuarios en sus plataformas.

1.2.2. Solución propuesta

Hacer un estudio de mercado para entender los factores básicos que hacen que los usuarios terminen o lleven al día un animé particular. Por lo tanto, se plantea hacer el análisis por medio de un software con aprendizaje automático que sea capaz de predecir si un animé va a ser completado o no.

1.2.3. Fuentes de datos

Para hacer este estudio de mercado, se van a utilizar tres conjuntos de datos tomados de Kaggle. Los 3 set de datos vienen del URL: https://www.kaggle.com/datasets/azathoth42/myanimelist. A pesar de provenir del mismo lugar, los 3 dataset se manejan de forma independiente.

Para este problema en particular, se van a usar los CSV que ya tienen un primer filtro (tiene el término *_filtered)*. A pesar de estar filtrados, hay una cantidad muy grande de datos: Uno de estos cuenta con más de 32 millones de instancias, por lo tanto, para efectos de este trabajo, ese CSV va a ser reducido (más detalles adelante). El autor de este post en Kaggle menciona que estos datos provienen MyAnimeList.net, el cual es solo uno de muchas opciones de streaming hoy en dia, asi que este estudio NO cubre otros servicios de streaming importantes como lo son Netflix o Crunchyroll.

1.2.4. Primer set: anime_filtered.csv

De entrada, este set de datos cuenta con 14474 instancias y 31 atributos distintos. Este número de atributos es manejable, pero muchos de ellos no aportan información valiosa. Por lo tanto, el primer paso es eliminar 19 de ellos.

1.2.5. Calidad de datos:

A continuación se listan y la razón por la cual se decidió eliminar del estudio:

1. *title_english:* Esta información es redundante con otro atributo llamado "title".
2. *title_japanese:* Esta información es redundante con otro atributo llamado "title", además de que no se quiere lidiar con caracteres en japones.
3. *title_synonyms:* Esta información es redundante con otro atributo llamado "title".
4. *favorites:* Este atributo es redundante con respecto a la variable objetivo, ya que un usuario que tenga un animé en su lista de favoritos muy probablemente lo va a completar de ver.
5. *image_url:* Es la imagen de la portada del animé, lo cual no es relevante para el estudio.
6. *broadcast:* Es la fecha en que se transmite el animé, ya esa en televisión japonesa o bien cuando sale un capitulo nuevo en el servicio de streaming. Este atributo esta sujeto a cambios continuos, por lo tanto, se volvería una variable difícil de estudiar.
7. *airing:* Indica si el animé esta en proceso de producción o no (si esta terminado o esta en curso). Este atributo es irrelevante dado que no es importante si está en producción o no, si no si los usuarios lo llevan al día o no.
8. *aired_string:* Se descarta por la misma razón anterior.
9. *aired:* Se descarta por la misma razón anterior.

10. *scored_by:* Este atributo indica quien le de dio una calificación al animé, lo cual es redundante con el usuario en sí.
11. *background:* Un poco mas de información del anime en si, pero sin ser relevante para el estudio o ser información que ya esta contenida en otro atributo.
12. *premiered:* Momento cuando el animé fue lanzado. De igual forma que con airing, lo importante es saber si fue completado o no, no cuando fue lanzando en sí.
13. *related:* Indica que otros animé son parecidos. Ya que se cuenta con una lista tan extensa de animé, no es necesario analizar esta relación.
14. *members:* Es el numero de usuarios que son miembros del animé. No queda claro a que se refiere este atributo, pero otro llamado favorite es semejante y da mejor información que éste.
15. *licensor:* Es el nombre de quien licencia la producción. Esta información es difícil de conseguir, por lo que alrededor del 77 % de este atributo tiene valores NaN, por lo que se decide descartarlo de entrada para no perder muchos de los datos.
16. *opening_theme:* Es la canción que suena durante la introducción de cada capitulo, lo cual no es relevante para el estudio actual.
17. *ending_theme:* Similar al anterior, es la canción que suena al final de cada capitulo. Se descarta por la misma razón anterior.
18. *studio:* Es el estudio quien desarrolla el animé. Similar a la propuesta anterior, es atributo no es de interés por parte del cliente dado que no quieren compararse contra las capacidad productivas de los competidores directos.
19. *rank:* Es el rango que tiene el animé, lo cual es redundante con la popularidad del mismo, por eso se descarta.

Después de este primer descarte de atributos, se quedan los siguientes 12:

1. *anime_id:* Es el ID de cada animé de la lista. **Este atributo se va a usar como llave de unión con el set de datos principal**
2. *title:* Nombre del animé, en ingles.
3. *type:* Predominan dos: TV y OVA, pero este atributo cuenta con mas de 6000 diferentes tipos. Por ahora no se va a analizar en profundidad y se va a tener que hacer una agrupación, pero se hará después de la unión con los otros dos set de datos (fuera de alcance para esta propuesta).
4. *source:* Indica cual fue la fuente de la historia del animé: una novela, un manga, otros.
5. *episodes:* La cantidad de episodios que tiene el animé.
6. *status:* Indica si la producción fue concluida o no.
7. *duration:* Cuanta es la duración promedio por episodio.
8. *rating:* Es la categoría del animé, por ejemplo: apto para todo público, mayores de cierta edad, para adultos, etc.
9. *score:* Es la calificación promedio del animé.
10. *popularity:* Un indicador de qué tan popular es el animé.
11. *producer:* El productor del animé. Cabe la pena resaltar que una producción de una animé normalmente requiere de varios productores a la vez.
12. *genre:* El genero del anime. Este atributo tiene la complejidad que un animé puede tener varios géneros.

Un analisis de NaN muestra que no hay mayor problema de esta índole (64 instancias en el atributo genre), por lo que las filas con NaN son descartadas.

1.2.6. Análisis del primer set de datos:

Después de este punto se hizo un análisis visual de los atributos, pero para efectos prácticos de esta propuesta, se muestran solo los resultados finales. Para mas detalle del estudio que se hizo, por favor referirse al Jupyter Notebook de la propuesta #2.

Estos resultados finales son:

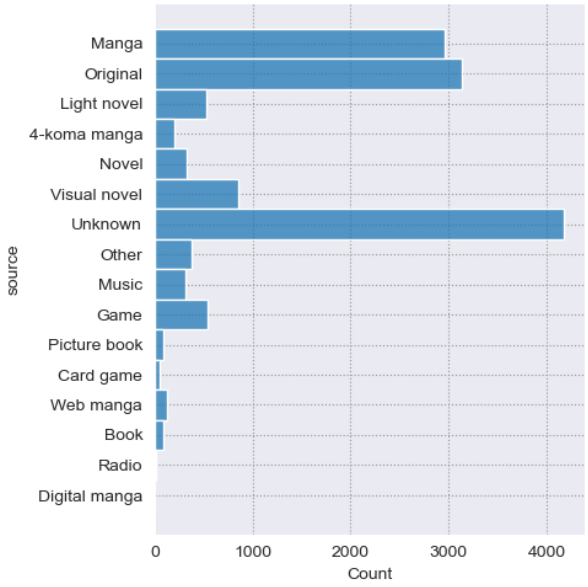

Figura 8: Histograma del atributo Source

En la figura 8 se nota que hay 3 principales sources para los animé: los mangas, historias originales o fuente desconocida. Para este atributo va a ser necesario hacer una estrategia de agrupamiento para lidiar con el desbalance. Dicho esto, esa estrategia no sera ejecutada en este momento de la propuesta.

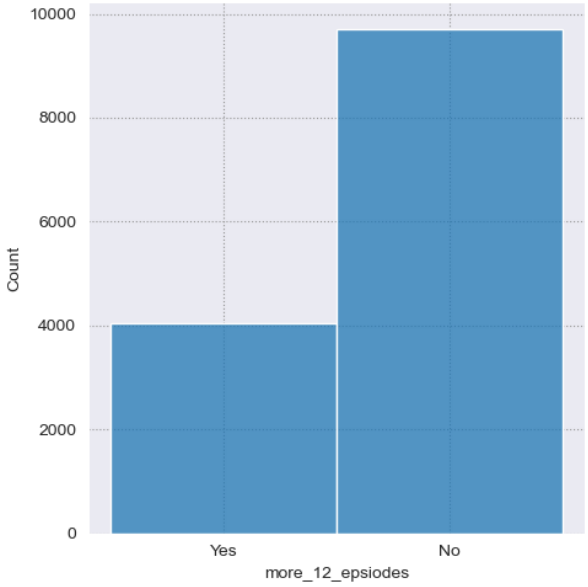

Figura 9: Histograma del atributo Episodes

Para el caso del atributo nuevo llamado more_12_episodes que fue creado a partir de episodes, la

figura 9 muestra un problema de desbalance en la cantidad de animes que tienen 12 o menos episodios. Como este set de datos es de soporte para el set de datos principal, todavía no se va a lidiar con este problema.

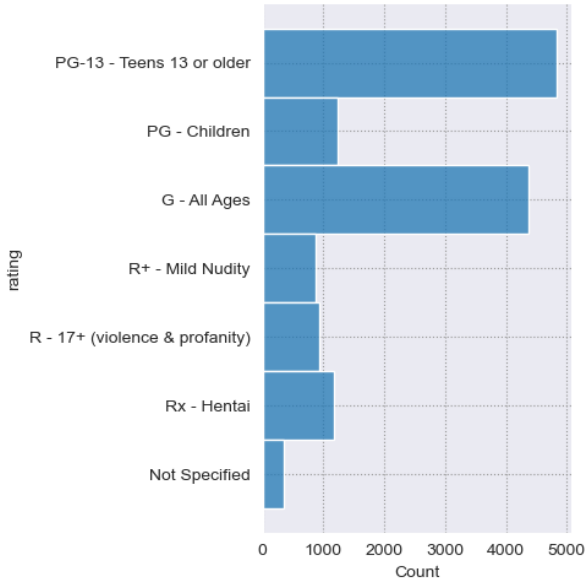

Figura 10: Histograma del atributo Rating

Finalmente, en la figura 10 se nota otro problema de desbalanceo con los datos. En este caso, se podría hacer una agrupación para evitar este problema y tener solo 3 categorías, pero para efectos de la propuesta, esto todavía no se va a realizar.

1.2.7. Segundo set: animelist_filtered_by_anime.csv

Este es el **set de datos principal** porque relaciona tanto los usuarios como el animé que vieron, y en cual estado lo tienen. Este CSV se creo a partir del que cuenta con mas de 23 millones de instancias. El set de datos cuenta con mas de 14 mil instancias y tiene 11 atributos, donde la mayoría están con valores de NaN o en cero, por lo que fueron descartados de entrada. Estos son: my_start_date, my_finish_date, my_rewatching, my_rewatching_ep, my_last_updated y my_tags.

Los atributos que quedan son los siguientes:

1. *username:* Este el nombre de usuario del servicio. **Esta es la llave de conexión el tercer ser de datos (por analizar más adelante)**
2. *anime_id:* ID del anime. **Esta es la llave de conexión con el primer set de datos (ya analizado a este punto)**.
3. *my_watched_episodes:* La cantidad de episodios visto por el usuario del animé en cuestión.
4. *my_score:* La calificación que el usuario le dio el animé.
5. *my_status:* Indica el estado del animé con respecto al usuario. Este atributo tiene 5 categorias: 1 = watching, 2 = completed, 3 = on hold, 4 = dropped y 6 = plan to watch. **De este atributo se va a crear el objetivo predictivo de esta propuesta.**

1.2.8. Calidad de datos:

En este punto, no es necesario trabajar en la calidad de los datos. Al eliminar las columnas sin información, el 100 % de los NaN se eliminar y el resto del set de datos es válido.

1.2.9. Análisis del segundo set de datos:

Para este set de datos se va a analizar solamente el objetivo predictivo. De forma individual, los estados se ven de la siguiente forma:

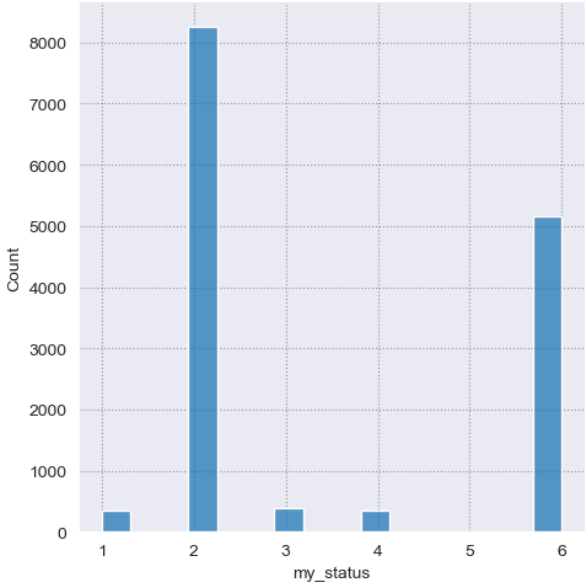

Figura 11: Histograma del atributo Status (individual)

De la figura 11 se concluye que hay dos estados que predominan sobre el resto: Completados (2) y Por Ver (6). El resto de los estados son bastante menores. Para crear el objetivo predictivo hay que agrupas todas la categorías juntos excepto la 2, para poder tener idea de como se ve la condición de completado contra el resto. Los resultados son los siguientes:

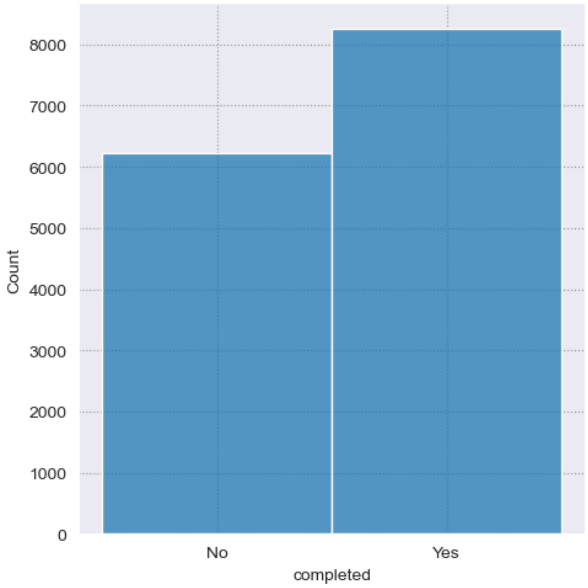

Figura 12: Histograma del atributo Status (agrupado)

Visto de esta forma, la figura 12 evidencia que el objetivo predictivo tiene un balance aceptable para poder hacer un categorización.

1.2.10. Tercer set: users_filtered.csv

Este último se de datos es la información de los usuarios, que cuenta con algunos atributos demográficos, y que tiene como finalidad dar mas atributos segundo set de datos.

Aquí se cuentan con mas de 116 mil instancias con 17 atributos. A continuación la lista de todos los atributos y en que consiste cada uno y si fue descartado o no:

1. *username:* Nombre del usuario. **Esta es la llave de unión con el segundo set de datos.**
2. *user_id:* ID del usuario. Se descarte porque es redundante con el nombre.
3. *user_watching:* Cantidad de animes que el usuario esta viendo actualmente. Esta cantidad no representa ninguna información importante, además de ser cubierto por el segundo set de datos, por lo tanto, se descarta.
4. *user_completed:* Cantidad de animes que el usuario ha completado. Esta cantidad no representa ninguna información importante, además de ser cubierto por el segundo set de datos, por lo tanto, se descarta.
5. *user_onhold:* Cantidad de animes que el usuario tiene en pausa. Esta cantidad no representa ninguna información importante, además de ser cubierto por el segundo set de datos, por lo tanto, se descarta.
6. *user_dropped:* Cantidad de animes que el usuario no va a continuar viendo. Esta cantidad no representa ninguna información importante, además de ser cubierto por el segundo set de datos, por lo tanto, se descarta.
7. *user_plantowatch:* Cantidad de animes que el usuario planea ver eventualmente. Esta cantidad no representa ninguna información importante, además de ser cubierto por el segundo set de datos, por lo tanto, se descarta.
8. *user_days_spent_watching:* Promedio de días que el usuario invierte viendo animé.
9. *gender:* Genero del usuario.
10. *location:* Localidad del usuario.
11. *birth_date:* Fecha de nacimiento del usuario.
12. *access_rank:* No se provee información, por lo tanto, se descarta.
13. *join_date:* Fecha en la que el usuario creó su cuenta en el servicio de streaming.
14. *last_online:* Última fecha registrada del usuario usando el servicio. Esta información no es relevante en el estudio y se descarta.
15. *stats_mean_score:* Promedio de calificación que da el usuario.
16. *stats_rewatched:* No queda muy claro de que se trata este atributo, se decide descartar.
17. *stats_episodes:* No queda muy claro de que se trata este atributo, se decide descartar.

1.2.11. Calidad de datos:

Al igual que con el set de datos anterior, no es necesario trabajar en la calidad de los datos dado que la cantidad de NaN es poca (6 en el peor de los casos), por lo que se decide eliminar esas instancias.

1.2.12. Análisis del tercer set de datos:

Este set de datos cuenta con casi 43 mil locaciones diferentes entre todos los usuarios, por lo tanto, no se puede mostrar un análisis visual de este atributo.

Veamos entonces los histogramas del género del usuario y su media de calificación:

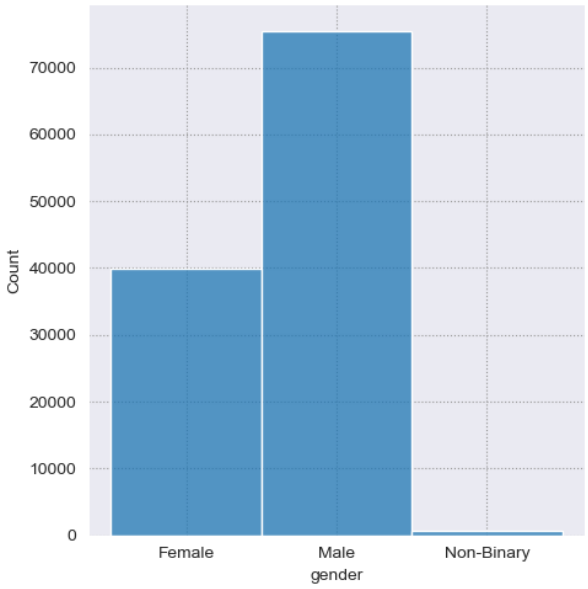

Figura 13: Histograma del género de los usuarios

Analizando los géneros, la figura 13 muestra una clara tendencia problemas de desbalanceo entre hombres y mujeres, y de estos dos contra No Binario. Como este es un set de datos de soporte, ya una vez unificado con los demás se podría evaluar esta condición de nuevo, dado que aquí hay 116 mil instancias pero en el set de datos principal solo se cuenta con mas de 14 mil, por lo tanto, este escenario podría cambiar.

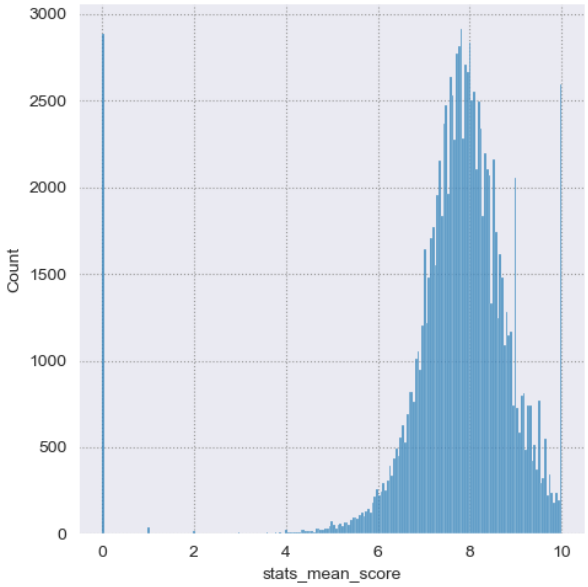

Figura 14: Histograma del mean score por usuario

Y para concluir, la figura 14 muestra 3 condiciones interesantes: una distribución normal de las medias entre 4 y 10, pero con una fuerte presencia de 0 o de 10. Siguiendo una estrategia similar al atributo anterior, es necesario unificar los datos para tener una mejor idea del escenario que se tiene con este atributo.

1.2.13. Conclusión de la segunda propuesta:

Esta propuesta tiene un objetivo predictivo con un balance aceptable, y cuenta con el reto de unir 3 set de datos distintos y evaluar condiciones de desbalanceo junto con atributos que presentan combinaciones de categorías, lo cual es necesario desempacar para tener una buena idea de los géneros y los productores, que a la vez generaría una cantidad muy importante de atributos por considerar.