NETFLIX

PROYECTO FINAL

PREDICTORES DE PRODUCCIONES

Presentado por: Fernando Garcias Corts y Ezequiel Taberner

CONTENIDOS

INTRODUCCION

Una película que hay que ver

Durante las siguientes diapositivas, veremos la necesidad y las posibles soluciones para realizar predicciones tanto en el contenido que preferira ver el publico asi como estimar la critica que podria llegar a tener una produccion que se agregue a la plataforma y si su contenido va a mantener el estandar de calidad de las producciones que se presentan.

SCIENTISTS

Ezequiel Taberner (Data Scientist)

Estudiante de Ingeniería Electrónica y apasionado por los datos. Decidió embarcarse de lleno en el apasionante mundo de los datos en el año 2022.

Junto a su compañero Fernando, eligieron trabajar en el set de datos de Netflix y abordar la resolucion de las problematicas presentadas

Fernando Garcias Corts (Data Scientist)

Desempeñandose en el sector de la automatizacion y control para la industria se encontró cara a cara con el mundo del ML y la IA. La curiosidad por este entorno lo llevo a incursionar de lleno con la ciencia de datos.

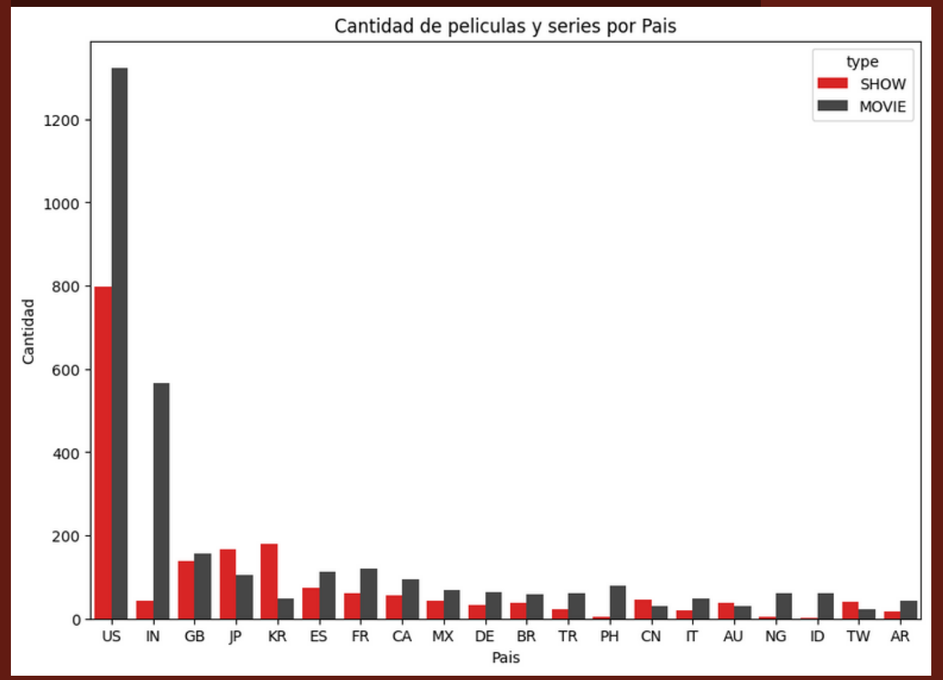
Fan del séptimo arte, no sorprendio su predisposición a trabajar con un set de datos de Series y Películas.

ANALISIS DE DATOS

01

Producciones por Pais

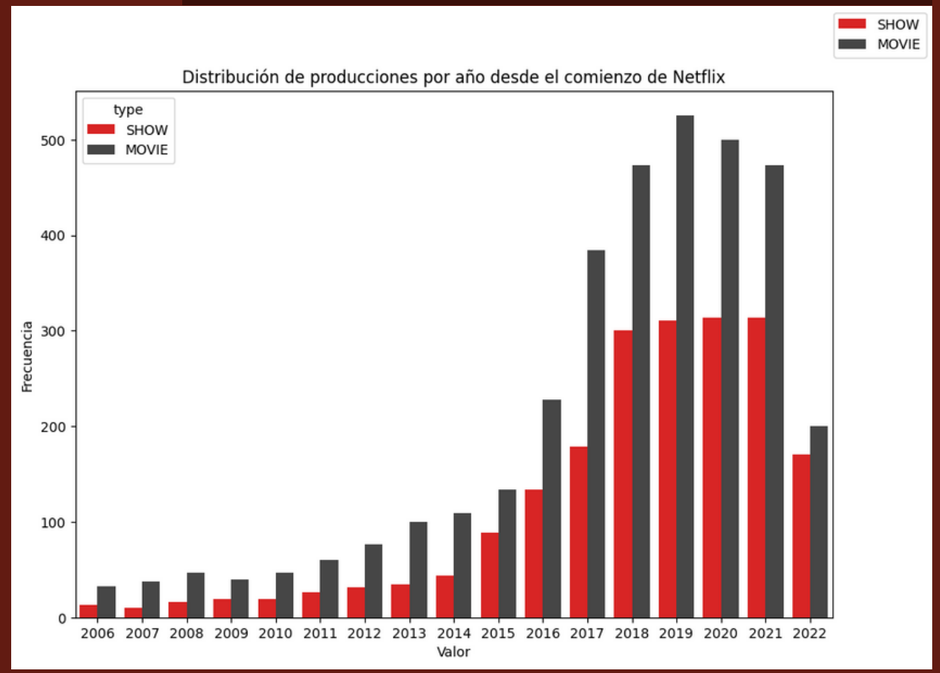
En este grafico podemos observar los 20 países que poseen mayor cantidad de producciones en la plataforma, de donde se destaca por amplia mayoría Estados Unidos



02

Cantidad por Año

Aquí se ven cuantas producciones se realizaron cada año desde el comienzo de Netflix. Como era de esperar, se ve un incremento considerable en los ultimos años.

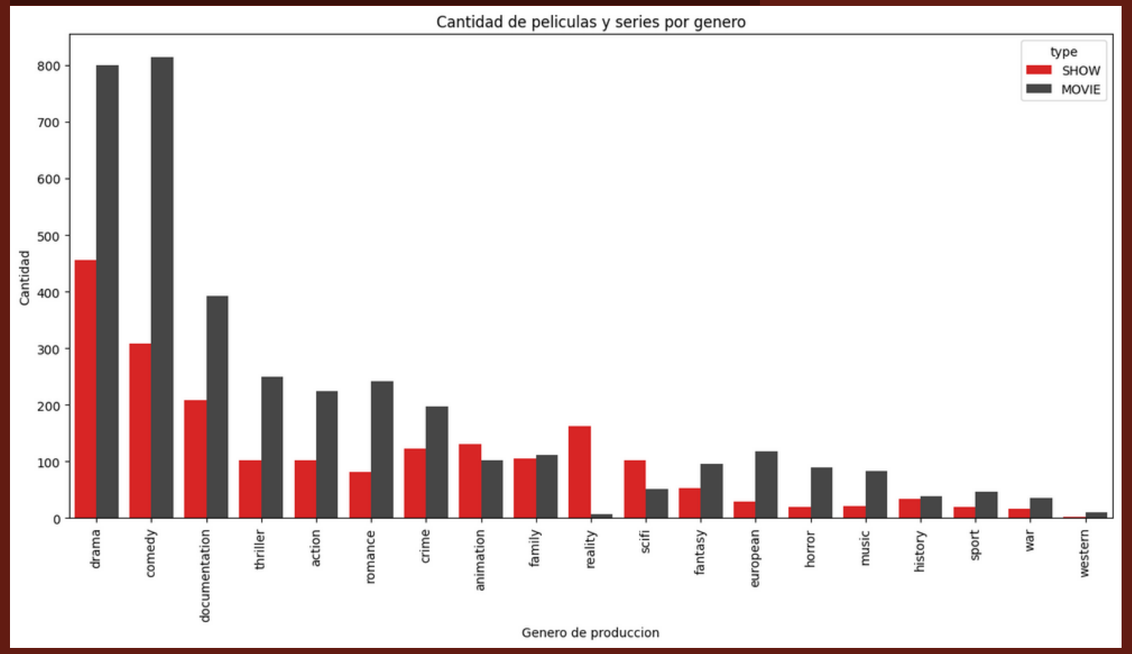


ANALISIS DE DATOS

03

Cantidad por Genero

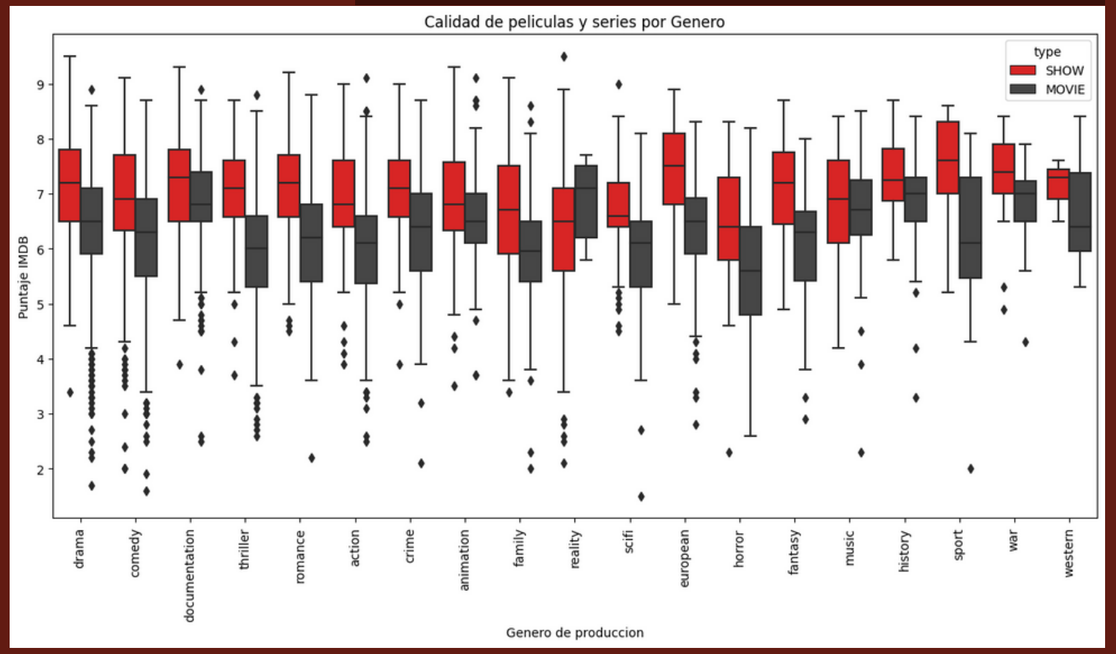
Si vemos la cantidad de producciones que se realizan de cada genero, vemos que claramente los dramas y las comedias son los poderdantes



04

Calidad por Genero

Como podemos ver, el genero de la producción es bastante decisivo a la hora del puntaje de la misma.. No así en todos pero si es marcado en la mayoria.

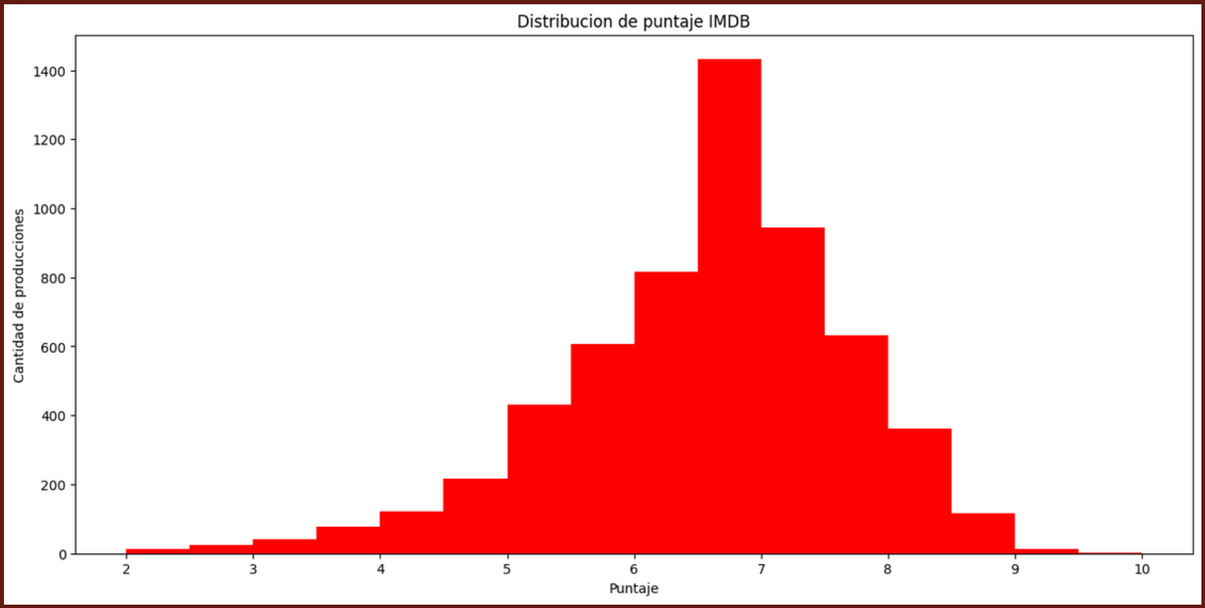


ANALISIS DE DATOS

05

Puntaje por producciones

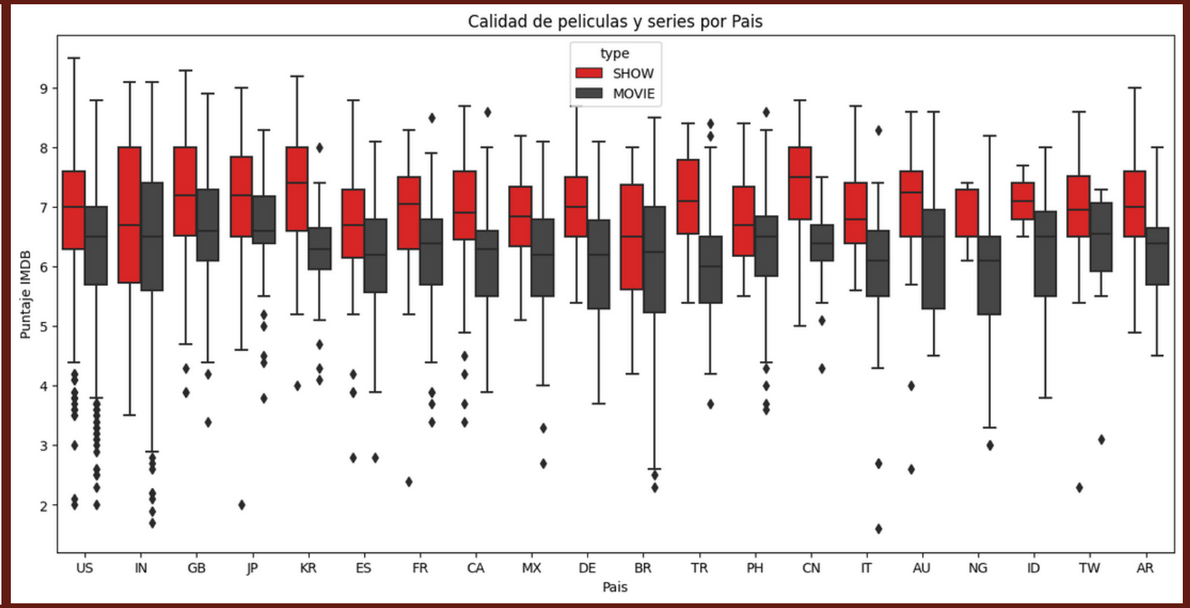
Como se observa, la mayoria de las producciones oscilan entre los 6 y los 8 puntos de IMDB, lo cual podriamos tomar como estandar de calidad



06

Calidad por Pais

De los 20 países con mas producciones, podemos ver que generalmente las series suelen tener mejor puntaje pero, las series de India y Gran Bretaña poseen mejores puntajes que las de Estados Unidos.

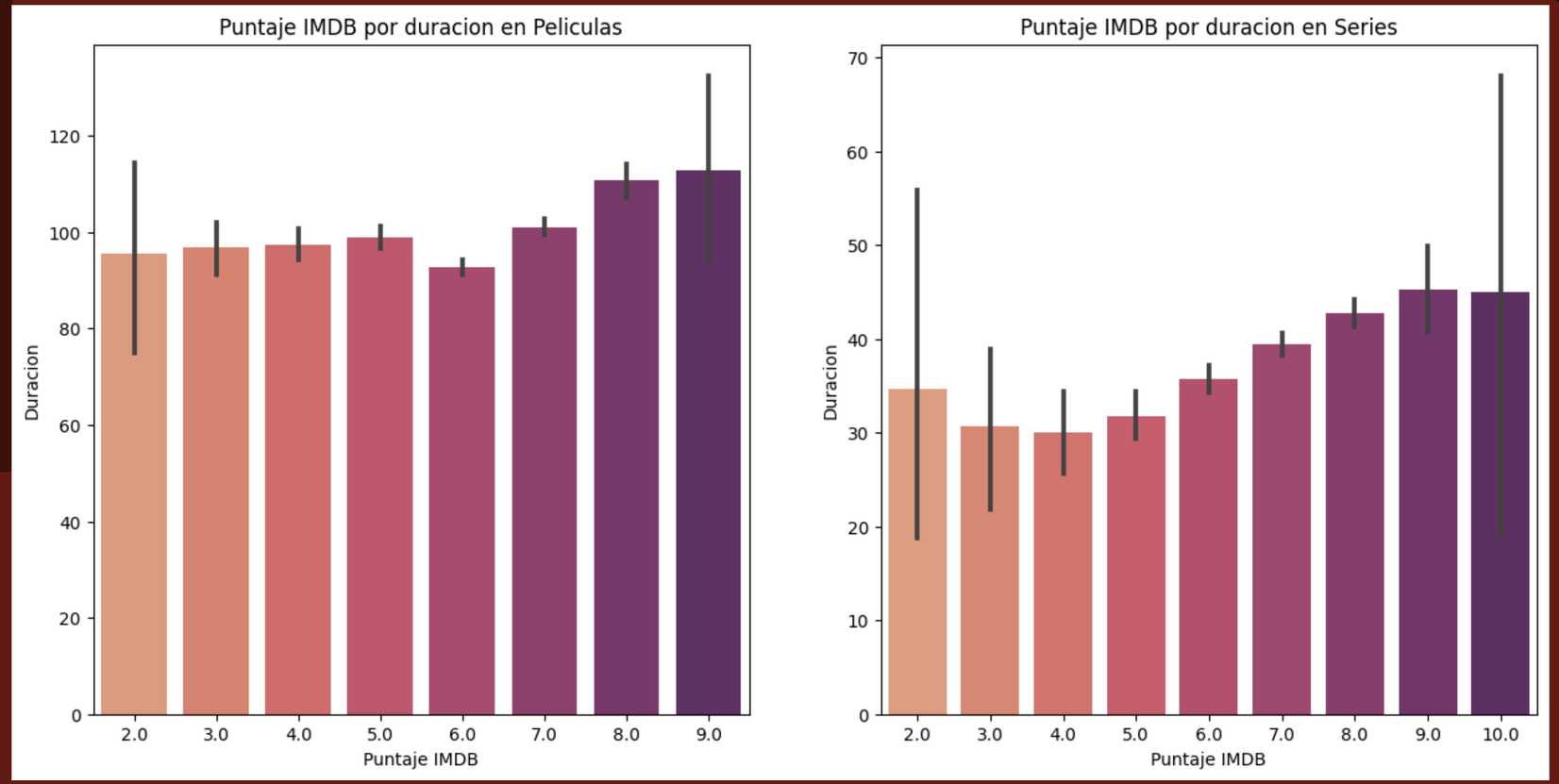


ANALISIS DE DATOS

07

Calidad segun duracion

Finalmente, podemos observar que la duración de la producción no tiene demasiada injerencia en la calidad de la misma, por lo que no es un parámetro ponderante para las predicciones.




RECOMENDADOR
CONTENIDO

IDEAS

Idea Principal

Cuando se inicie se solicitará que se elijan algunos títulos que le gusten al usuario. Usaremos esos títulos para el primer proceso de "configuración" de las recomendaciones. En base a esas elecciones se podrá agrupar que títulos podrían ser de agrado para el usuario, esta agrupación tendría en cuenta la información sobre los títulos como ser género, categorías, actores, año de lanzamiento y demás características.

Hibrido

Se podria decir que es un sistema hibrido ya que en principio es un recomendador basado en contenido pero una vez enfocado en el grupo de titulos se basa en popularidad para recomendar los primero 5 titulos dentro del grupo(cluster)

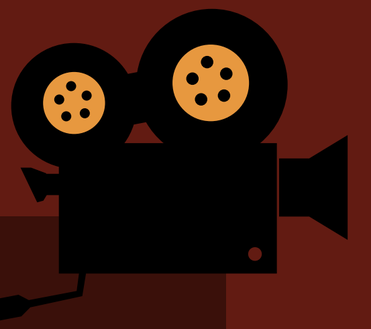
Algoritmo

El algoritmo usado para el agrupamiento es HDBSCAN el cual esta basado en DBSCAN convirtiéndolo en un algoritmo de agrupamiento jerárquico y luego usa una técnica para extraer un agrupamiento plano basado en la estabilidad de los agrupamientos.

Arranque en frio

Para crear el perfil de usuario se solicitará que se elija entre los géneros existentes de la BD, una vez elegido el género la interfaz mostrara los 10 títulos que tengan el mejor IMDB de ese género, el usuario elegirá 1 título por género, 6 géneros en total.

CONCLUSIONES DEL RECOMENDADOR



Este recomendador esta centrado principalmente en una recomendacion por contenido, hoy en dia las aplicaciones utilizan varios recomendadores al mismo tiempo, como ser los mas votados, los mas vistos, lo mas visto por perfiles de usuarios iguales a los tuyos, los de tendencias y hasta veces haciendo un mix de ellos.

- El dataset utilizado solo contiene informacion al respecto de los titulos con lo cual hace reducido su alcance.
- Son necesarios mas datos para tener una aproximación mejor.
- Seria ideal poder contar con datos sobre el comportamiento de los usuarios dentro de la aplicacion para poder ampliar la recomendacion

COOMING SOON: RECOMENDADOR 2.0

(TAMBIEN LLAMADO POSIBLES MEJORAS)

01 Teniendo los datos de varios perfiles de usuarios se puede generar una recomendacion colaborativa, es decir, generar recomendaciones de titulos basado en los titulos que le gusten a perfiles como el tuyo

02 Este arranque en frio se hace a travez de selecciones del usuario, seria optimo poder contar con la informacion del comportamiento real del usuario dentro de la aplicacion como ser, titulos que realmente vea o titulos que pause sin llegar a verlos completos asi como titulos que vea dos o mas veces.

03 Una vez este en uso el recomendador se podra utilizar el feature de "me gusta" o likes que de el usuario a los titulos para generar un indice de puntuacion propio del usuario.


PREDICTOR
PUNTAJE

IDEAS

Idea Principal

Mediante el uso de la BD con los títulos en la plataforma, se planea realizar una herramienta para poder predecir, basándonos en ciertos parámetros previamente vistos, el puntaje de una nueva producción que se agregue a la plataforma y así decidir si mantiene los estándares de calidad de la plataforma o no.

Herramientas

Para realizar el predictor, se utilizaran herramientas de machine learning para definir los puntajes, dentro de las mismas se encuentran distintos métodos de regresión y con los mismos se buscara obtener una predicción lo mas certera posible.

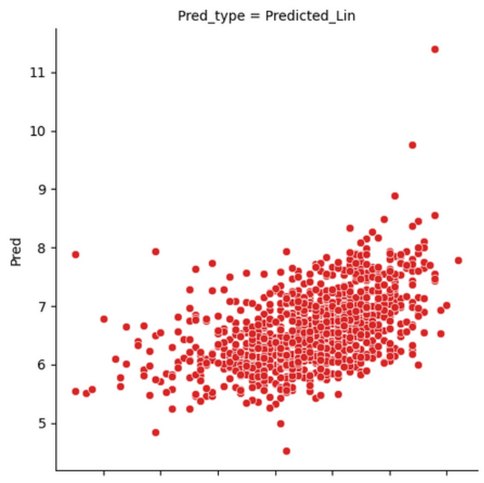
Parámetros

Debido a ciertos problemas que surgen al momento de realizar las predicciones con ciertos parámetros, se realizo una limpieza de los mismos tomando, por ejemplo, producciones que se encuentren dentro del 75% de los mas votados

Optimizacion

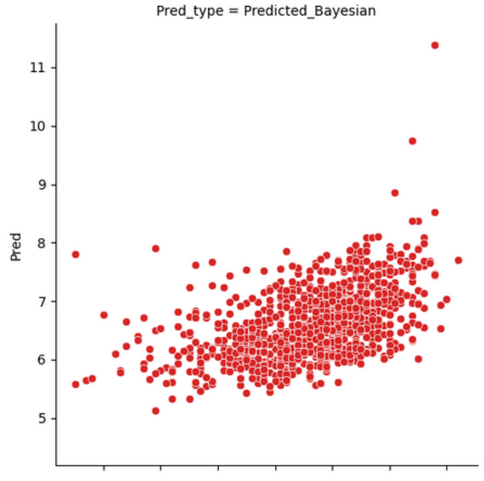
Dentro de los modelos utilizados, algunos fueron optimizados mediante una búsqueda de grilla de valores calculados para encontrar los mejores parámetros a utilizar en los modelos.

PREDICCIONES Y METODOS



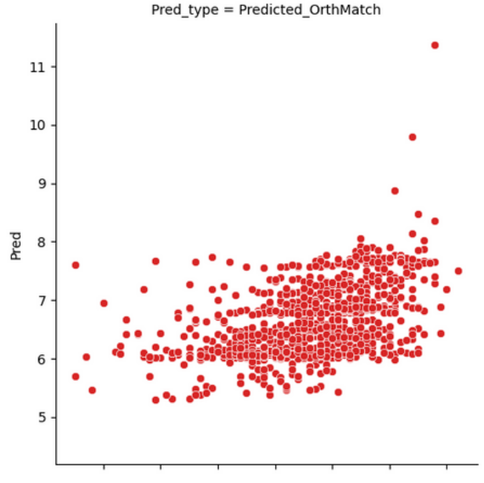
Regresión Lineal

Modelo de relación lineal clásico, como se ve, la aproximación de la predicción tiende a sobre calificar las producciones.



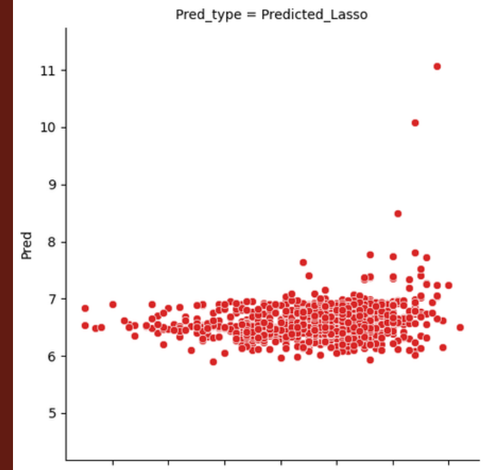
Regresión Bayesiana

Similar al anterior pero agregando un enfoque Bayesiano, vemos una mejora en la tendencia pero sigue existiendo una sobrecalificacion.



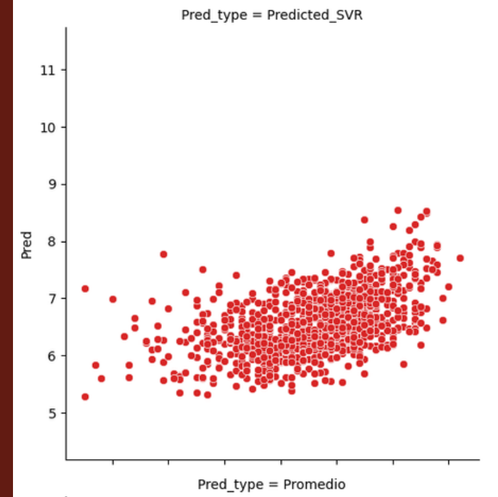
Regresión OMP

Es un algoritmo voraz iterativo que selecciona en cada paso la columna que está más correlacionada con los residuos actuales.



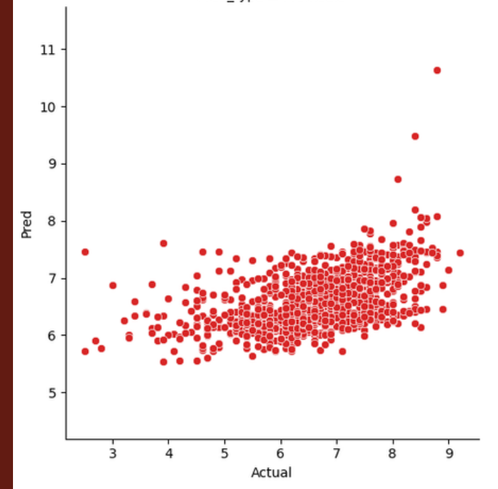
Regresión Lasso

Realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este. Claramente no es el mejor



Regresión Vectorial

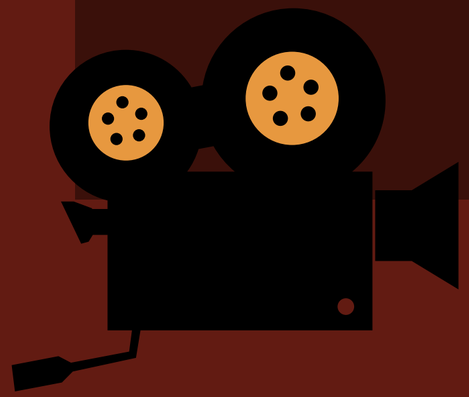
Regresión de Vectores de Soporte, es un sistema de regresión vectorizado, similar a una SVM.



Promedio de Regresiónes

Es el promedio realizado entre los 5 modelos de regresión, se ve que al unir todos los métodos puede acercarse a una versión mas exacta.

TODO MUY INTERESANTE, PERO QUE ACERTIVIDAD TIENE REALMENTE.



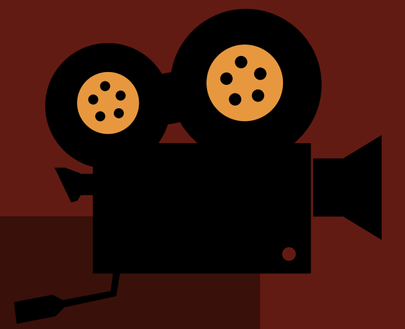
PUNTOS CLAVE A TENER EN CUENTA

01 La cantidad de parámetros a tener en cuenta para realizar una predicción exacta es demasiado extensa. Mucho mas que la del dataset utilizado

02 Si bien hay producciones que trascienden generaciones, dependen mucho de la época en la que fueron hechas para tener en cuenta si será o no un éxito

03 Mas allá de las distribuciones observadas, el promedio de desviación en los puntajes si tomamos en cuenta los 5 modelos es de un 0.4% y de un 4.8% de mediana

CONCLUSIONES DEL PREDICTOR



Si bien no es un método 100% confiable, las herramientas de Machine Learning utilizadas pueden dar una aproximación bastante fidedigna sobre la calidad que tendrá una producción añadida a la plataforma teniendo en cuenta los parámetros tomados en cuenta

- Las herramientas de ML por si solas resultan ineficientes, se deben usar varias para obtener buenos resultados.
- Son necesarios mas datos para tener una aproximación mejor.
- Los métodos de calificación propios pueden no ser una buena guia.


MUCHAS GRACIAS