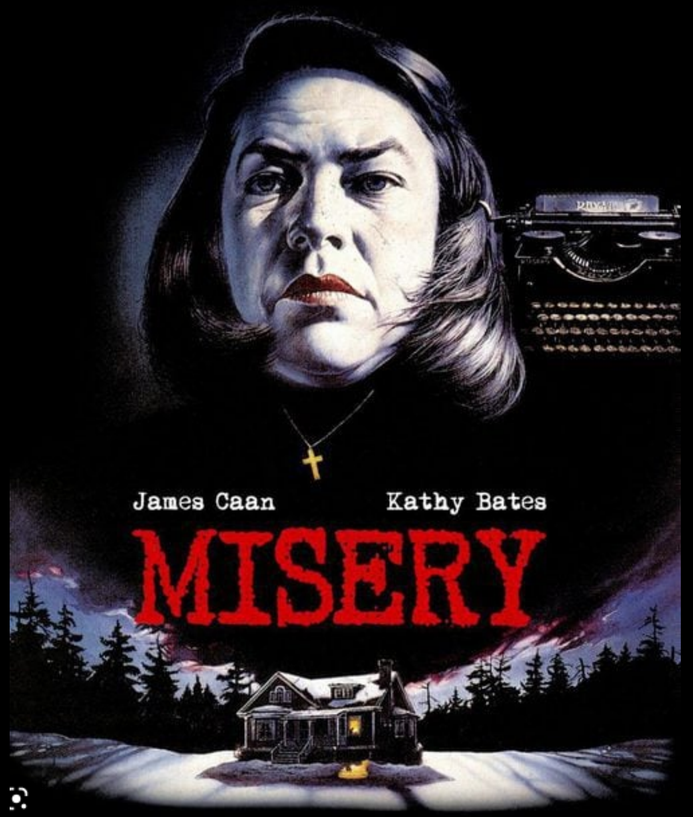
Braindead

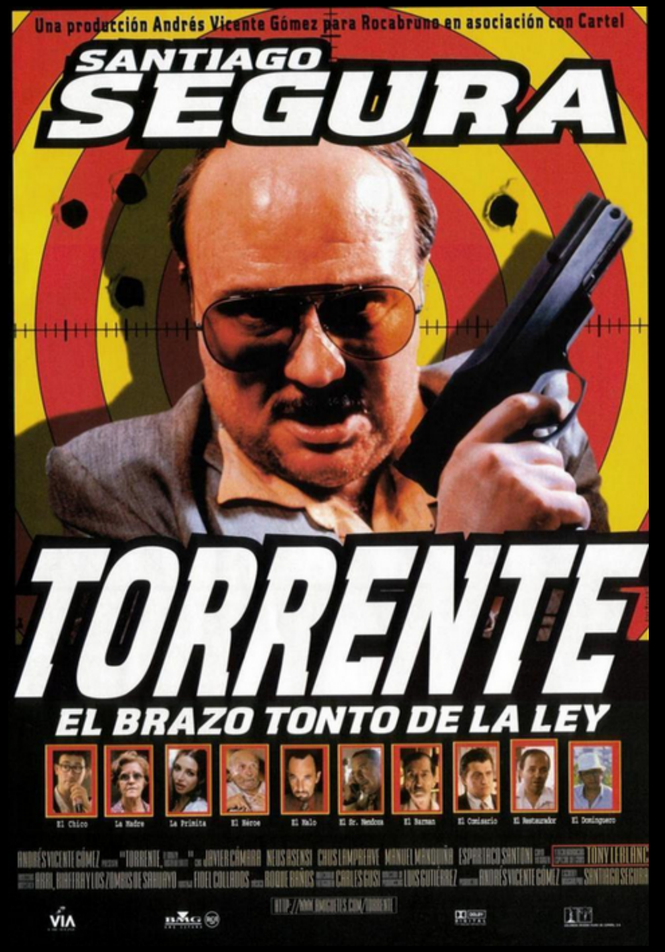
Proyecto de ML para The Bridge

Aída Márquez

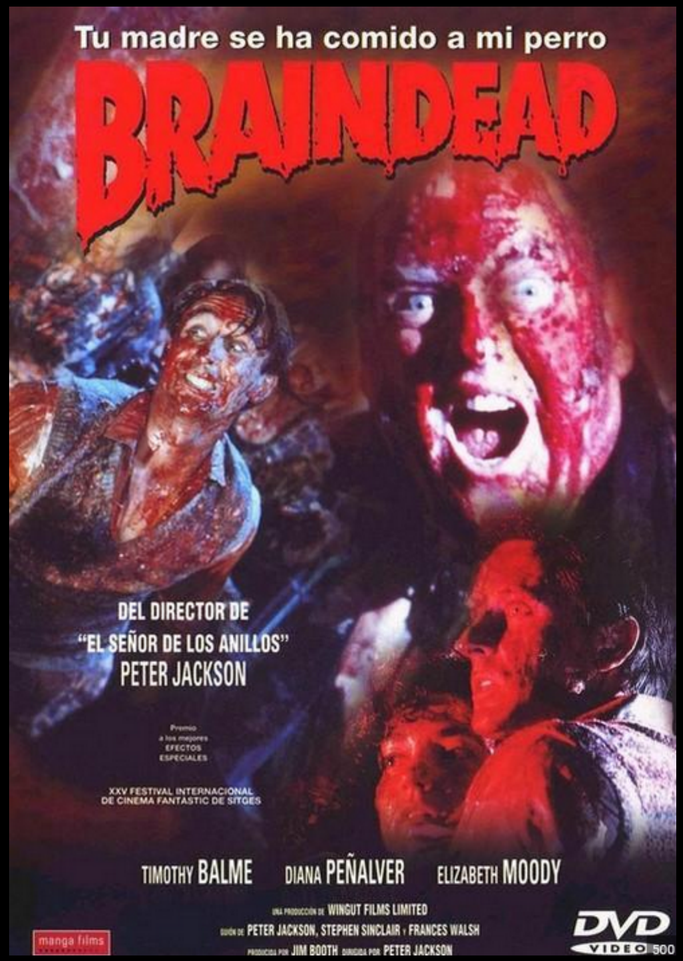
Las películas de miedo me dan miedo

Y las comedias, a veces, también





Me pregunto si se seguirán haciendo comedias de terror



Cambiar de dirección en mitad de un salto es mala idea casi siempre.

Datos para hacer el estudio

Extraídos de Imdb con la API de Tmdb (algunos de Wikipedia fueron descartados)

Películas: 10.000

Características:

Descartadas: "overview", "backdrop_path", "spoken_languagess"

Numéricas: "popularity","vote_count","vote_average","budget","duration"," revenues"

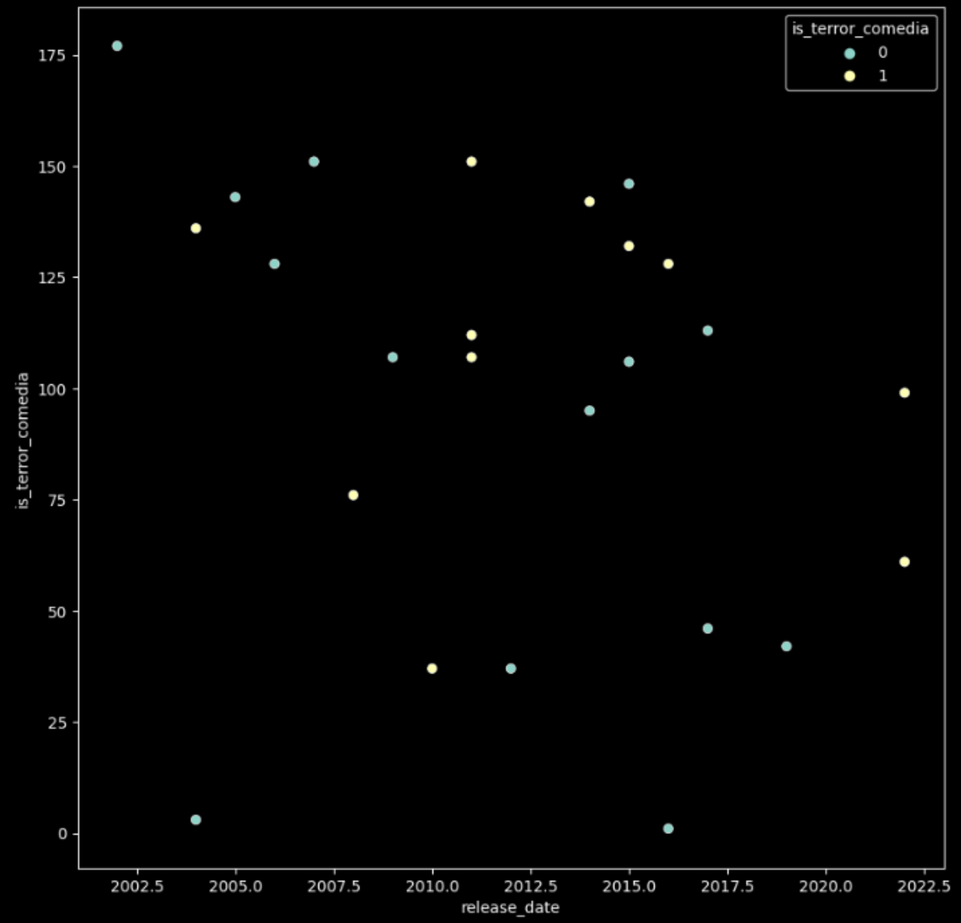
Categóricas: genre_ids

Añadidas: quantile, reciente, is_terror_comedia, val_count

Los datos agregados tienen una distribución bastante irregular

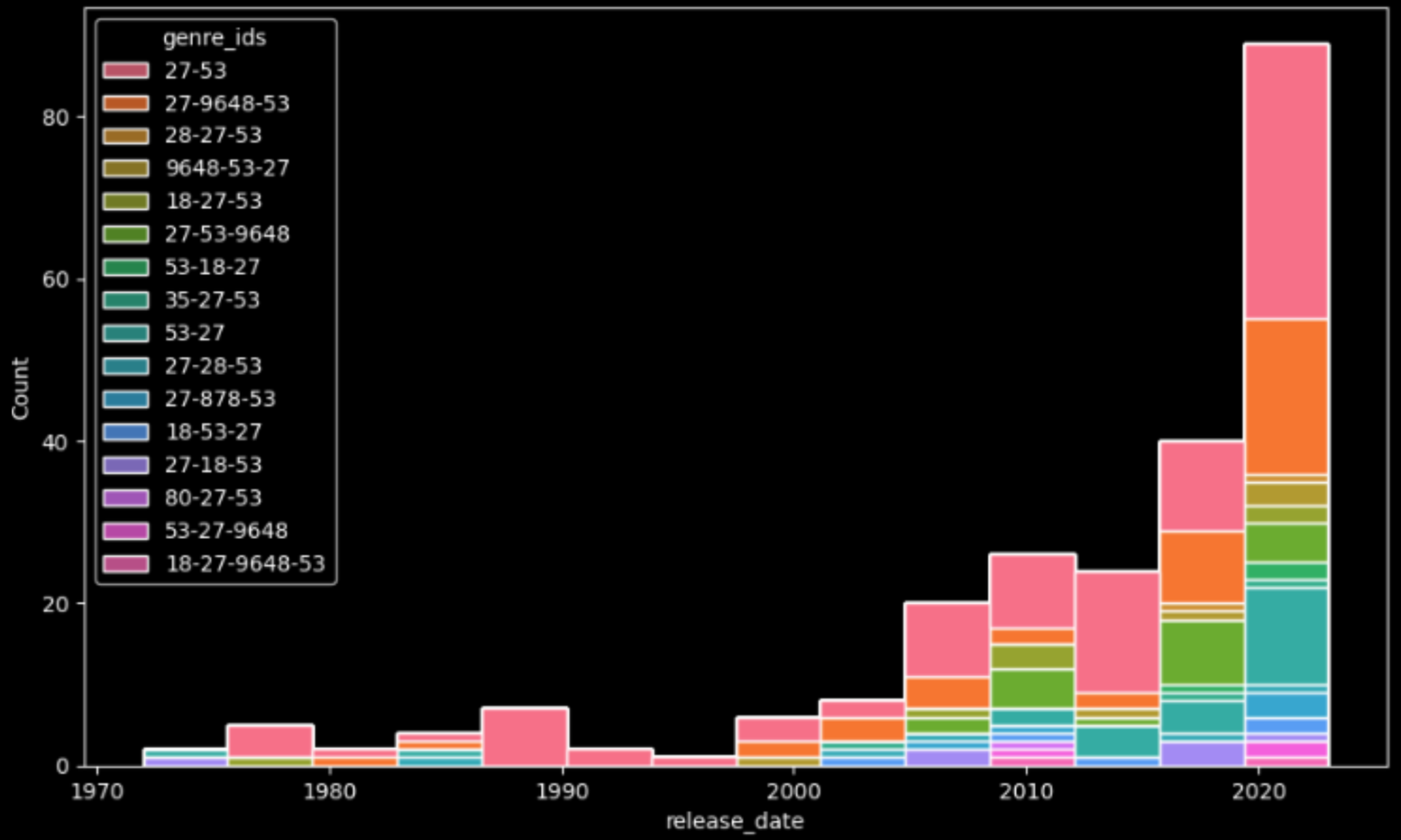
Estrenos por año desde el año 2000

En amarillo las películas de terror
que además son comedias,
en azul las que no,
son de terror y otros géneros



Existen muchas combinaciones

Lo más práctico en este caso será separar en dos categorías



Estrategia inicial

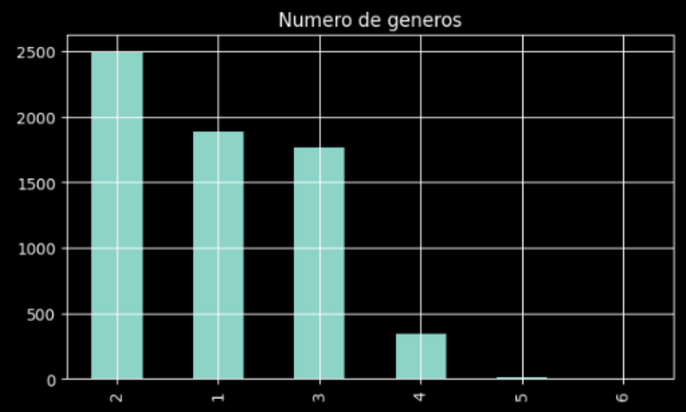
Cómo modelar los datos, qué categorías me interesan

Divido el dataset entre películas de comedia-terror y las demás. La columna is_terror_comedia contiene el dato binario.

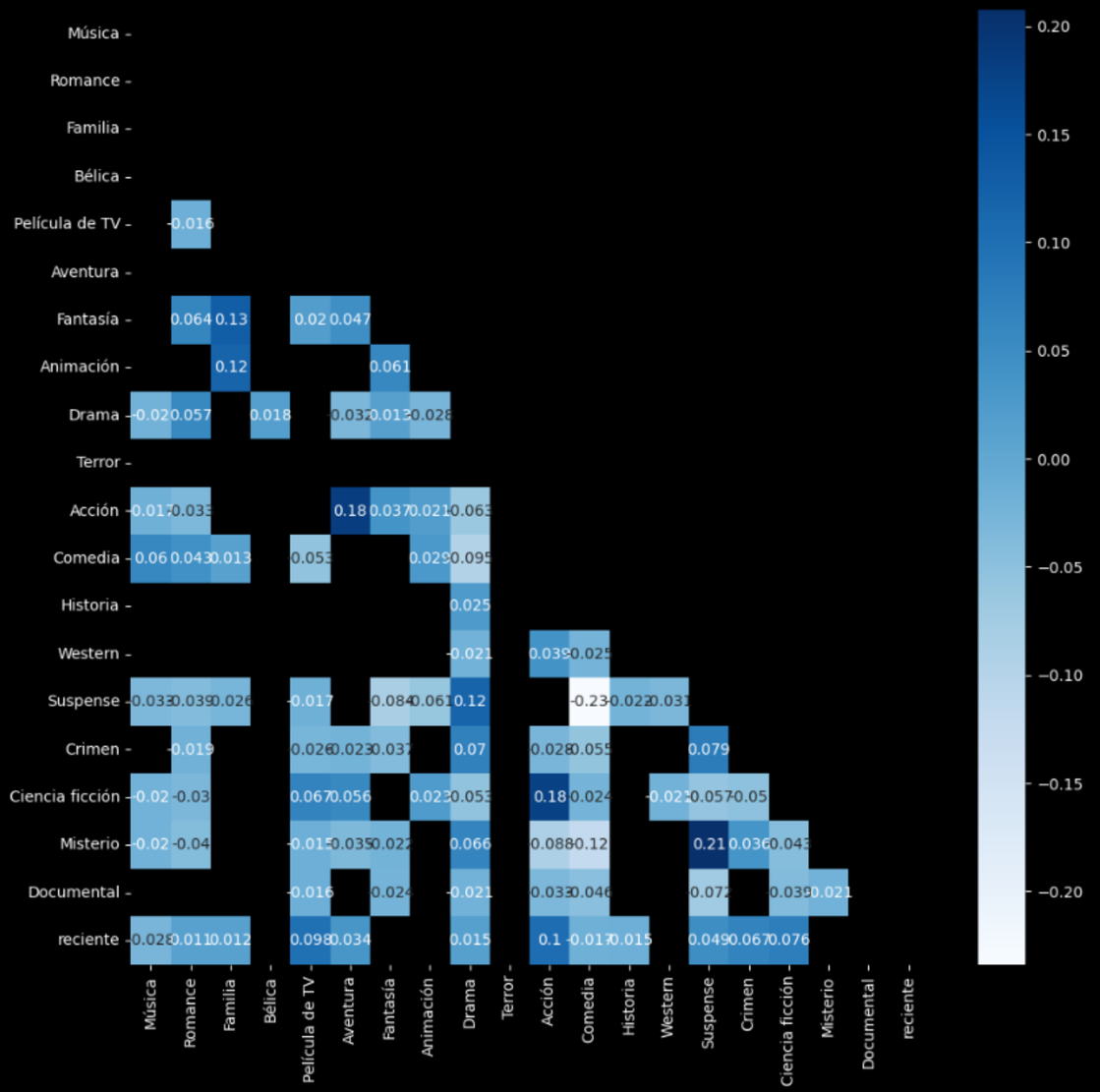
Las columnas numéricas serán las que utilice como X

Trabajando con los datos
Feature engineering

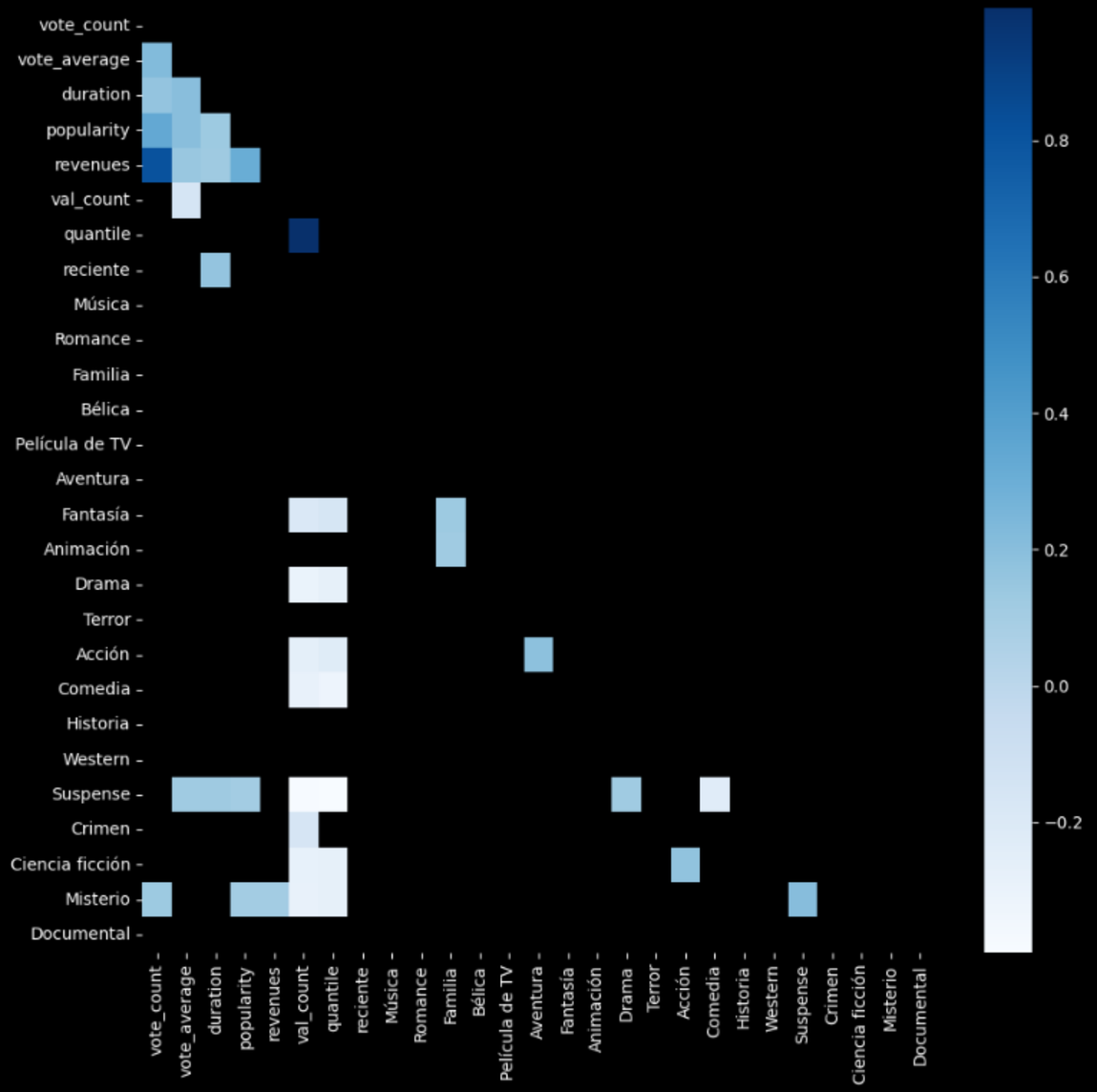
- Hay muy pocas películas con popularidad mayor a 500
- Y muy pocas pertenecientes a idiomas minoritarios, una o menos
- Las que pertenecen a más de 4 géneros también con poco representativas
- Existe un problema para analizar los datos, la recaudación no es significativa porque es un género que se distribuyó mucho en vídeo

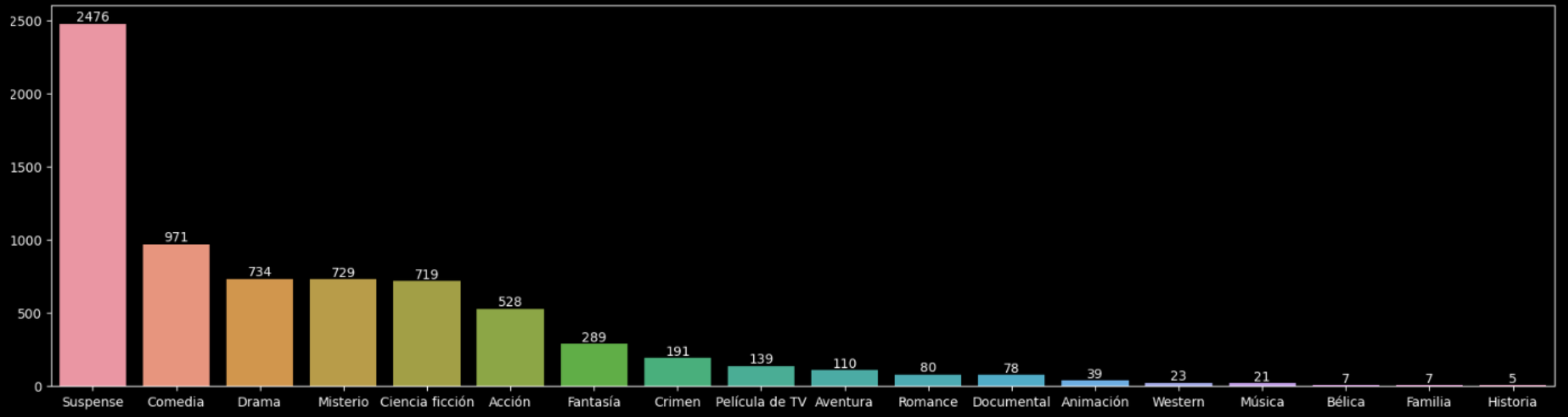


- Por curiosidad miramos qué géneros están más correlacionados.



- Y filtramos el heatmap para ver un poco mejor.




2500
2000
1500
1000
500
0
2476
971
734
729
719
528
289
191
139
110
80
78
39
23
21
7
7
5
Suspense
Comedia
Drama
Misterio
Ciencia ficción
Acción
Fantasía
Crimen
Película de TV
Aventura
Romance
Documental
Animación
Western
Música
Bélica
Familia
Historia

Primera aproximación a algoritmos de clasificación

Feature importances

Probamos varios algoritmos, otros los descartamos por diferentes razones. Pero el objetivo siempre es predecir la categoría (¿pertenece a terror/comedia?)

- Regresión Logística
- Árboles de decisión
- Catboost
- Kmodes
- MultioutputRegressor

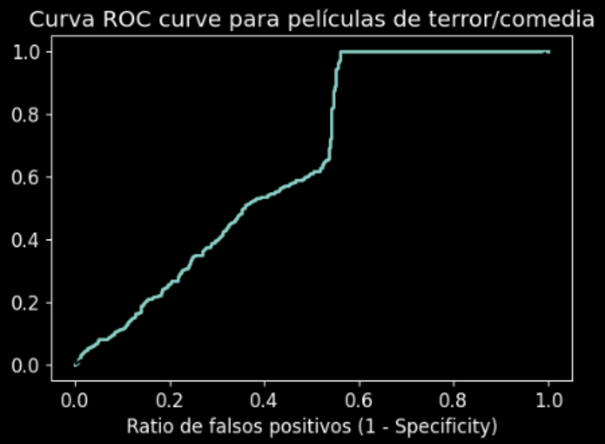
Score de los diferentes modelos

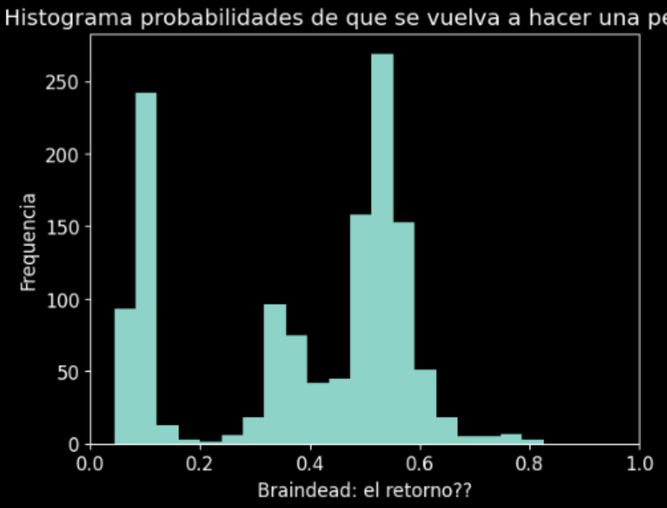
- Regresión Logística: 0.6025 (C=1, solver='saga')
- Árboles de decisión: 0.905805 (n_estimators=100, eval_metric = 'logloss')
- Catboost: 0.91(iterations=100, learning_rate=0.05, depth=6, eval_metric='Accuracy')
- Kmodes: Descartado
- MultioutputRegressor: R2 score: 0.95

Observaciones tras probar modelos

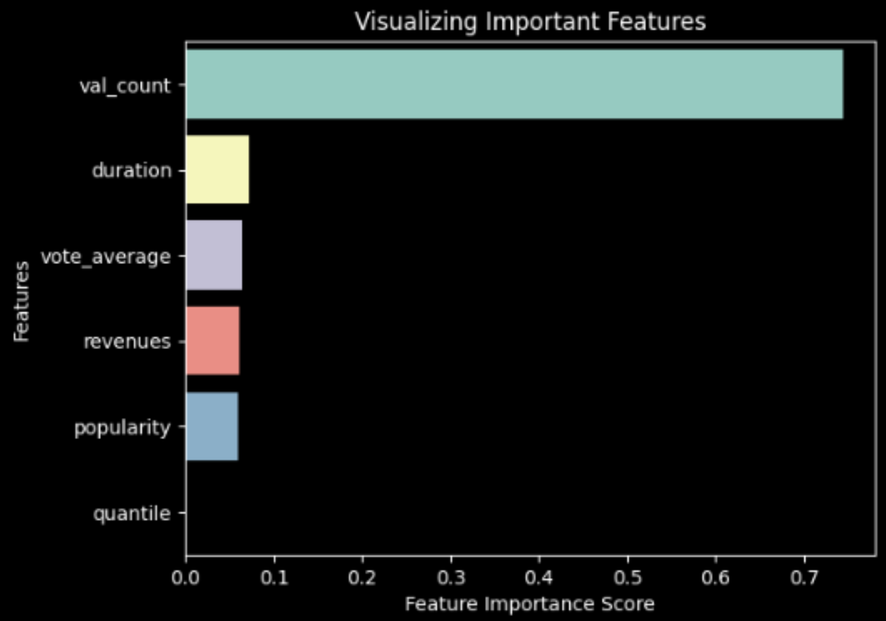
La regresión logística es pesimista

Y no se ajusta demasiado bien

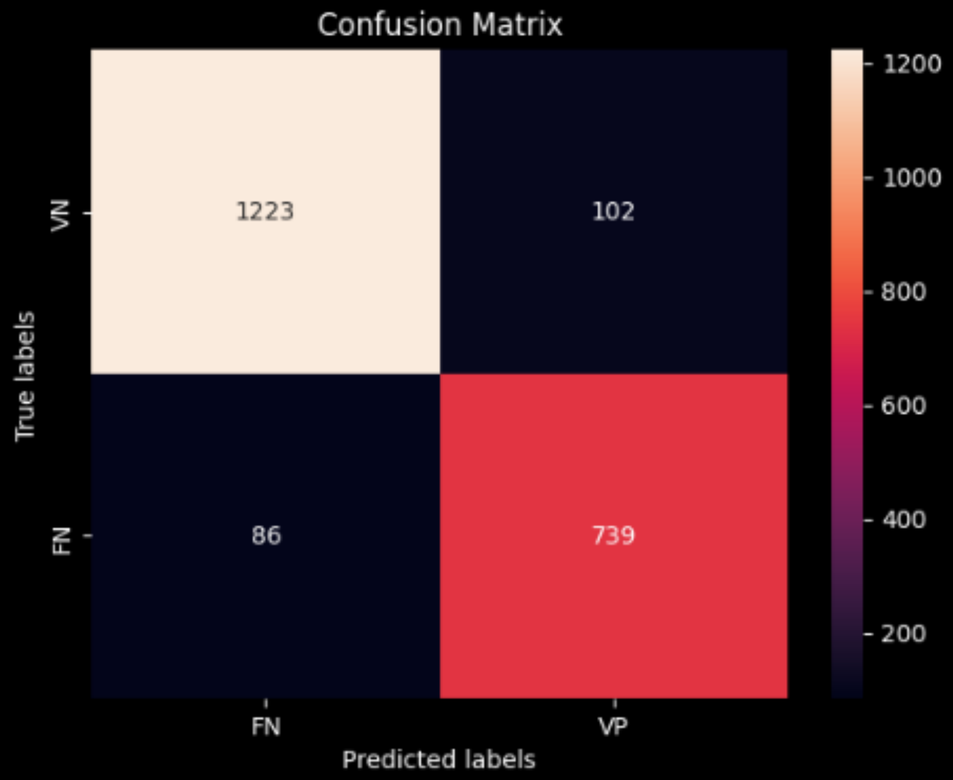


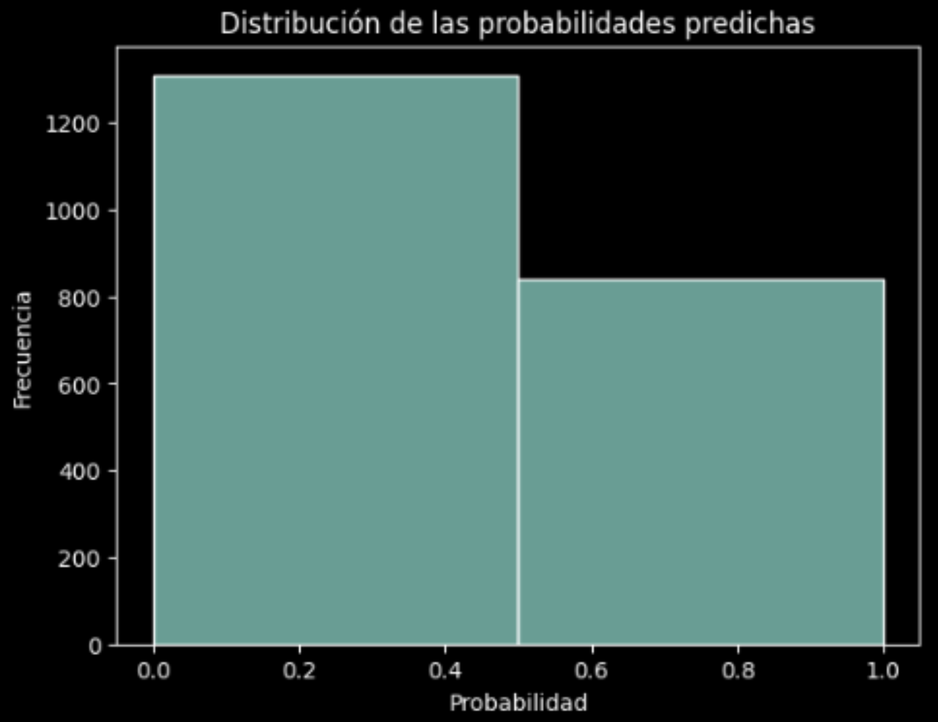


- Los árboles se adaptan mejor al problema
- XGBoost nos da buenos resultados
- Y algunas características ordenadas por importancia
- Val_count representa el número de apariciones de esa combinación de géneros en el dataset. Parece que sí nos da información.



- Parece que generaliza bastante bien.




Distribución de las probabilidades predichas
Frecuencia
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad

- Y Catboost sería la otra opción interesante. Nos devuelve estos indicadores

Precision class=0: 0.9076

Recall class=0: 0.9241

F1 class=0: 0.9158

Opciones descartadas que podrían haber sido interesantes

Series temporales, redes neuronales

Tras un par de pruebas con Fbprophet ... descartado

Estrategia elegida

Riesgo de overfitting. ¿Regularización?

No he visto riesgo de overfitting, aunque habría que estudiar más profundamente las features elegidas.

Quizás habría que usar más categorías, tener en cuenta la industria del vídeo, coetánea de épocas en las que el género fue muy popular.

Conclusiones

Con los mismos parámetros que tenemos en películas de Terror/Comedia de los 80, la probabilidad de que los sea ahora es de 0.002

Estos parámetros podrían no ser válidos, influyen temas como por ejemplo el vídeo, las plataformas, la manera de votar ...

Los datos

Ay! los datos

Quizás hay otros enfoques que pudieran enriquecer el resultado del proyecto.

Merecería la pena explorarlos en el futuro.

Gracias

!