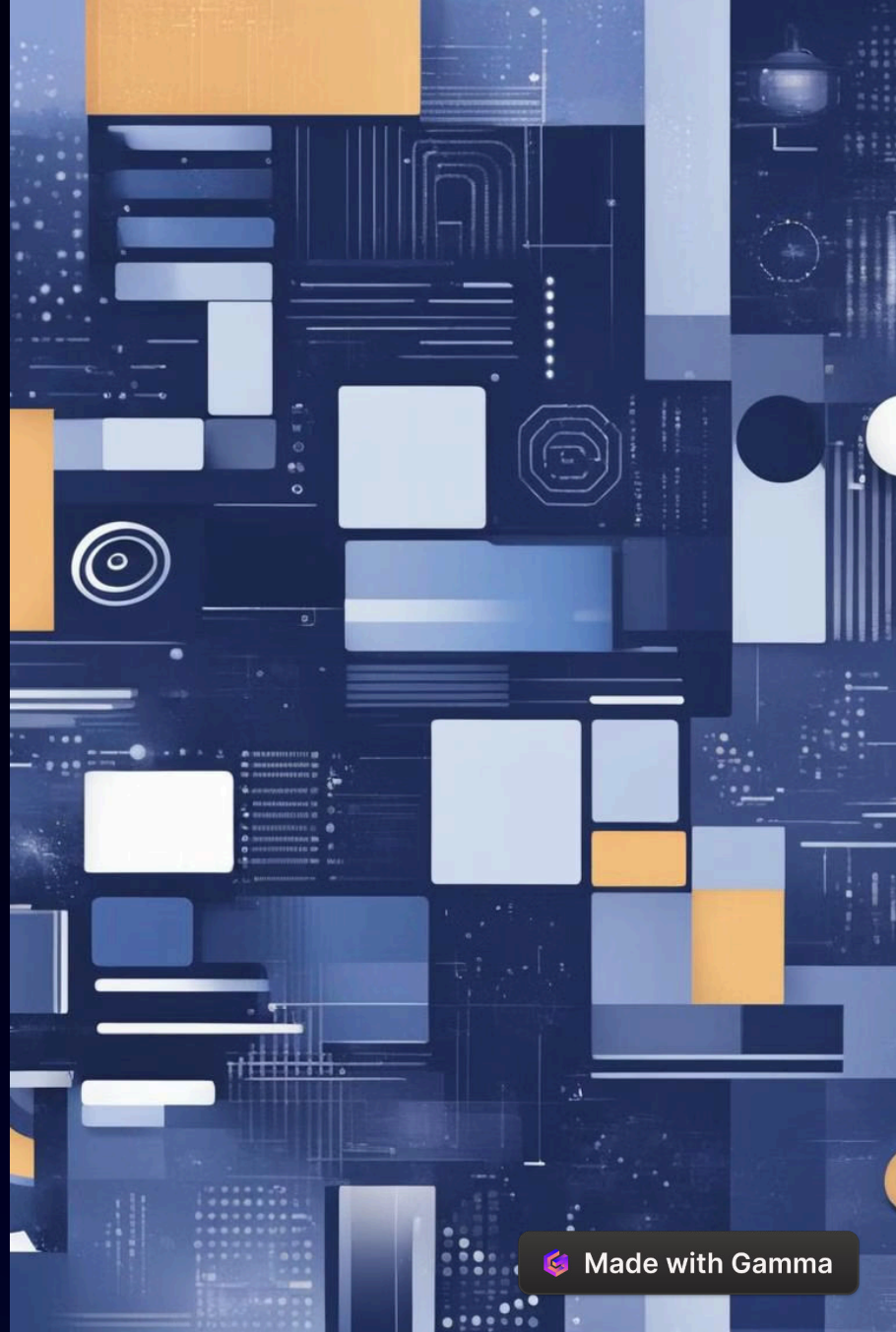
Introducción a la predicción de géneros de película

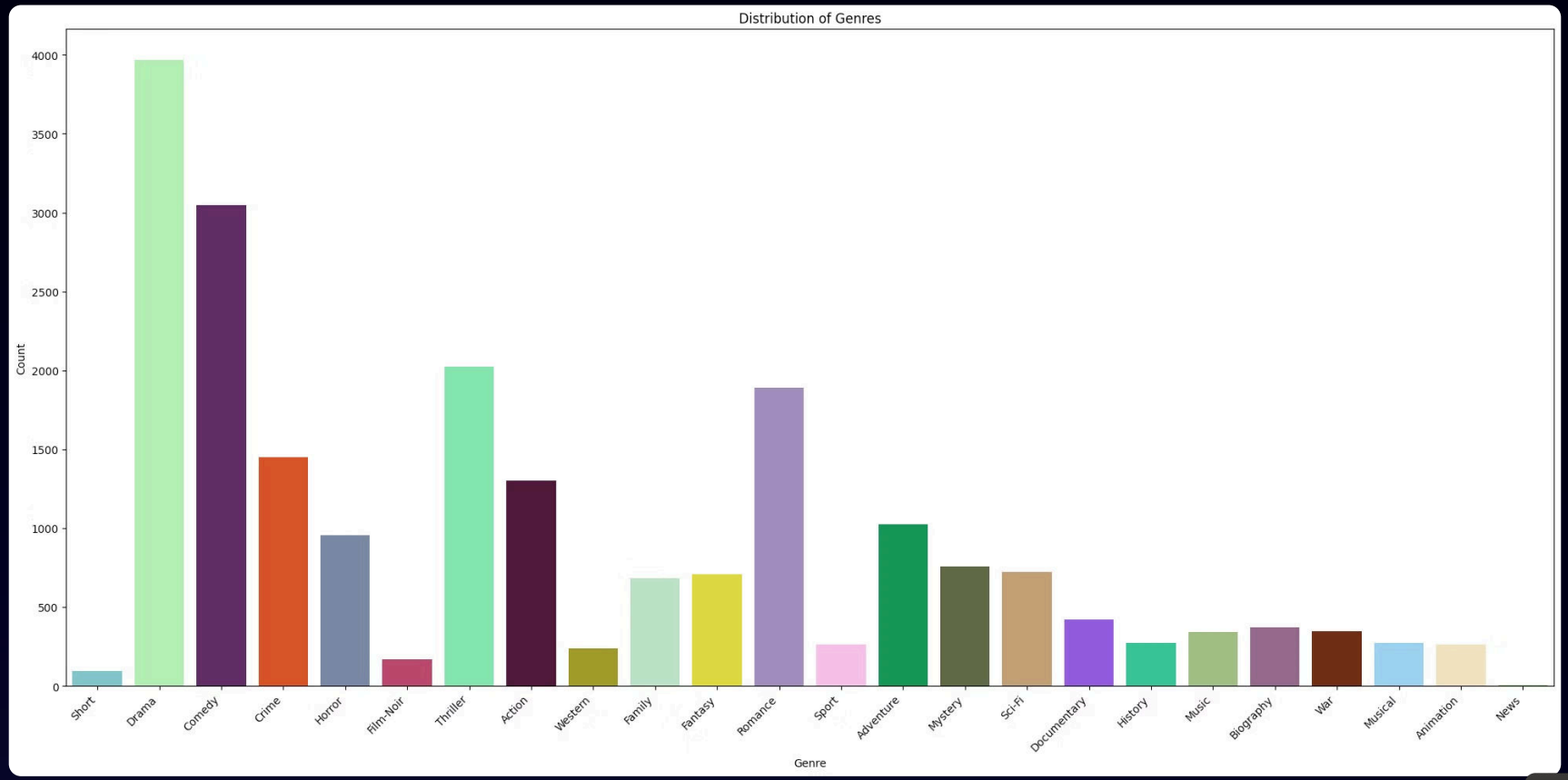
En esta presentación, la predicción de géneros de película utilizando modelos de aprendizaje automático. Analizaremos cómo los algoritmos pueden identificar patrones en los datos para predecir con precisión los géneros de las películas.



Distribución de Datos de Entrenamiento

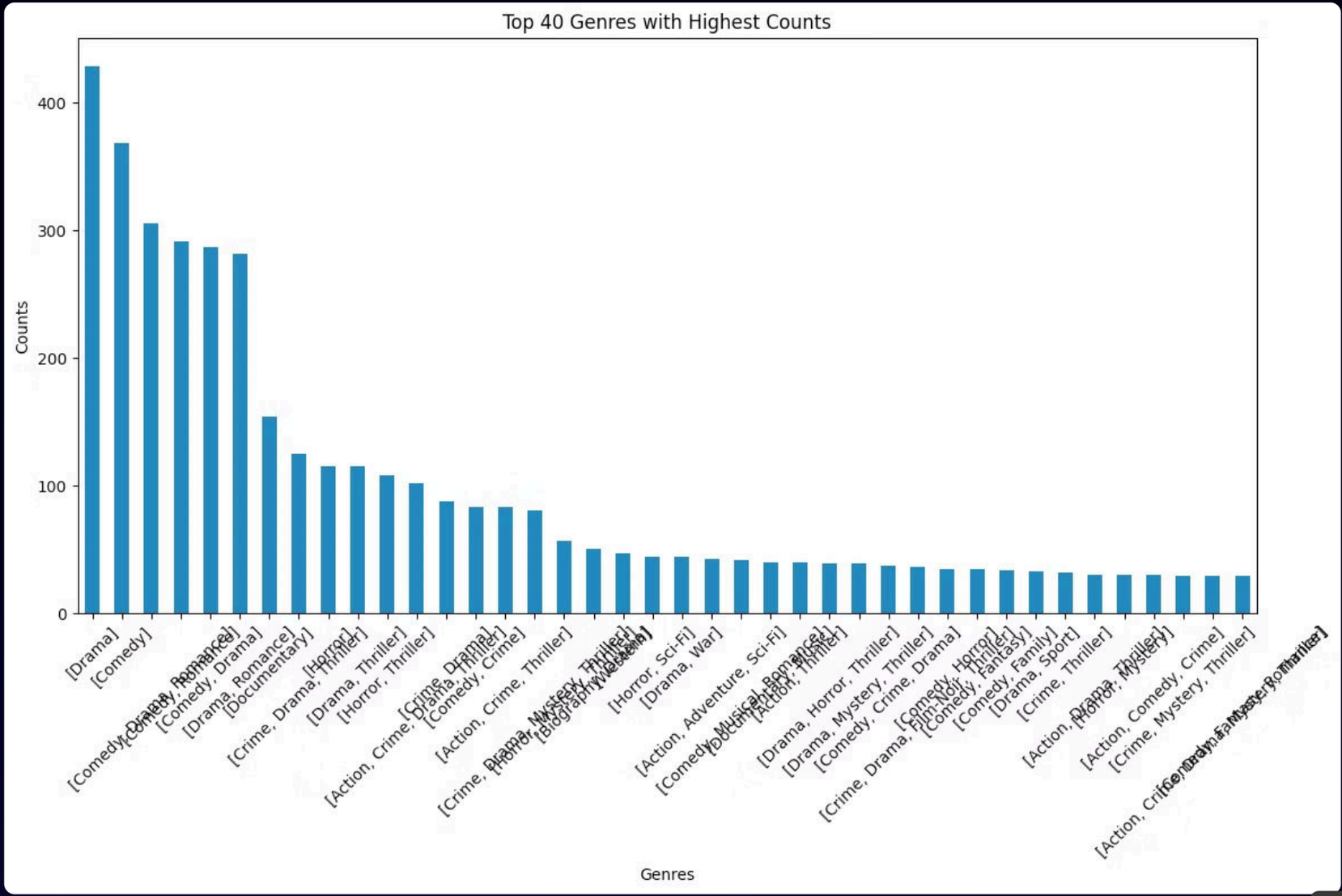
7,895 entradas de películas

24 diferentes tipos de genero de película con esta distribucion por genereo que implica unos targets desbalenciada



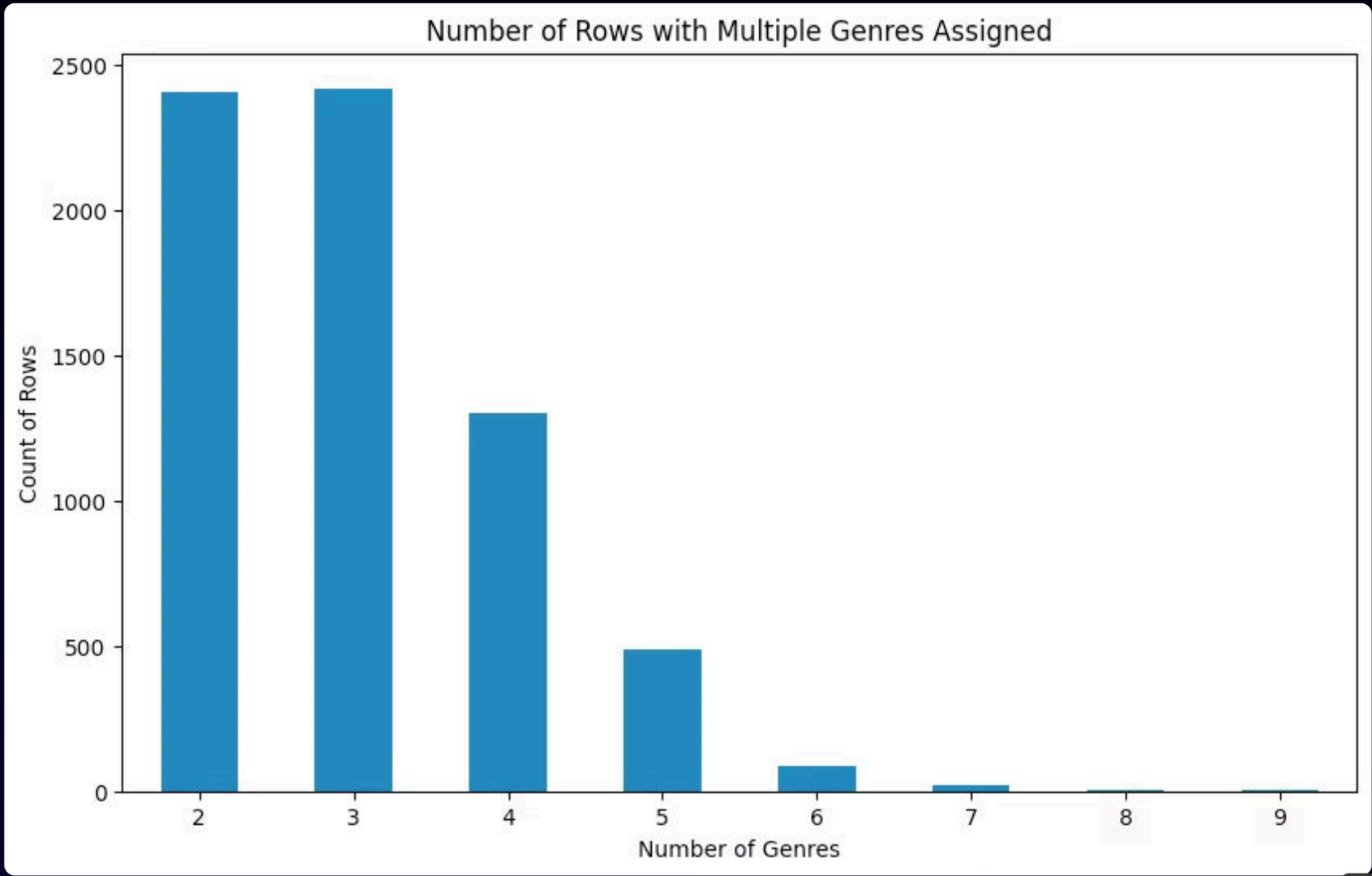
Distribución de Datos de Entrenamiento

Cada entrada o película puede estar asignado mas de un tipo de genero.



Distribución de Datos de Entrenamiento

El top es 9 géneros por entrada de película que solo cuenta con 2 entradas



Modelos de Random Forest como Base Case

depth_level = 10

CountVectorizer con features = 1000

ROC = 0.7812262183677007

ROC no fue muy afectedo por depth_level

Aplicamos TF-IDF con paramentos

stop_words='english',ngram_range=(1, 2), min_df=2, max_df=0.95

Resulto con un aumento del ROC

0.8107723036823163

Modelos de regresión logística

Aplicamos TF-IDF con los mismo parametros que se usaron en Random Forest.

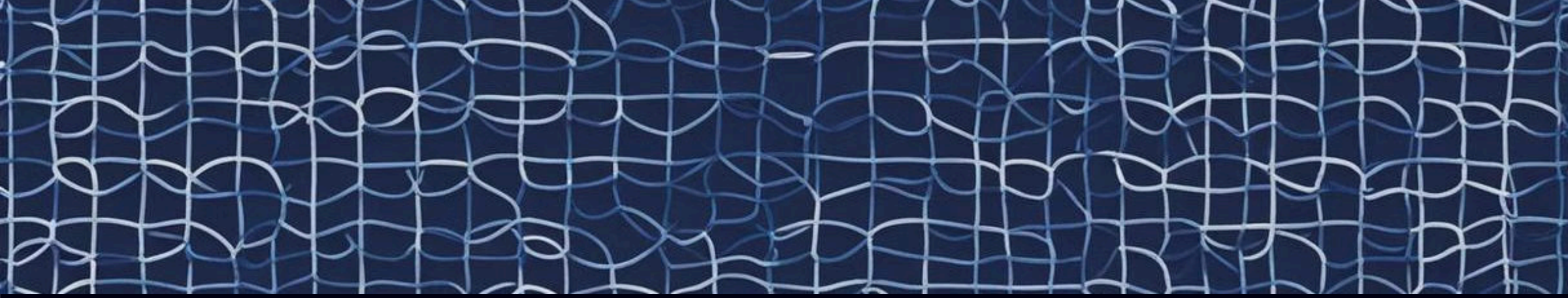
ROC = 0.875801684876896

Usando GloVe Embeddings o Count Vectorizer o otros parametro de TF-IDF no resulta en mejor rendimiento del ROC. El peor de estos fue aplicando el GloVe embedding

Modelos de SVM/ Naive Bayes

Entrenamos un modelo SVM y MultiNomial Naive Bayes con Count Vectorizer y TF-IDF pero el rendimiento no supero los modelos anteriores

ROC = 0.5270409587370896



Modelos de GRU (Gated Recurrent Unit)

Se uso los embeddings de GloVe y se hizo una buscada de arquitectura de la red neuronal recurente. La buscada se realizo usando Keras Tuner y resulto con una arquitectura :

> GRU(units=64, dropout=0.4, recurrent_dropout=0.1)

Durante el entrenamiento , se demostró que el modelo no llega a converger en 100 epocas, en donde solo llega a una metrica:

```
ROC = 0.8461974367709132
```

Comparación de los modelos utilizados

Modelo	Sensibilidad al tamaño de datos	ROC
Regresión Logística	Medio	Alta
SVM	Alta	Bajo
Random Forest	Medio	Medio
GRU	Medio	Medio
Naive Bayes	Bajo	Bajo

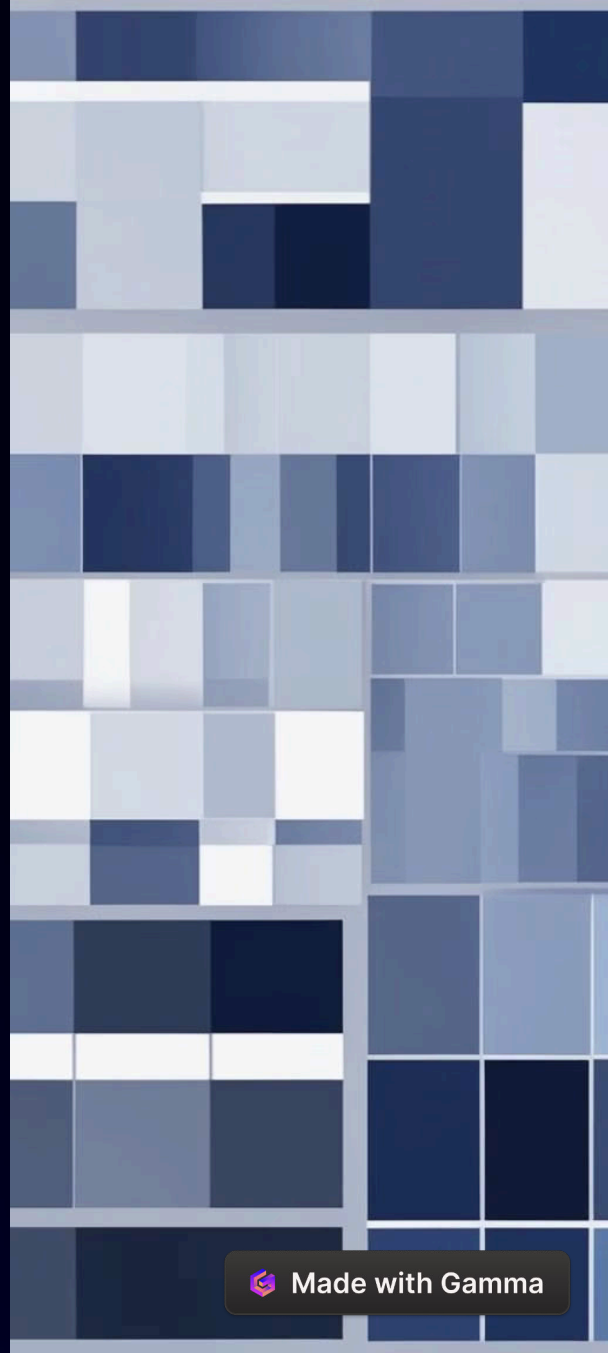
Resultados y conclusiones

Mejor Modelo

La regresion logistica resulto con el mejor ROC con 87.5%.

Lecciones Aprendidas

El tamaño de los datos se pueden incrementar pero tener en cuenta que unos modelos se saturan mas rapidos que otros



Preguntas y respuestas

Preguntas Abiertas

Es el momento perfecto para plantear preguntas y fomentar la discusión interactiva.