



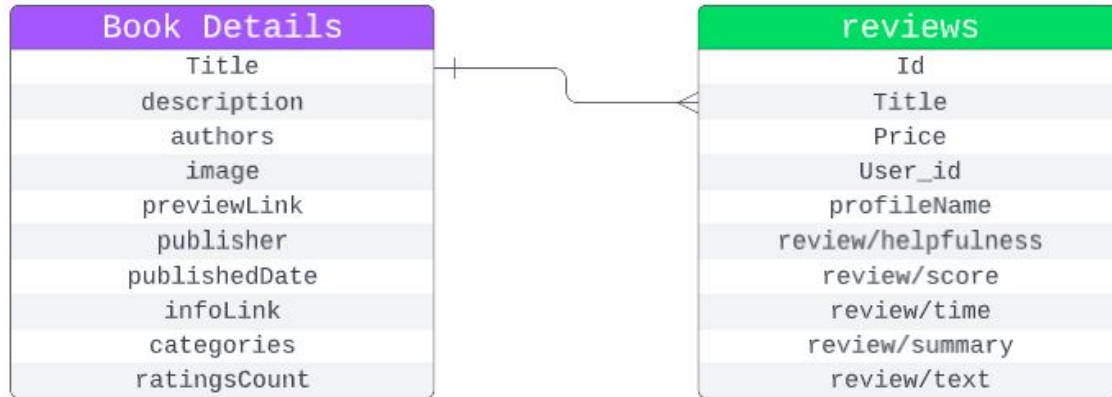
Sistema de recomendación de libros

Luis Manuel Ambrocio Loreto




Dataset

- El conjunto de datos contiene reseñas de libros de amazon, este conjunto de datos contiene 2 archivos, como se ve en la siguiente figura
- Se tienen 3 millones de reviews y 212404 libros



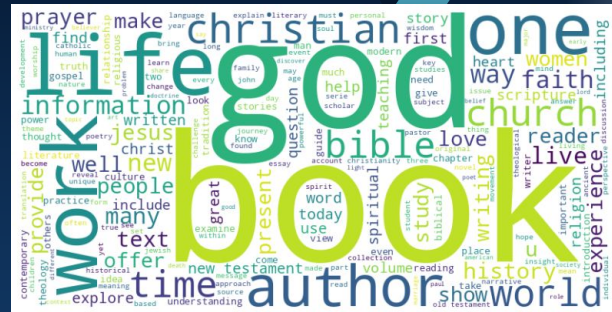
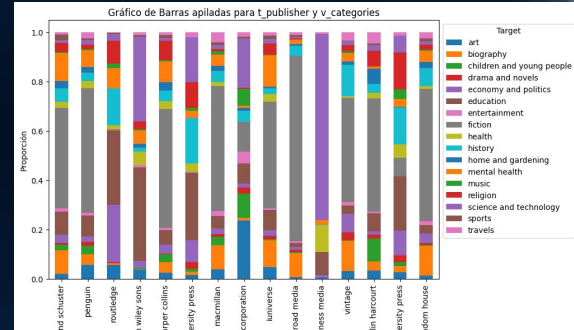
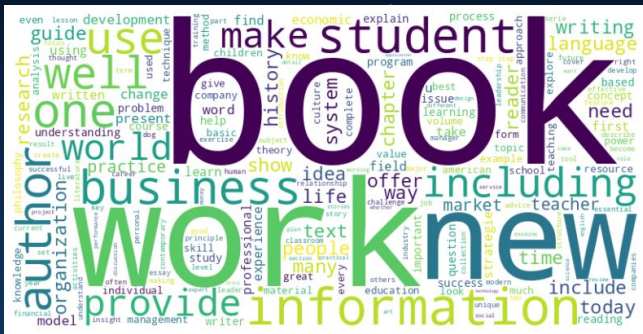
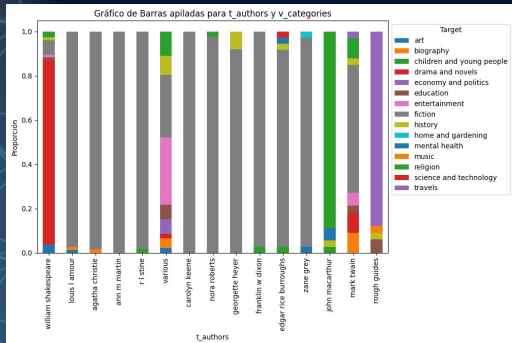


Objetivo

- El objetivo al final del diplomado es tener un sistema de recomendación que ayude a las personas a descubrir nuevos libros basado en sus gustos, entre la información más importante que se requiere para entrenar un sistema de recomendación se encuentra:
 - Tener información sobre calificaciones que usuarios hacen a los libros
 - Tener formas de agrupar a los libros y/o usuarios
 - Los objetivos de este parte del proyecto son:
 - Tener una representación numérica de la reseña que el usuario dio a un libro
 - Tener toda la columna de la categoría del libro llena (actualmente tiene muchos datos sin información)
 - Se crearán 2 modelos que nos ayuden a cumplir estos objetivos
- 

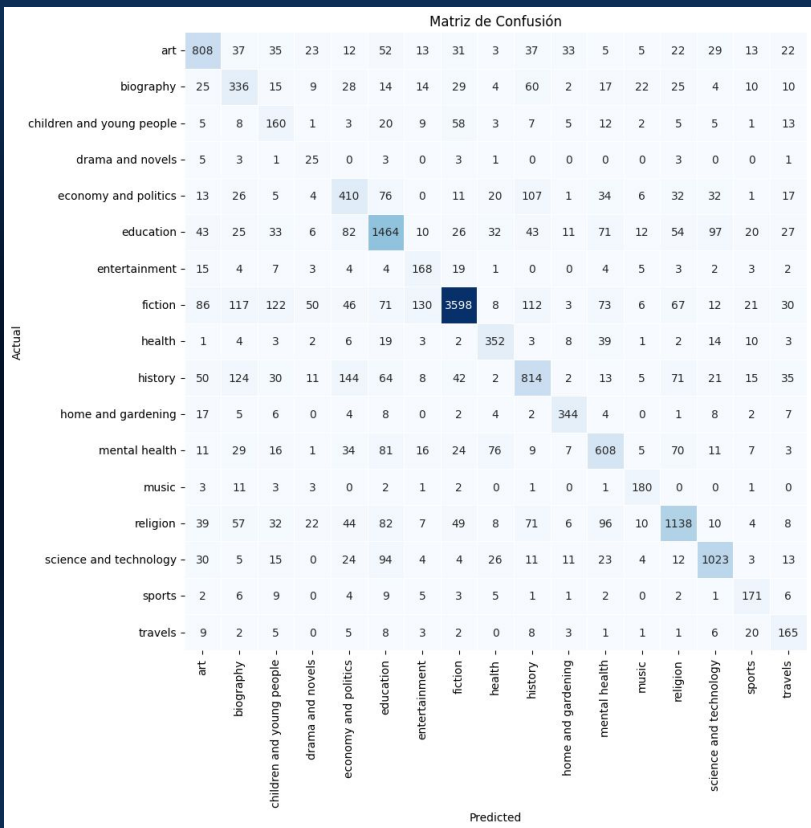
Modelo 1

- Se detectó que la variable de clasificación del libro (ficción, educación etc.) estaba muy sucia, esta variable es de mucha importancia al momento de crear recomendaciones.
- Una vez limpios los datos se detectó cerca del 12% de los datos sin información para esta variable, así que se decidió crear un modelo ML para imputar estos datos faltantes
- Al hacer el EDA las variables que más influyen en la clasificación son:
 - Autor
 - Editorial
 - Descripción y título del libro
 - Fecha de publicación



Modelo 1

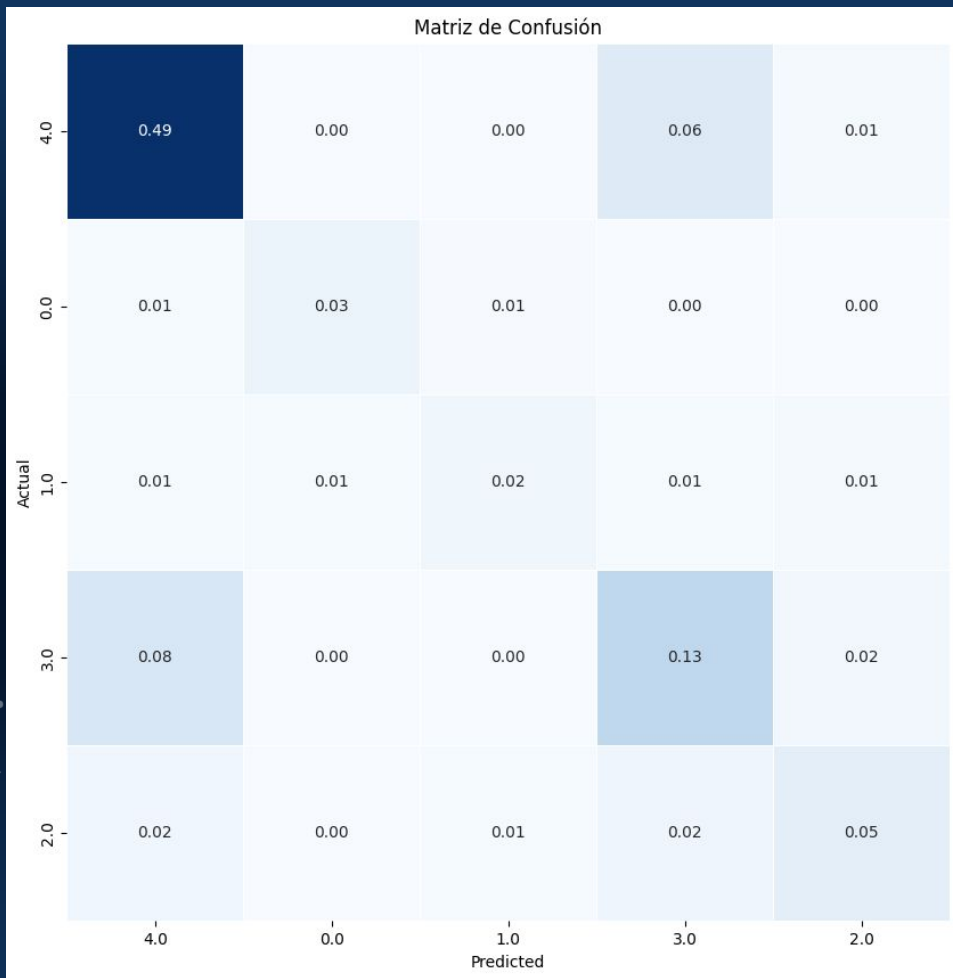
- Entre los modelos utilizados se encuentran:
 - Naive Bayes
 - Random Forest
 - Gradient Boosting
 - Ada Boost
 - Redes Neuronales Artificiales
- El mejor modelo fue una Red Neuronal donde se obtuvo un Accuracy de 0.71 y la arquitectura fue la siguiente:
 - La primera capa con 3 entradas diferentes:
 - Descripción tokenizada seguida de una capa embedding de dimensión 300
 - Título Tokenizado seguido de una capa de embedding dimension 300
 - Autor , editorial y fecha en One hot encoding
 - Se concatenaron las 3 entradas y se pasaron a una capa oculta con 64 unidades y funcion de activacion ReLu
 - Capa Dropout con parámetro 0.5 para regularización
 - Capa de salida con 17 unidades y función de activación softmax
 - Optimizador Adama con Learning rate de 0.00003





Modelo 2

- En el modelo 2 se busca predecir el Score que un usuario dio al libro a partir de su reseña.
- El objetivo es tomar la salida de la última capa oculta de una red neuronal como una representación numérica de la reseña.
- Se obtuvo un Accuracy con la siguiente arquitectura e hiperparametros
 - Review tokenizado como entrada:
 - Vocab_size: 200000
 - Max_sequence_length: 300
 - Capa embedding de dimension 300
 - Capa oculta de 64 unidades y función de activación ReLu
 - Dropout con parametro 0.4
 - Capa de salida con 5 unidades y función de activación softmax
 - Optimizador Adam con learning rate de 0.00002



Modelo 2

- Se combinaron los score 1 y 2 como score 1 (no le gusto), el 3 como score 2 (más o menos), y el 4 y 5 como score 3 (le gusto).
- Se mantuvo la misma arquitectura, solo cambiando la última capa de 5 unidades a 3 unidades, se tuvo un accuracy de 0.87

