UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PIA
APRENDIZAJE AUTOMÁTICO

Profesor

JOSÉ ANASTACIO HERNÁNDEZ SALDAÑA

Alumna

JUDITH CAROLINA LUCIO GARZA

Matrícula: 1877415

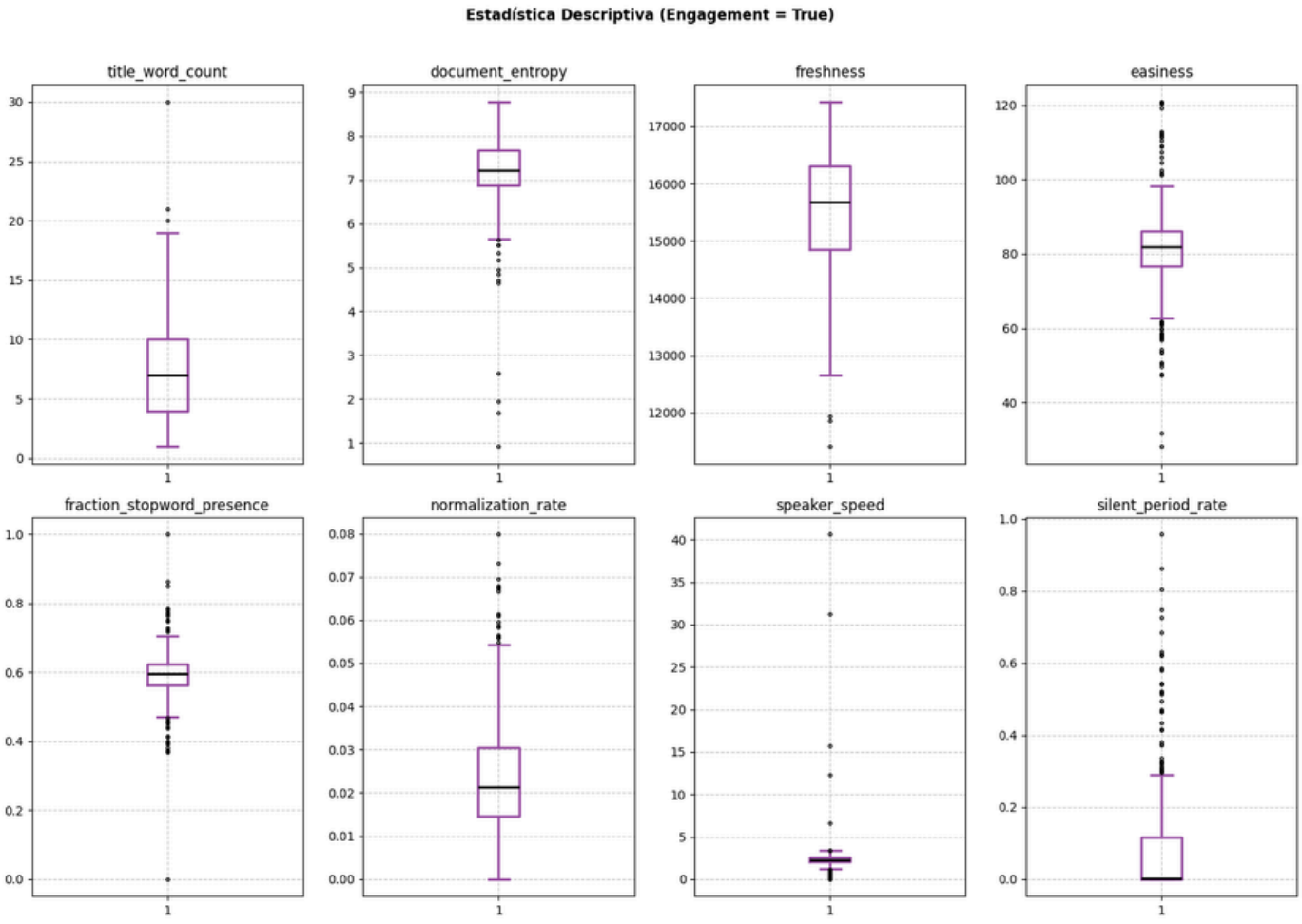
NUESTRA BASE DE DATOS...

El presente proyecto utiliza la base de datos proporcionada, la cual consta de 8 variables numéricas referentes a las características de cierto material digital, y una variable categórica llamada “engagement”, la cual es el indicador de clasificación, ya que señala si el material despertó interés e interacciones.

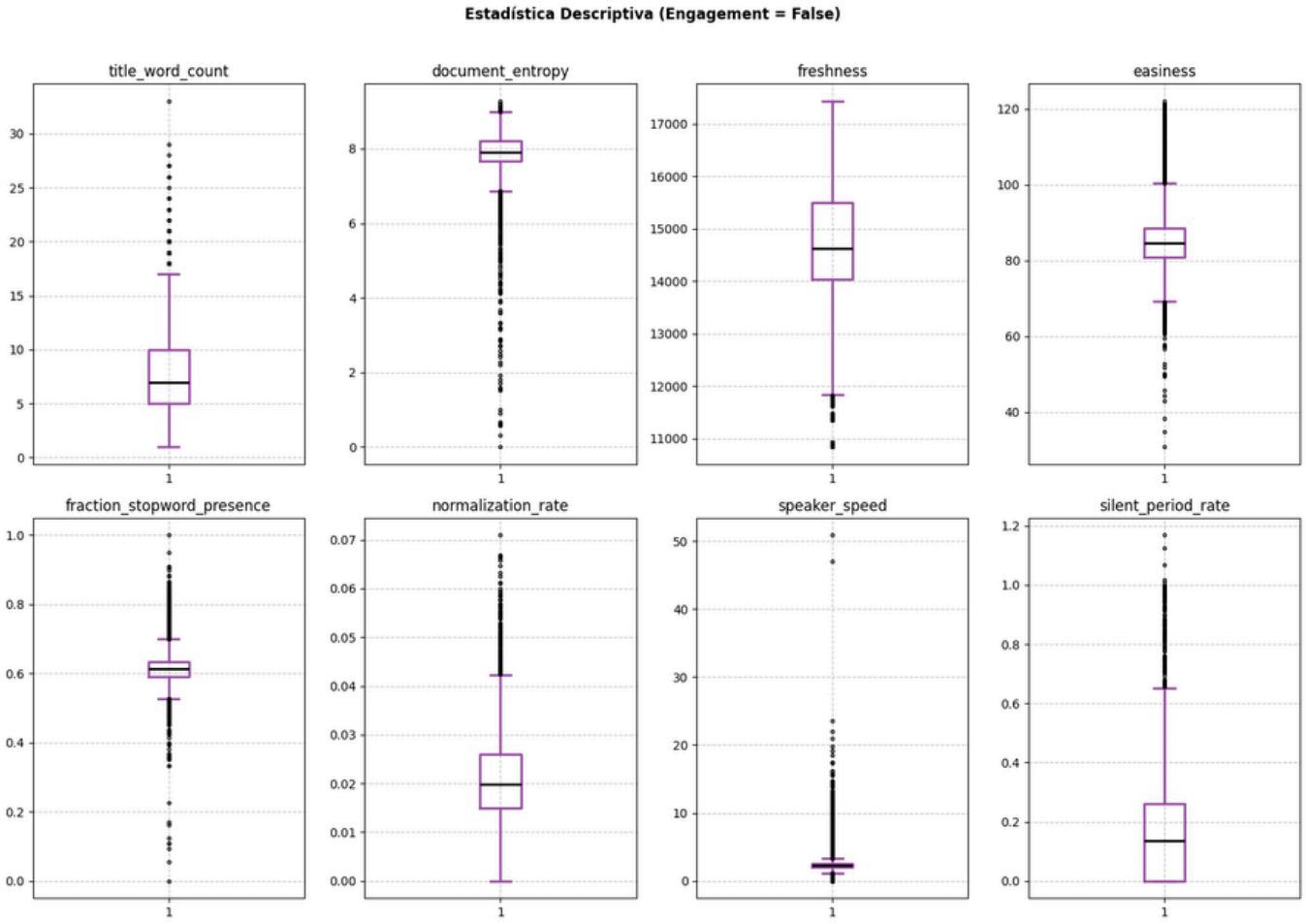
Se cuenta con dos archivos con datos, y se tomó el primero para entrenamiento y prueba. El 25% de estos registros se utilizará para hacer el test.

Analizando nuestros datos de entrenamiento, encontramos lo siguiente:

De un total de 6,929 registros, solo 661 tuvieron “engagement”. A continuación se muestran diagramas de cajas y bigotes para cada variable predictora:



El resto de los registros, es decir 6,268 no tuvieron “engagement”. Sus variables se muestran descritas estadísticamente en las siguientes gráficas:



A grandes rasgos se puede observar mayor dispersión en los casos negativos. Además, se nota diferencia respecto a los positivos, particularmente para las variables de document_entropy y silent_period_rate.

A continuación, se realizó la selección de parámetros y modelo a partir de los datos antes descritos, utilizando RandomForest.

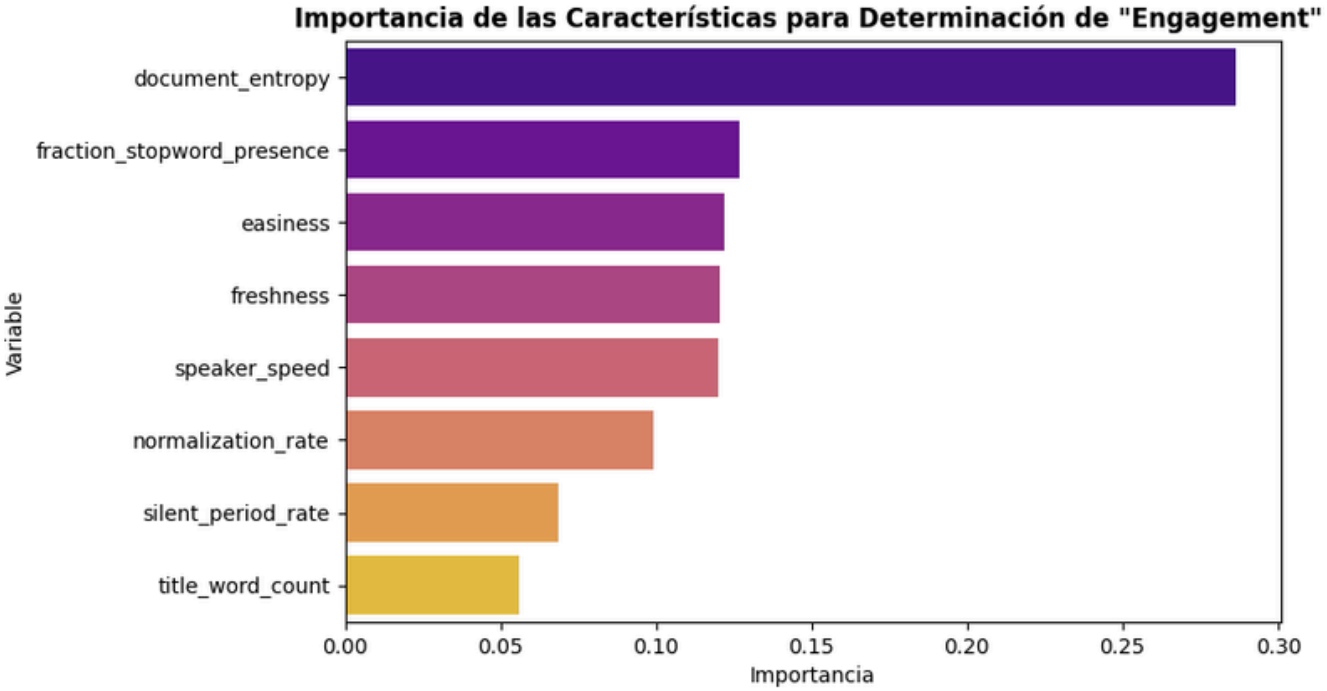
MEJORES PARÁMETROS

- Cantidad de árboles: 500
- Cantidad mínima de muestras: 2
- Límite de profundidad: 10

MEJOR ROC_AUC

89.41%

A través de un análisis de importancia para las características consideradas, se encontró que document_entropy tiene un rol de mayor peso en la clasificación del material. El resto de las variables tienen una relevancia similar, con excepción de silent_period_rate y title_word_count que calificaron con un porcentaje menor.

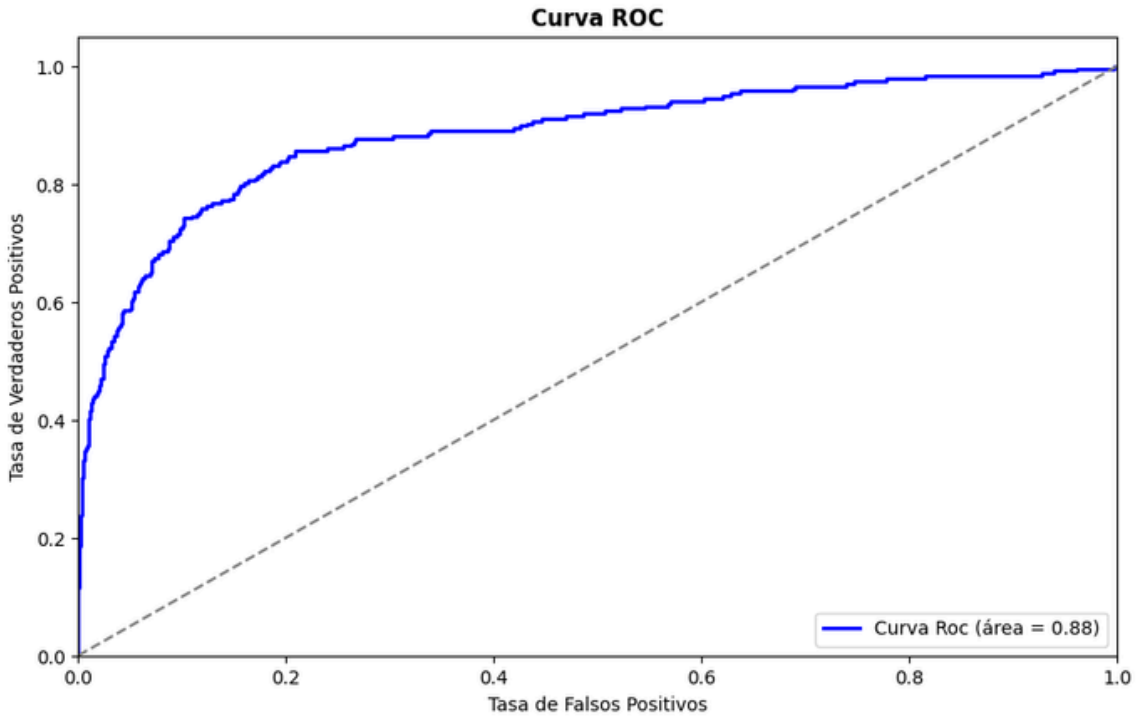


Testing

Considerando el modelo elegido, se realizó la prueba con las datos previamente seleccionados, teniendo el siguiente resultado para el criterio ROC_auc:

ROC_AUC

88.40%.



El porcentaje de puntuación, así como la gráfica del área bajo su curva indican un ajuste considerado como bueno respecto a los datos reales.

Nuevos Datos

Finalmente, tomando los registros del segundo archivo proporcionado, se hizo una clasificación de acuerdo al modelo elegido.

De un total de 2,309 elementos clasificados, se identificaron 140 como positivos para la categoría analizada, mientras que 2,169 se ubicaron como sin “engagement”

A continuación se expone una muestra de los resultados, donde “predictions” se refiere a la clasificacion del ítem, y “probability” a la probabilidad de que dicho material provoque el interés del usuario,

id	predictions	probability
9240	FALSO	1.92%
9241	FALSO	3.60%
9242	FALSO	7.96%
9243	VERDADERO	83.42%
9248	FALSO	8.82%
9249	FALSO	1.14%
9250	FALSO	1.45%
9251	FALSO	2.28%
9264	FALSO	1.64%
9272	FALSO	2.81%
9407	FALSO	3.68%
9413	FALSO	4.39%
9414	FALSO	35.51%
9415	VERDADERO	81.11%
9416	FALSO	13.57%
9417	FALSO	3.33%
9418	FALSO	0.83%
9419	VERDADERO	50.94%
9420	FALSO	4.13%
9421	FALSO	26.19%
9422	FALSO	0.90%
9423	FALSO	6.61%