Identificación de hojas de especificación

Alvaro Pequeño Mondragón

28 de Enero de 2025

Introducción

Debido al proceso manual de clasificación de documentos que se está haciendo actualmente, se tiene un cuello de botella a la hora de brindar una respuesta hacia nuestros clientes por lo que poder identificar las hojas de especificación con mayor velocidad nos permitiría tener mejores tiempos de respuesta.

Objetivo

Se quiere realizar un clasificador de documentos que nos permita identificar las hojas de especificación de entre otros documentos, el objetivo de este documento es poder identificar aquella combinación de variables que nos permitan tener una mayor precisión al momento de realizar esta clasificación usando el algoritmo de random forest, las variables a utilizar son:

- Número de palabras
- Número de caracteres
- Densidad de las palabras

Una vez identificada la combinación de variables con mejor desempeño se considerará una variable extra la cual es ver si el link de descarga del documento contiene la palabra "spec" y ver si esta información extra nos permite hacer una mejor identificación.

Por último si hay una mejora se procederá a buscar aquellos parametros del modelo de random forest que nos den el mejor resultado posible.

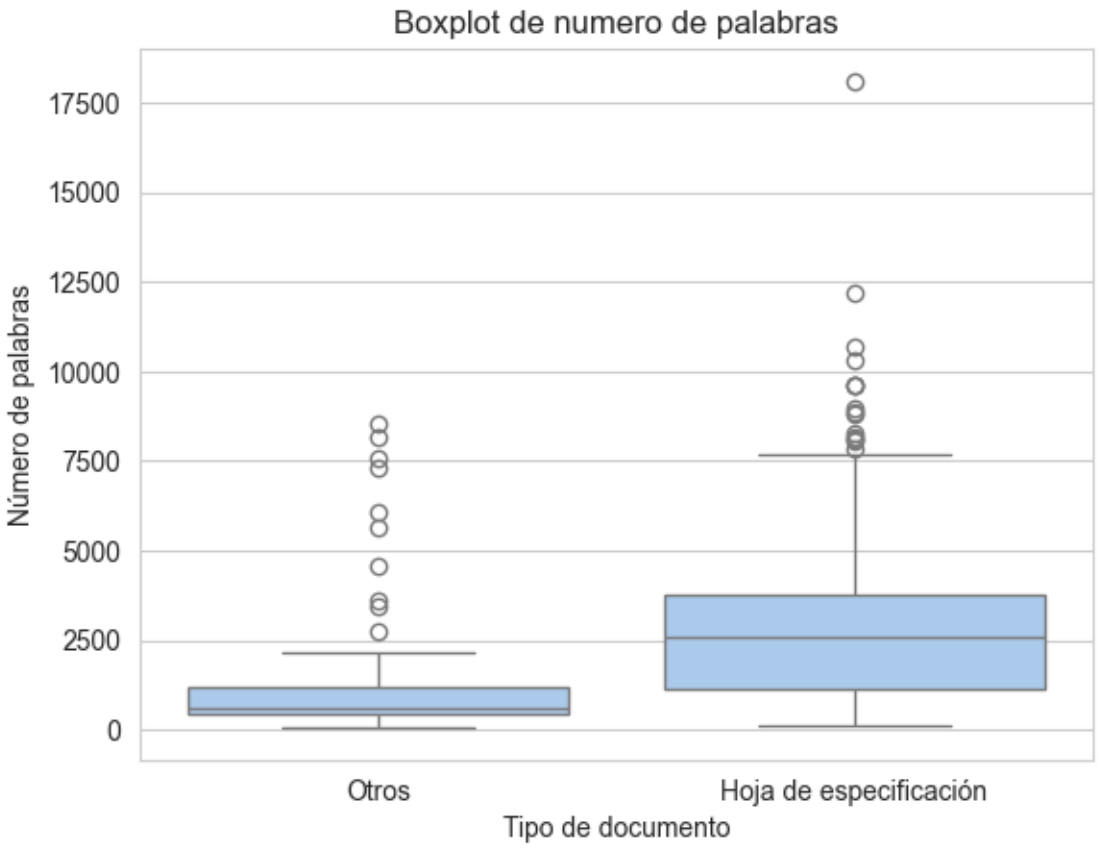
Todas las evaluaciones del modelo se harán con la métrica de precisión.

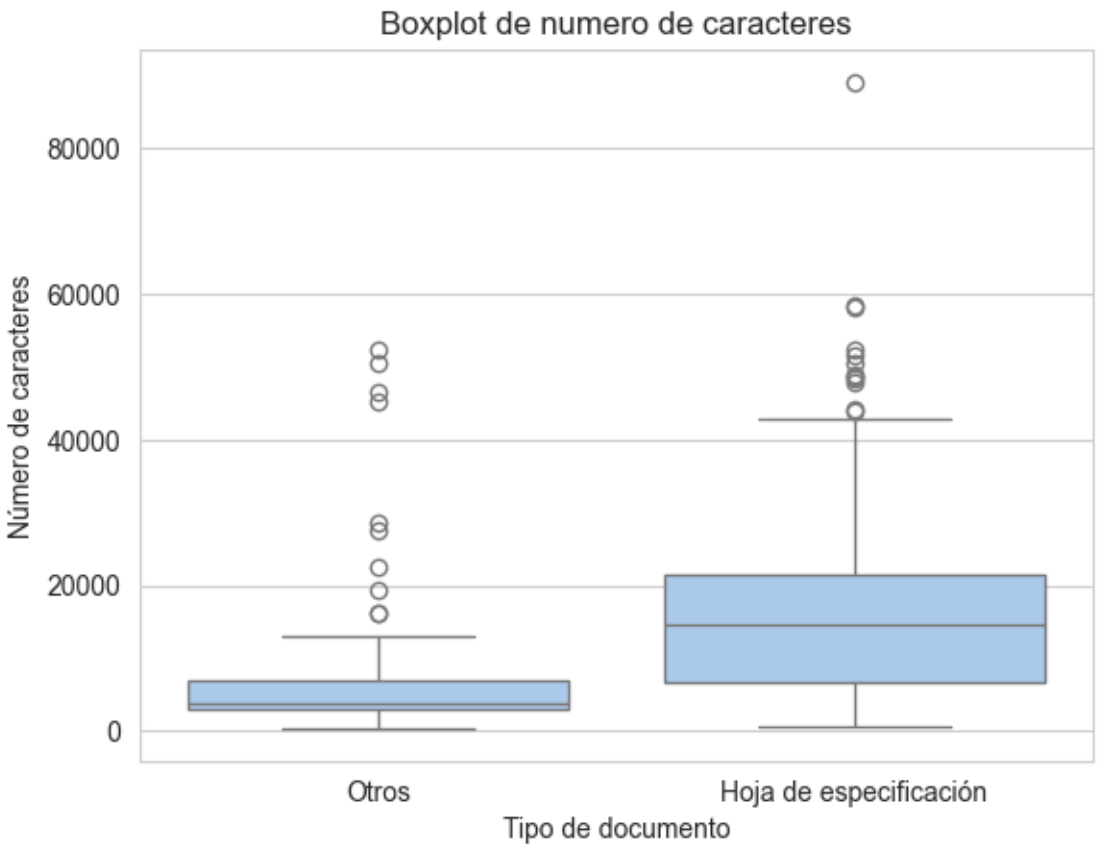
Experimentos

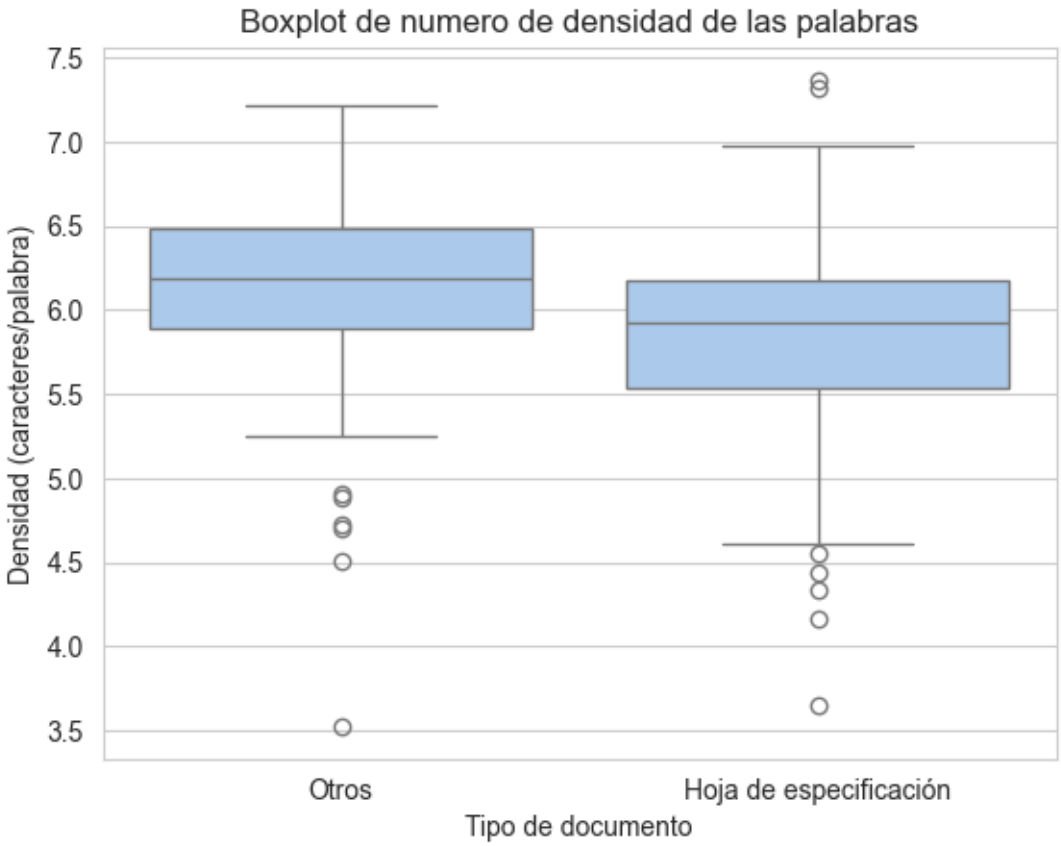
A continuación se muestra una tabla donde se visualizan las diferentes combinaciones de variables y las nomenclaturas con las que se identificaran dichas combinaciones:

	Número de palabras	Número de caracteres	Densidad de las palabras
Número de Palabras	W	-	-
Número de caracteres	WC	C	-
Densidad de las palabras	WD	CD	D

Antes de pasar a los resultados de los modelos se muestran unas graficas de caja y bigotes de las 3 variables para poder observar su distribución y sus medidas descriptivas:




Boxplot de numero de caracteres
Número de caracteres
80000
60000
40000
20000
0
Otros
Hoja de especificación
Tipo de documento



Viendo las graficas anteriores podemos observar que entre cada una de las clases hay diferencias visibles en las variables que se tomarán a cuenta para ejecutar los modelos.

Resultados

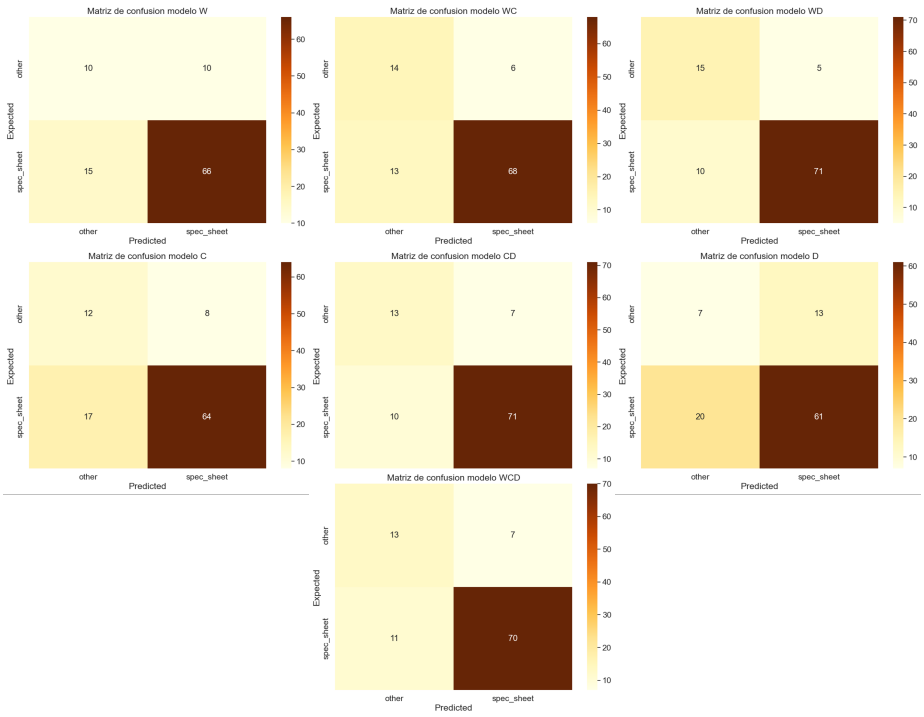
Como se mencionó en la sección de objetivo se realizaron varias iteraciones de modelos de random forest en los cuales se usa como entrada las diferentes combinaciones que se pueden hacer con las variables mencionadas con anterioridad. Una vez ejecutados los diferentes modelos se obtuvieron los siguientes resultados en la métrica de precisión:

	Número de palabras	Número de caracteres	Densidad de las palabras
Número de Palabras	0.7524	-	-
Número de caracteres	0.8118	0.7524	-
Densidad de las palabras	0.8514	0.8316	0.6732

El modelo que contiene las 3 variables (WCD) obtuvo una precisión de:

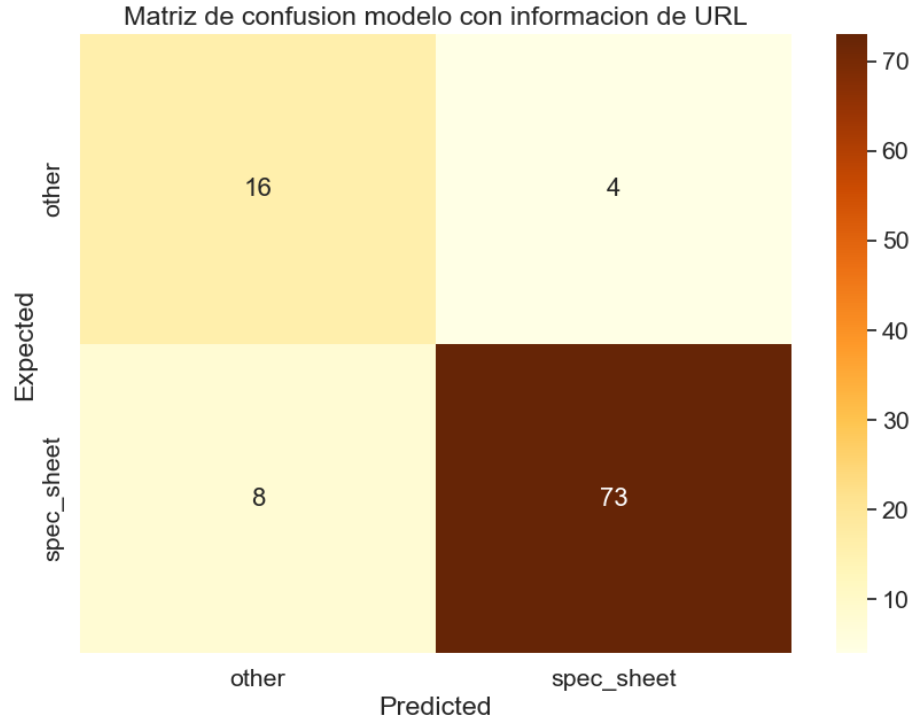
0.8217

A continuación se muestran las matrices de confusión para cada uno de los modelos:



Con los resultados anteriores observamos que la combinación que tiene una mejor precisión es la WD que son las variables de número de palabras y la densidad de las palabras. El siguiente paso fue incluir en el modelo información especifica del URL de descarga, se determina si el URL contiene la palabra "spec" y esto se utiliza como una variable de entrada adicional. El modelo con esta variable adicional obtuvo una precisión de: 0.8818

La matriz de confusión del modelo es la siguiente:



Podemos ver que agregar la información del URL de descarga nos mejora alrededor de un 3% por lo que se hará un gridsearch para encontrar aquellos parámetros óptimos que nos entregue la mejor precisión, los parámetros que se están buscando son los de "max_depth" y "n_estimators" donde los valores óptimos serian 9 y 50 respectivamente. Estos parámetros óptimos nos dan una precisión de: 0.8875 lo cual no es realmente una mejora respecto al modelo inicial.

Conclusiones

Observando los resultados de los diferentes modelos se puede observar que el número de variables que se utilizan en los modelos si tiene efecto en el modelo, pero contrario a lo que se pudiera pensar no siempre usar mas variables significará un mejor desempeño del modelo. Con los resultados obtenidos en los modelos podemos ver que el modelo que contiene las 3 variables tiene peor desempeño que 2 modelos de 2 variables.

Por ultimo se observo que al agregar una variable con características externas al texto nos puede ayudar a tener una mejora en el desempeño del modelo. Esta mejora es pequeña pero tal vez con una selección de alguna variable más significativa para la clasificación del documento se pueda tener un mejora más grande en el desempeño del modelo.