UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería
CC3085 – Inteligencia Artificial
Sección 30
Ing. Javier Fong

Proyecto #2
Filtro SPAM/HAM usando bayes

Carlos Daniel Estrada Vega - 20853

GUATEMALA, 04 de marzo del 2024

Análisis de datos exploratorio (EDA).

Para comenzar, se realizó un análisis exhaustivo de las columnas presentes en la base de datos, identificando un total de cinco columnas, denominadas v1, v2, y tres en blanco.

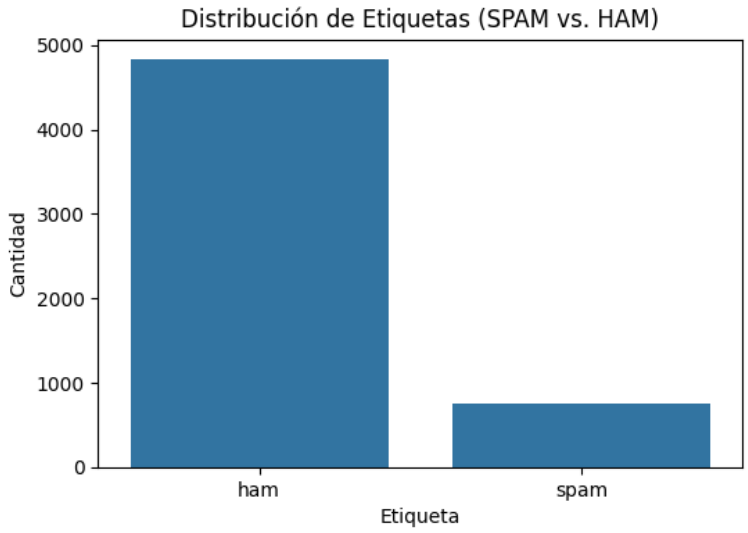
```
     v1                                                 v2 Unnamed: 2  \   Unnamed: 3 Unnamed: 4
0   ham  Go until jurong point, crazy.. Available only ...        NaN    0        NaN        NaN
1   ham                      Ok lar... Joking wif u oni...        NaN    1        NaN        NaN
2  spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN    2        NaN        NaN
3   ham  U dun say so early hor... U c already then say...        NaN    3        NaN        NaN
4   ham  Nah I don't think he goes to usf, he lives aro...        NaN    4        NaN        NaN
```

Se proporcionó una descripción básica del data frame, detallando la cantidad de datos no nulos en cada columna.

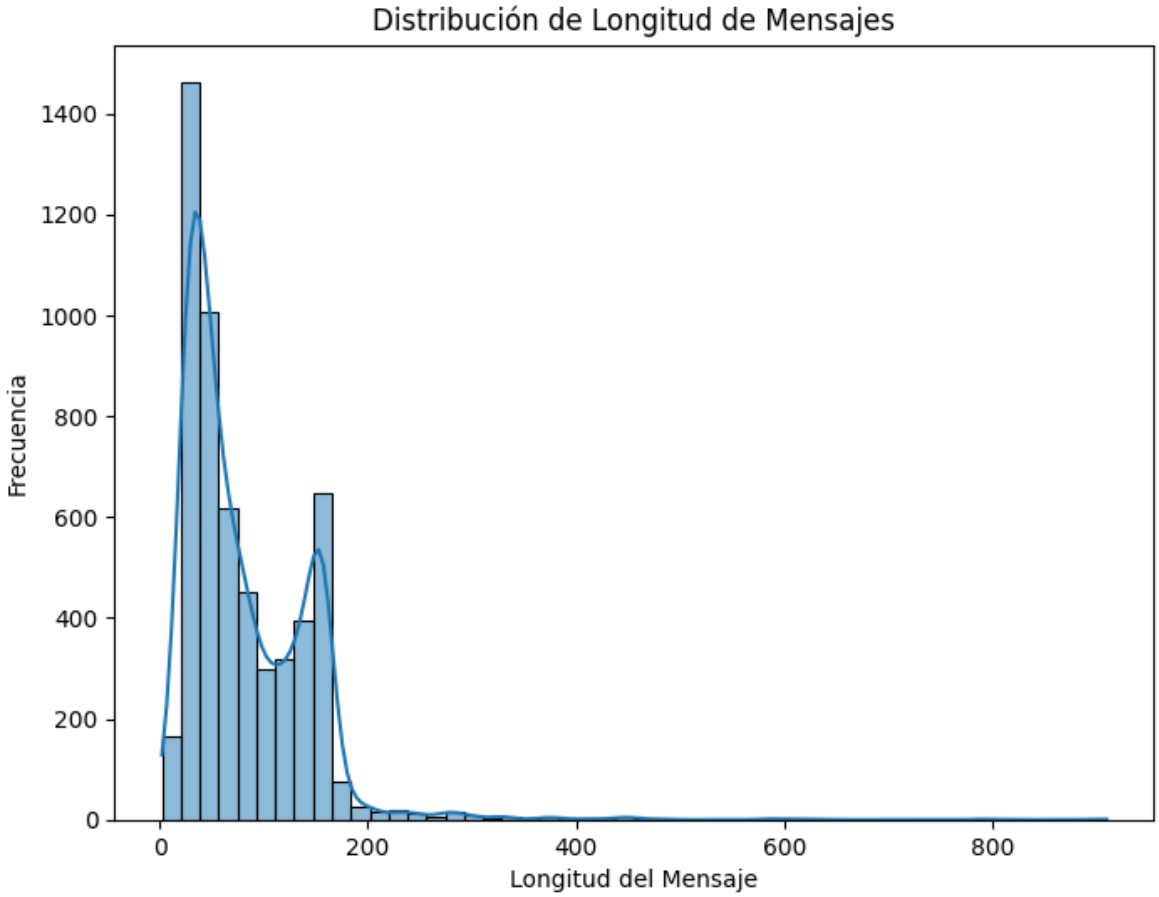
```
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

Además, se presentó un resumen de la clasificación de mensajes como SPAM y HAM en la base de datos, evidenciando que la mayoría de los mensajes pertenecen a la categoría HAM.

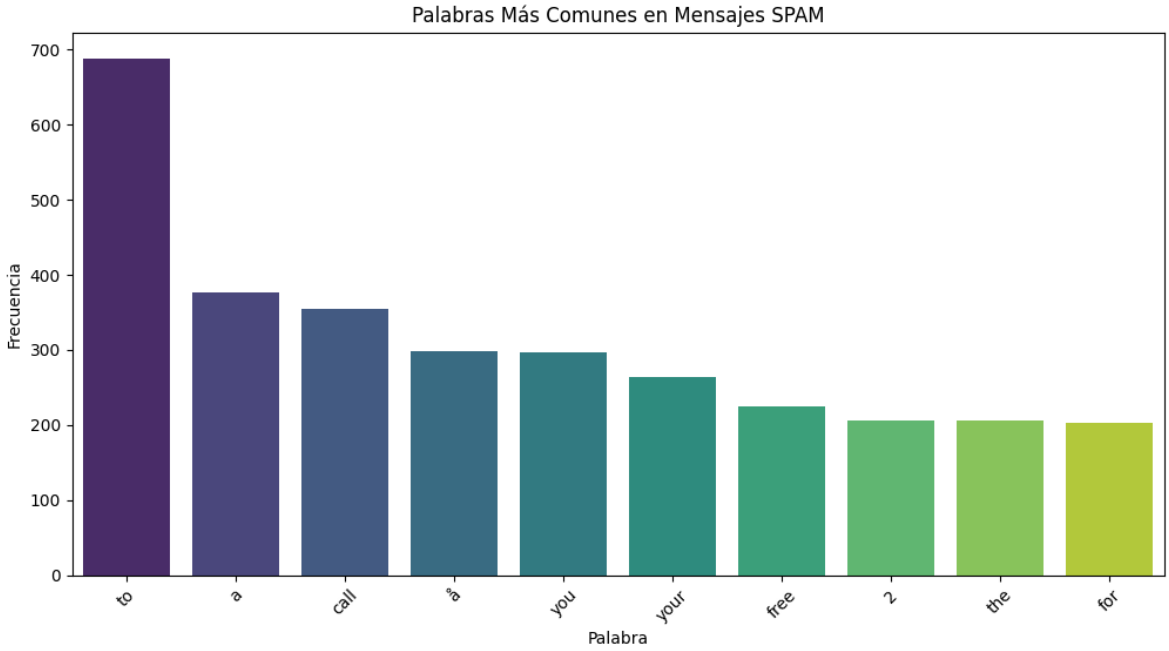
```
v1
ham     4825
spam     747
Name: count, dtype: int64
```

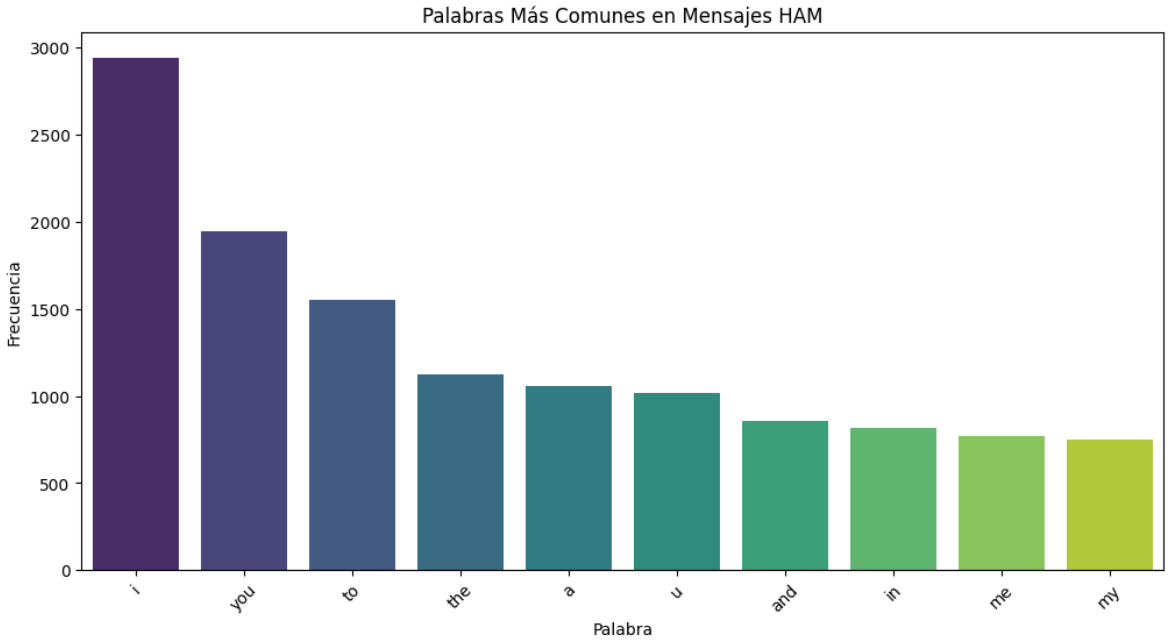


Se procedió a calcular la longitud de los mensajes y su distribución. Se observó que a medida que la longitud del mensaje aumenta, la frecuencia disminuye considerablemente. Se destacó la escasez de mensajes con longitudes superiores a 200 caracteres, lo que sugiere que los mensajes cortos son más frecuentes en la base de datos.

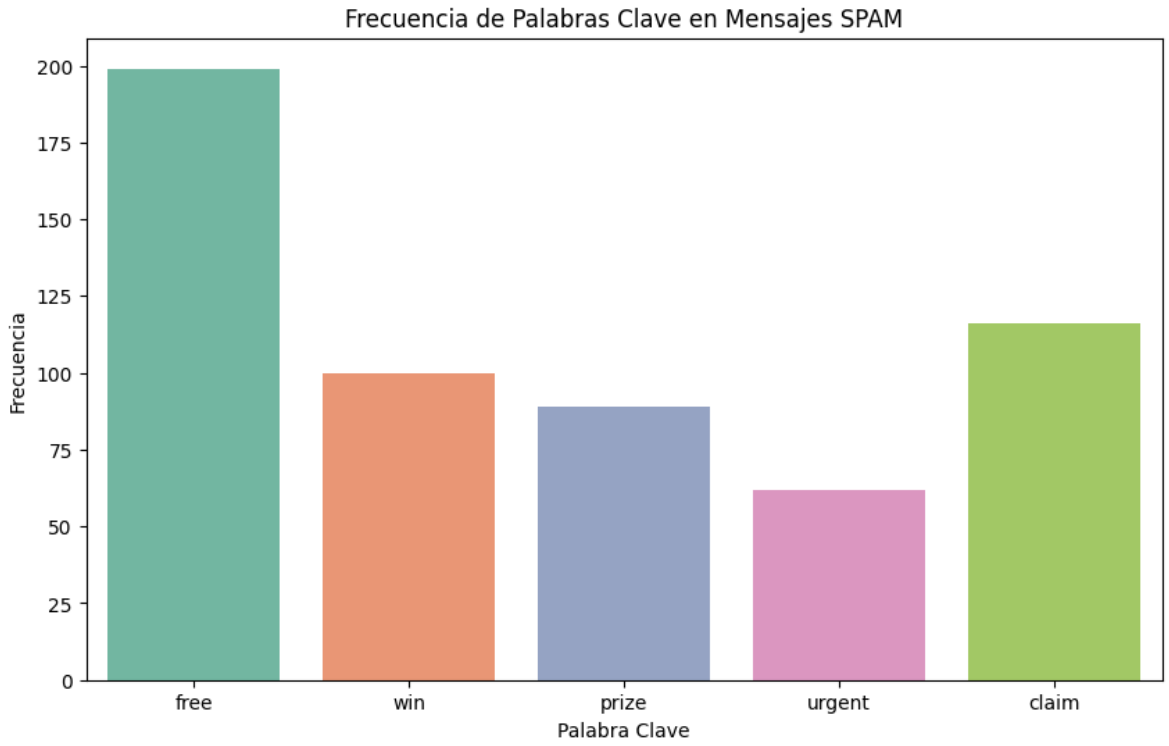


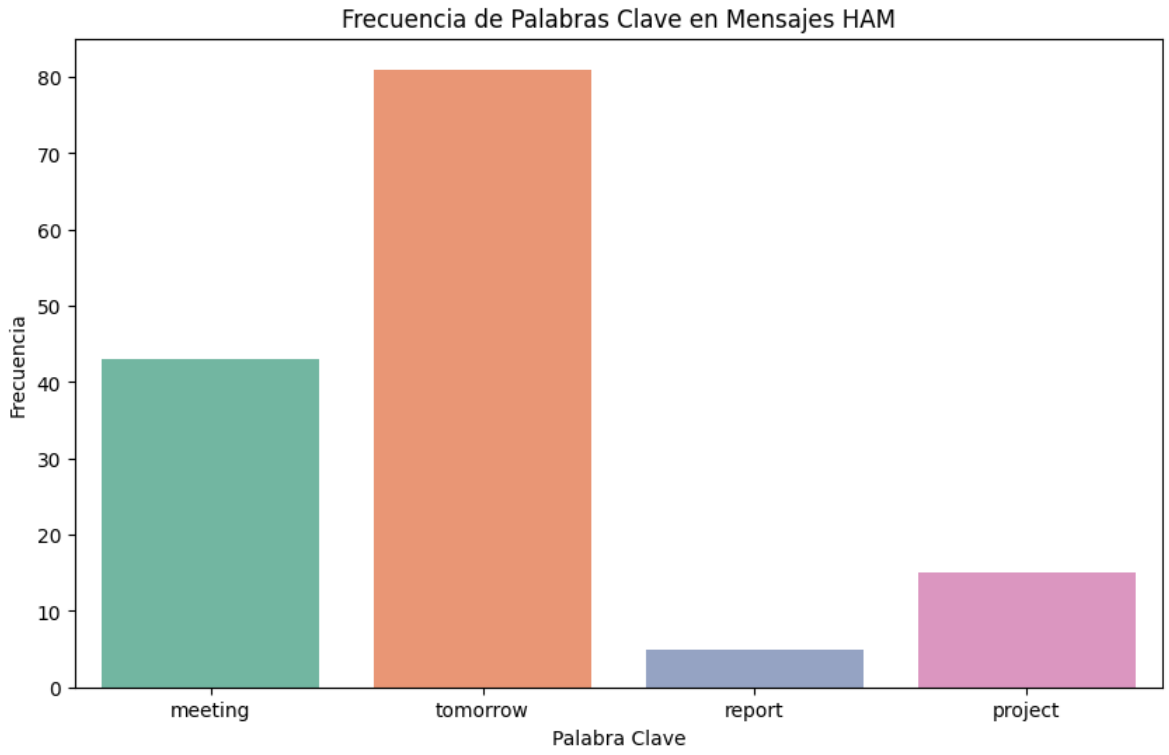
También se llevó a cabo el cálculo y la representación gráfica de las 10 palabras más frecuentes en los mensajes clasificados como SPAM y HAM, así como la frecuencia de palabras clave presentes en ambos tipos de mensajes.





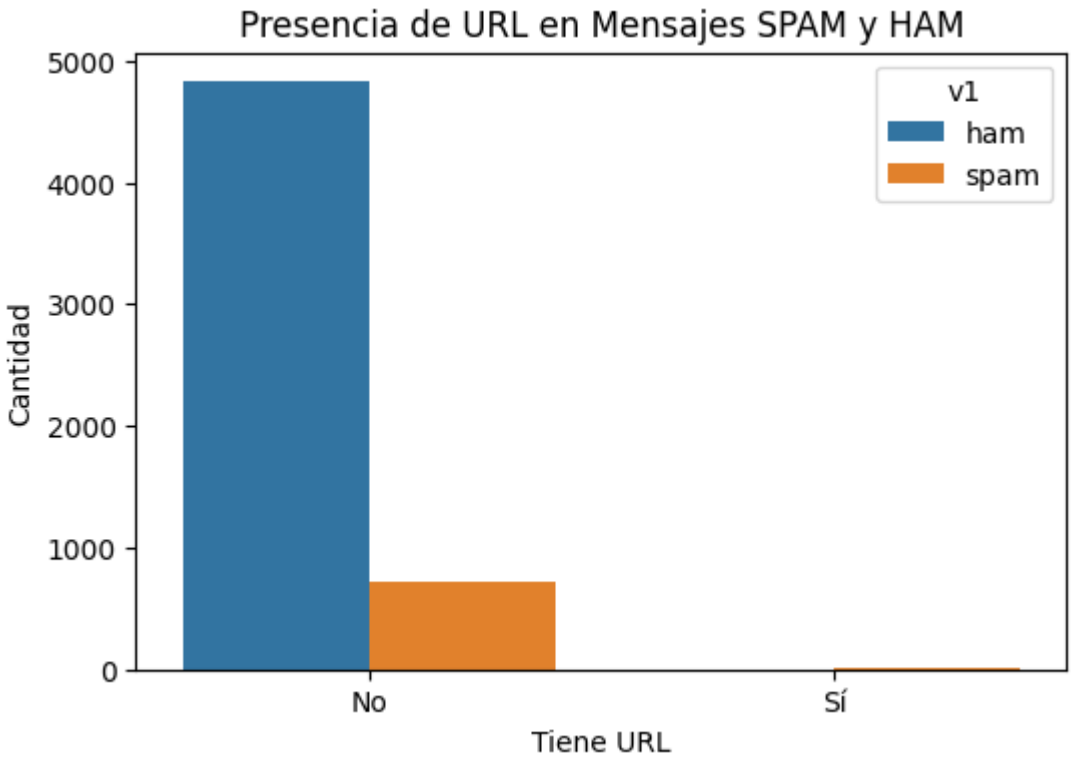
Aquí se aprecia que las palabras que más se repiten son palabras conectoras, como i, you, to, etc, las cuales no son palabras relevantes para predecir si el mensaje es spam o ham, sin embargo, que se repita la palabra call y free muchas veces en los spam si es un dato relevante.



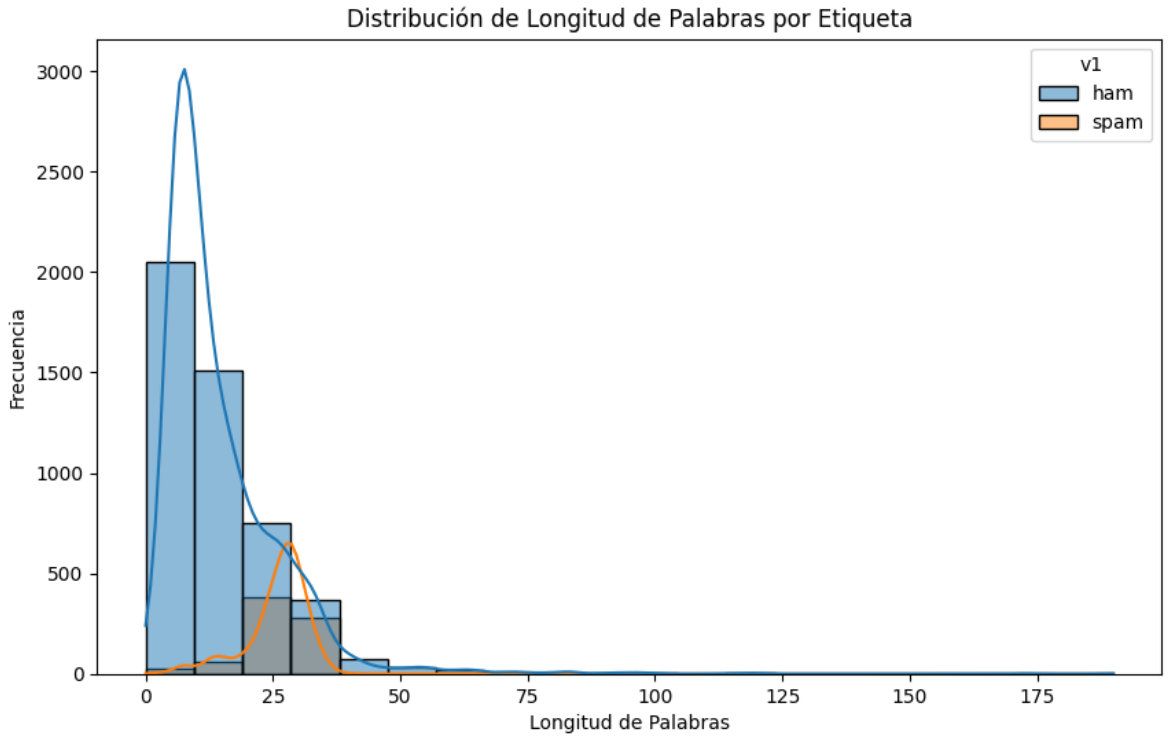


Por otro lado, se detectaron algunas palabras claves que se repiten muchas veces en los mensajes de cada tipo, siendo free, claim y free las más relevantes en los mensajes de tipo spam. Conocer estas palabras clave podría ayudar en la detección de SPAM o aumentar la conciencia sobre los patrones de los mensajes.

Se evaluó la presencia de enlaces URL en los mensajes y se analizó la longitud promedio de las palabras.

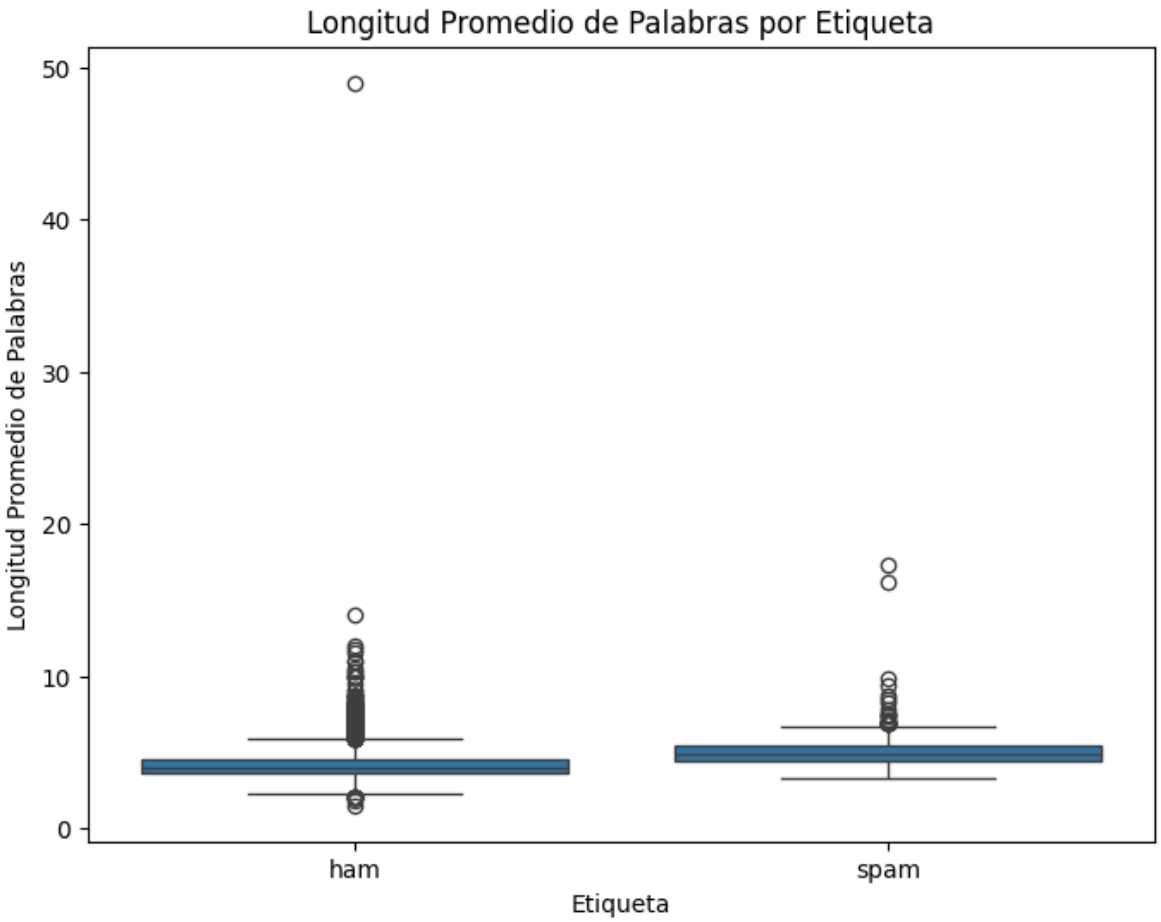


No se pudo interpretar más información de esta gráfica, más que no existe una relación entre los mensajes de tipo spam y los que contienen url para concluir que uno sea indicativo del otro.

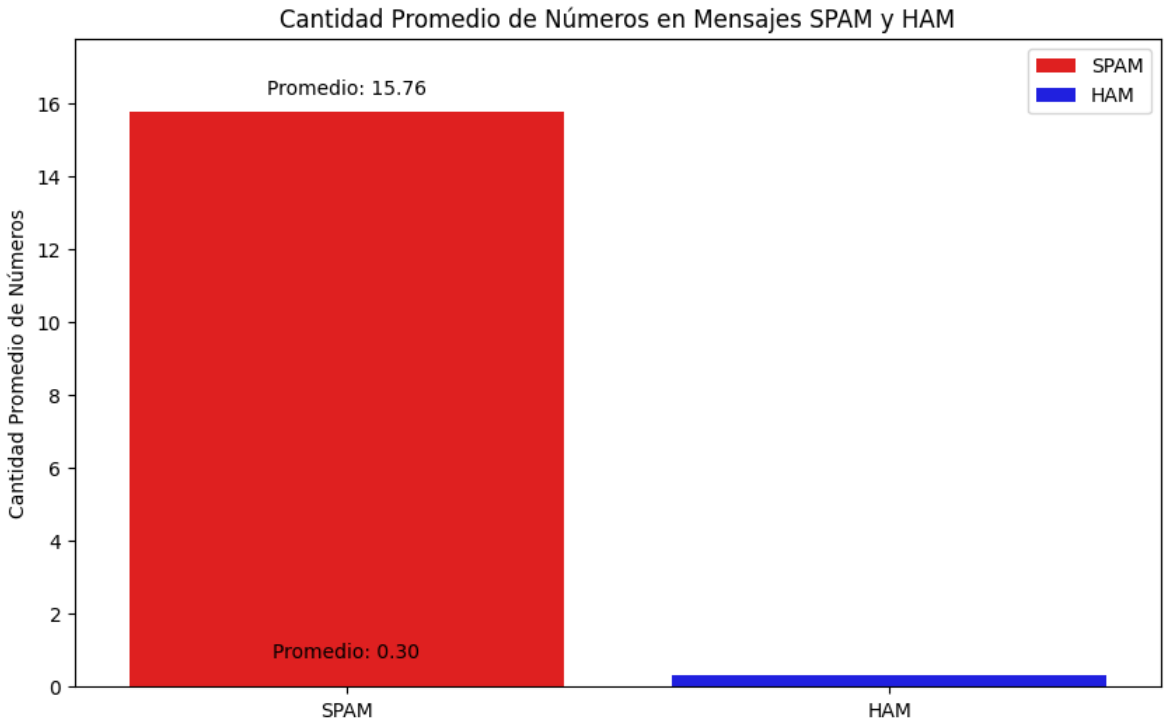


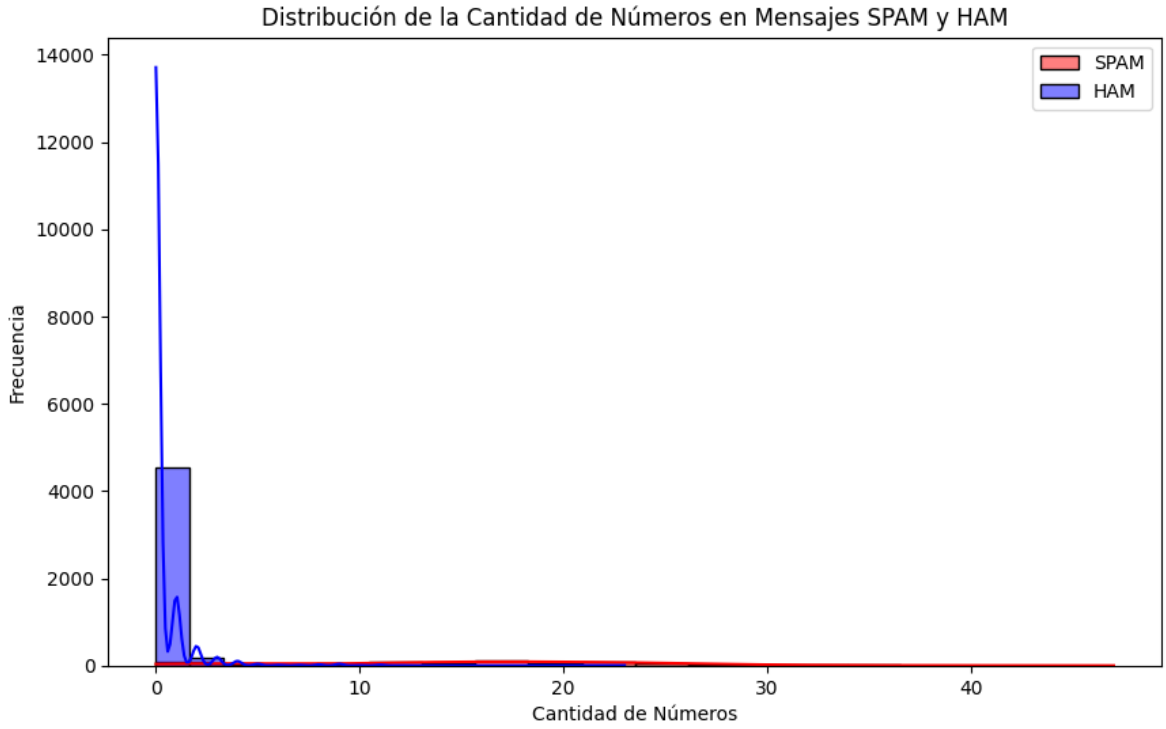
La mayoría de las palabras caen dentro de un rango más pequeño de longitudes, con muy pocas palabras que alcanzan longitudes superiores a 50 caracteres.

Los mensajes ‘ham’ y ‘spam’ siguen una tendencia similar, pero los ‘spam’ parecen tener una frecuencia ligeramente más alta en palabras más largas.

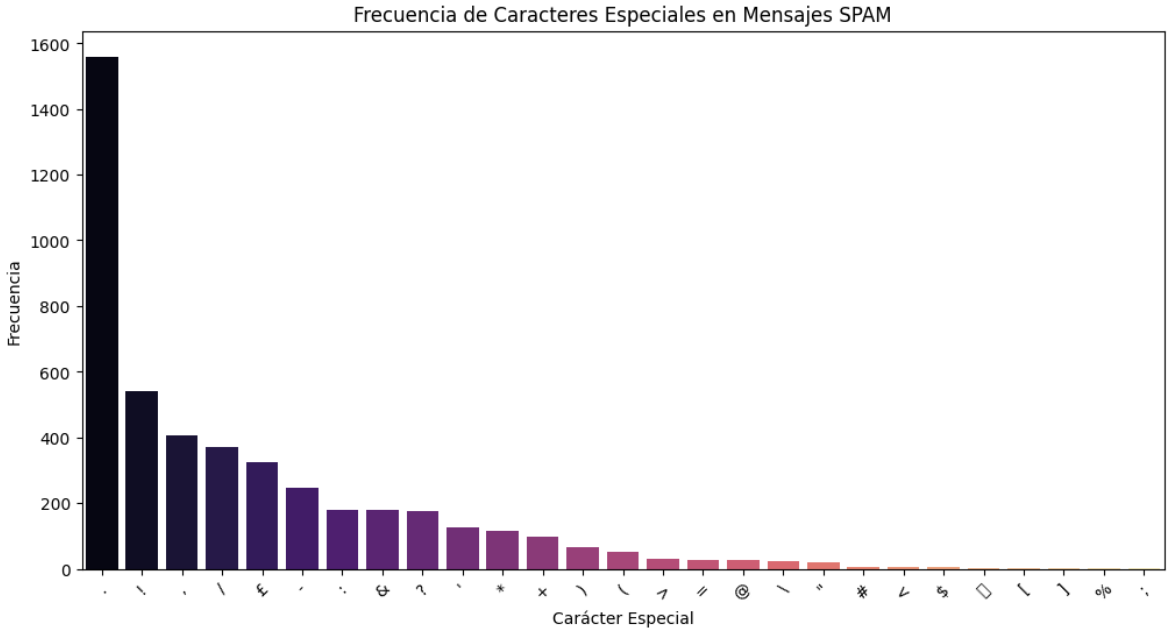


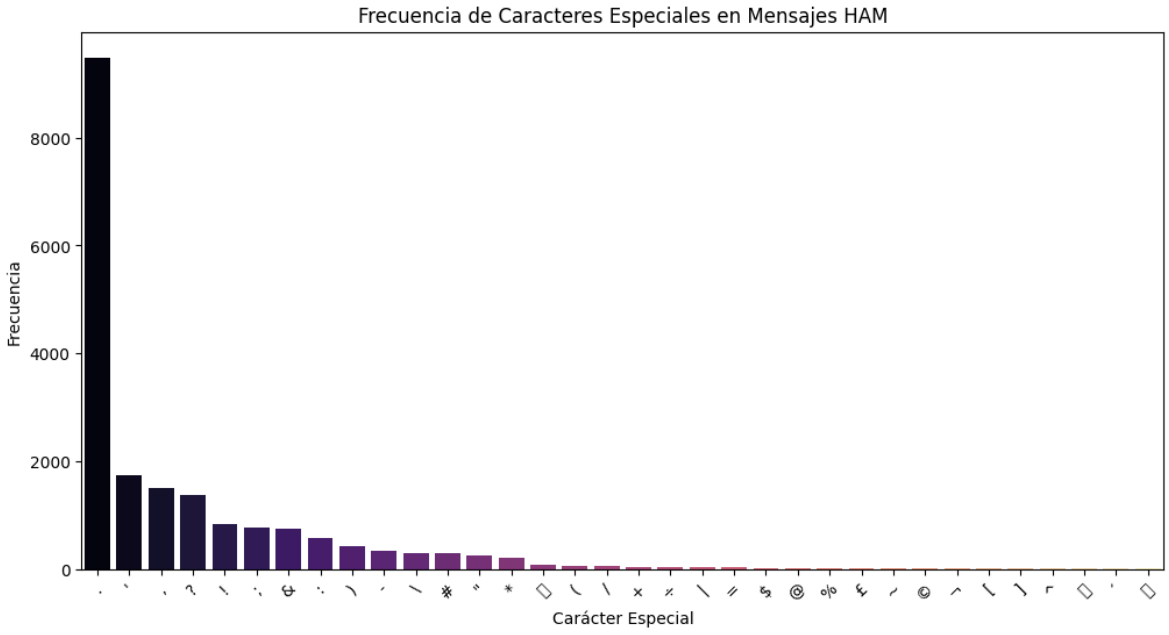
Finalmente, se examinó la cantidad de números y caracteres especiales presentes en los mensajes, diferenciando entre aquellos clasificados como HAM y SPAM.





Estas gráficas sugieren que los mensajes clasificados como SPAM tienden a tener un promedio significativamente mayor de números en ellos en comparación con los mensajes clasificados como HAM.





Los caracteres especiales más comunes en los mensajes HAM y SPAM son aquellos que se utilizan para separar palabras o frases, como guiones, puntos y comas. Debido a que estos caracteres especiales no nos ayudan al propósito de clasificar los mensajes y que en vez nos pueden generar ruido en el modelo, lo recomendable es eliminar estos caracteres cuando se haga la limpieza de los datos.

Limpieza de datos.

- Expansión de abreviaturas comunes:

Se reemplazan las abreviaturas comunes por sus equivalentes completos.

Beneficios: Ayuda a normalizar el texto y a asegurarse de que las abreviaturas no afecten negativamente a los modelos de análisis de texto, permitiendo que se capturen mejor los significados de las palabras.

- Eliminación de emoticones:

Se eliminan los emoticones del texto.

Beneficios: Los emoticones pueden no tener un significado semántico y pueden interferir con los modelos de análisis de texto, por lo que eliminarlos puede ayudar a mejorar la precisión del análisis.

- Tokenización:

Se divide el texto en tokens o palabras individuales.

Beneficios: Facilita el procesamiento del texto a nivel de palabra, lo que permite realizar operaciones específicas en cada palabra, como eliminación de puntuación, filtrado de stop words, lematización o derivación.

- Conversión a minúsculas y eliminación de palabras repetidas:

Se convierten todas las palabras a minúsculas y se eliminan las palabras repetidas.

Beneficios: Normaliza el texto y reduce la complejidad, asegurando que las palabras se traten de manera consistente independientemente de su formato y eliminando repeticiones innecesarias.

- Eliminación de puntuación:

Se eliminan los caracteres de puntuación del texto.

Beneficios: Permite que el modelo se centre en las palabras y su significado, ignorando los caracteres que no contribuyen directamente al contenido del texto.

- Filtrado de stop words:

Se eliminan las palabras comunes que no aportan un significado sustancial al texto, como "the", "is", "in", etc.

Beneficios: Reduce el ruido en el texto y ayuda a centrarse en las palabras más importantes, lo que puede mejorar la precisión de los modelos de análisis de texto al eliminar palabras irrelevantes.

- Lematización:

Se reduce cada palabra a su forma base o lema.

Beneficios: Ayuda a normalizar las palabras y reduce la dimensionalidad del espacio de características, lo que facilita el análisis al agrupar palabras con significados similares.

Modelo.

Pasos que sigue el programa

Carga de datos y preparación: Se cargan los datos limpios que previamente se han sometido a limpieza de texto para eliminar ruido y normalizar el formato. Se agregan características como la longitud del mensaje y la frecuencia de palabras clave asociadas con SPAM.

Entrenamiento y prueba del modelo: Los datos se dividen en un conjunto de entrenamiento y un conjunto de prueba. Utilizamos el conjunto de entrenamiento para calcular las probabilidades relevantes para nuestro modelo bayesiano.

Cálculo de probabilidades condicionales: Se calculan las probabilidades condicionales de que una palabra aparezca en un mensaje SPAM o HAM, así como la probabilidad de cada palabra en el conjunto de entrenamiento.

Predicción de SPAM o HAM para nuevos textos: Cuando llega un nuevo mensaje, se calcula la probabilidad de que sea SPAM utilizando las probabilidades condicionales previamente calculadas. Si la probabilidad es mayor que 0.5, se clasifica como SPAM; de lo contrario, se clasifica como HAM.

Cálculo de probabilidades condicionales de palabras dadas las etiquetas SPAM o HAM (P(W|S) y P(W|H)):

- Se cuenta cuántas veces aparece cada palabra en los mensajes etiquetados como SPAM y como HAM en el conjunto de entrenamiento.
- Se calcula la probabilidad condicional de que una palabra específica aparezca en un mensaje SPAM (P(W|S)) dividiendo el número de veces que aparece esa palabra en los mensajes SPAM por el total de mensajes SPAM.
- Se realiza el mismo cálculo para las palabras en los mensajes HAM (P(W|H)).

Cálculo de la probabilidad prior de SPAM y HAM (P(S) y P(H)):

- Se cuenta el número de mensajes etiquetados como SPAM y como HAM en el conjunto de entrenamiento.
- Se calcula la probabilidad prior de SPAM (P(S)) dividiendo el número de mensajes SPAM por el total de mensajes en el conjunto de entrenamiento.
- Se calcula de manera similar la probabilidad prior de HAM (P(H)).

Cálculo de la probabilidad de que una palabra sea SPAM dado un mensaje (P(S|W)):

- Dado un mensaje, se tokeniza en palabras individuales.
- Para cada palabra en el mensaje, se calcula la probabilidad de que sea SPAM dado que aparece esa palabra (P(S|W)) utilizando el teorema de Bayes:
- P(W|S) es la probabilidad de que la palabra aparezca en un mensaje SPAM (calculada en el paso 1).
- P(S) es la probabilidad prior de SPAM (calculada en el paso 2).
- P(W|H) es la probabilidad de que la palabra aparezca en un mensaje HAM (calculada en el paso 1).
- P(H) es la probabilidad prior de HAM (calculada en el paso 2).

Predicción de la etiqueta SPAM o HAM para el mensaje:

- Se multiplica todas las probabilidades P(S|W) calculadas para cada palabra en el mensaje para obtener la probabilidad conjunta de que el mensaje sea SPAM dado esas palabras.
- Si la probabilidad conjunta es mayor que 0.5, el mensaje se clasifica como SPAM; de lo contrario, se clasifica como HAM.

Pruebas de rendimiento.

```
Accuracy: 0.9300448430493273
Precision: 0.6698564593301436
Recall: 0.9395973154362416
F1-score: 0.7821229050279329
Confusion Matrix:
[[897  69]
 [  9 140]]
PS C:\Users\estra\OneDrive\Documentos\UVG\Inteligencia\proyecto2>
```

Discusión de resultados.

Nuestra métrica de **exactitud (Accuracy)** evalúa qué proporción de las predicciones de nuestro modelo son correctas en comparación con el total de predicciones realizadas. En este caso, el modelo logra una exactitud del 93.00%, lo que significa que aproximadamente el 93% de las predicciones son correctas. Este resultado positivo se debe al proceso de limpieza de datos y al modelo de clasificación, que han mejorado la capacidad del modelo para distinguir con precisión entre mensajes SPAM y HAM.

La **precisión**, por otro lado, se enfoca en la proporción de predicciones positivas correctas sobre el total de predicciones positivas hechas por el modelo. Aquí, la precisión es del 66.99%, lo que indica que alrededor del 67% de los mensajes clasificados como SPAM por el modelo son realmente SPAM. Esto sugiere que las decisiones tomadas durante el proceso de limpieza de datos y el modelado, como el uso de características relevantes como la longitud del mensaje y la frecuencia de palabras clave, han contribuido a mejorar la precisión.

El **recall**, por su parte, mide la proporción de casos positivos reales que el modelo identifica correctamente sobre el total de casos positivos reales en los datos. En este caso, es del 93.96%, lo que significa que el modelo identifica correctamente aproximadamente el 94% de todos los mensajes SPAM presentes en los datos. La eliminación de palabras comunes y la lematización podrían haber ayudado a capturar de manera más efectiva los términos clave asociados con el SPAM.

El **F1-score** representa un equilibrio entre precisión y recall, siendo útil especialmente cuando las clases están desequilibradas, como en el caso del SPAM, donde hay muchos más mensajes HAM que SPAM. En este caso, el F1-score es del 78.21%, lo que indica un buen equilibrio entre precisión y recall, sugiriendo que nuestro modelo funciona bien en general.

Por último, en lo que respecta a la matriz de confusión, se puede observar que, 140 mensajes que fueron correctamente clasificados como SPAM **(Verdaderos positivos (TP))**, 897 mensajes que fueron correctamente clasificados como HAM. **(Verdaderos negativos (TN))**, 69 mensajes que fueron incorrectamente clasificados como SPAM cuando en realidad eran HAM (**Falsos positivos (FP))** Y 9 mensajes que fueron incorrectamente clasificados como HAM cuando en realidad eran SPAM (**Falsos negativos (FN))**

La alta cantidad de verdaderos positivos (140) y de verdaderos negativos (897), indica que el modelo es efectivo para identificar mensajes que son realmente SPAM y que son HAM correctamente.

La tasa de falsos positivos y falsos negativos son relativamente baja, ya que solo una pequeña parte de los casos negativos se clasificó incorrectamente como positivos, lo cual indica un buen rendimiento en la identificación de casos positivos.

Código:

https://github.com/DannyEst6109/Proyecto2_InteligenciaA