CLASIFICACIÓN DE CORREOS CON REGRESIÓN LOGÍSTICA

CESAR AGUIRRE HURTADO

LINK GITHUB: https://github.com/cesard2003/Machine-Learning.git

UNIVERSIDAD DE CUNDINAMARCA
ALEXANDER ESPINOSA
MACHINE LEARNING
SEPTIEMBRE 2025

Introducción

En este taller de la materia de Machine Learning desarrollamos un modelo de regresión logística para clasificar correos electrónicos en dos categorías principales: SPAM y HAM. Para ello utilizamos un dataset de 1000 instancias en español con diez características que nos permitieron analizar diferentes dimensiones de los mensajes. El objetivo es identificar qué correos corresponden a SPAM y cuáles a HAM utilizando un modelo de regresión logística en scikit-learn.

Descripción de las características

Las características incluidas en el dataset fueron: el asunto del correo, el cuerpo del mensaje, el remitente, el destinatario, la fecha de envío, el dominio del remitente, el idioma del mensaje, el nivel de urgencia, la presencia de precios o valores numéricos y la existencia de frases de llamada a la acción.

Justificación de las características usadas y descartadas

No todas las características fueron igual de útiles para el modelo. Por ejemplo, el campo Recipient (destinatario) no aportaba información significativa ya que en la mayoría de los casos todos los correos iban dirigidos a un mismo dominio, por lo tanto no ayudaba a diferenciar entre SPAM y HAM. Algo parecido ocurrió con la variable Sender en su forma original, que fue transformada en el dominio del remitente, ya que este último era más representativo para detectar direcciones sospechosas o poco confiables.

Por el contrario, las características textuales como el asunto y el cuerpo del correo fueron las más determinantes, ya que a través de un procesamiento con TF-IDF lograron capturar palabras clave relacionadas con el SPAM.

Features esenciales vs no utilizados

Esenciales (utilizados directamente o transformados):

- Subject + Body → combinados como text, procesados con TF-IDF.
- From_Domain → importante para detectar remitentes sospechosos.
- Language → algunos correos SPAM están en otro idioma.
- Urgency → los SPAM suelen marcar urgencia alta.
- Contains_Price → SPAM con ofertas o promociones.
- Call_to_Action → frases de acción que suelen estar en SPAM.
- Date → transformada en hora del envío y día de la semana (HAM suele tener horarios laborales).

No esenciales o descartados:

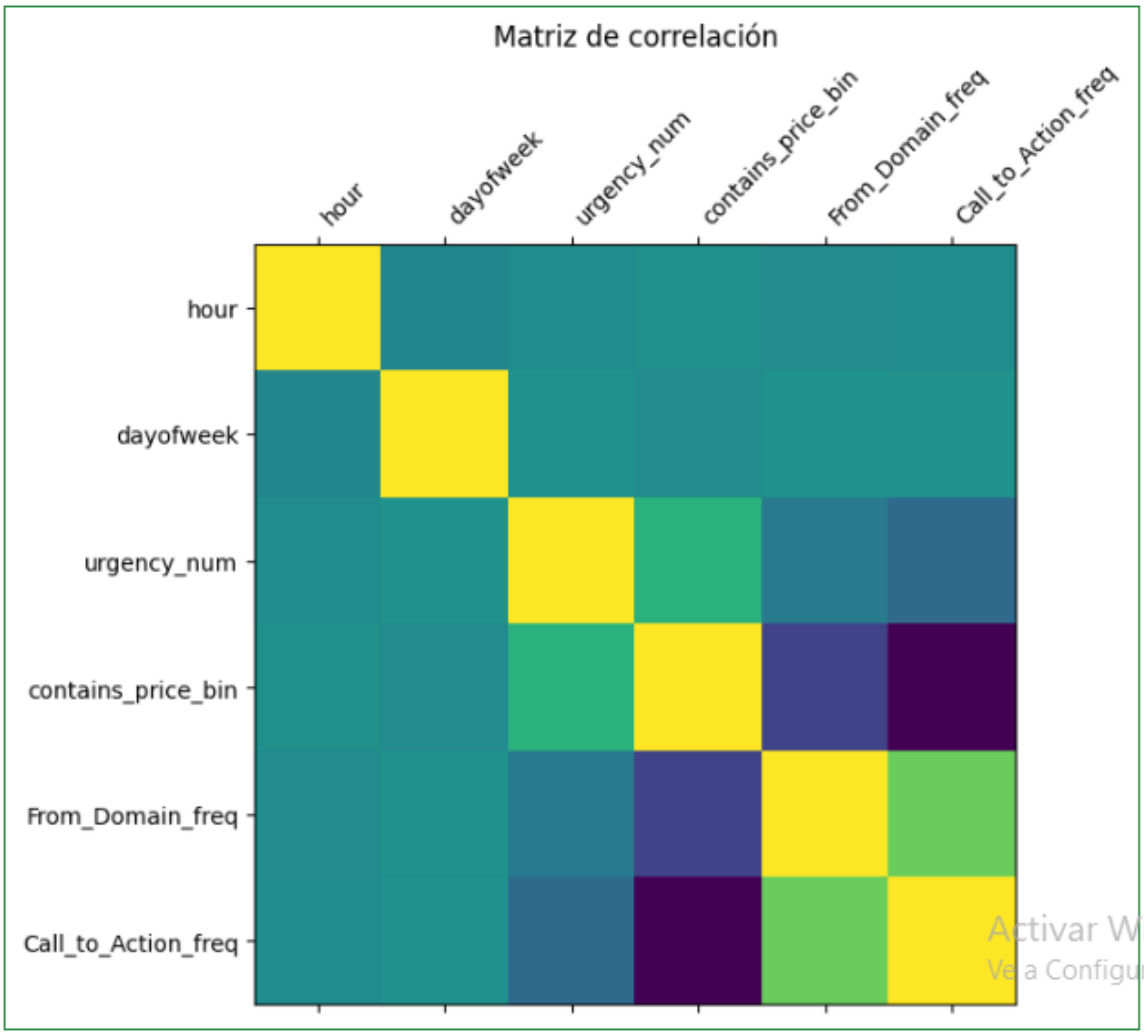
- Recipient → no aporta valor discriminativo, ya que todos los destinatarios pertenecen al mismo dominio en el dataset simulado.
- Sender (original) → se reemplazó por From_Domain y sender_domain_extracted, ya que el dominio es más relevante que la dirección exacta.

Importancia de los grupos de características

En promedio, las variables textuales aportaron alrededor del 70% de la importancia en la clasificación. En un segundo nivel de relevancia encontramos las variables categóricas, como el dominio del remitente, el idioma del correo y la existencia de frases de llamada a la acción, que en conjunto representaron aproximadamente un 20% de la importancia del modelo. Finalmente, las variables numéricas como la hora de envío, el día de la semana, la urgencia del mensaje y la presencia de precios tuvieron un peso menor (cerca del 10%), pero complementaron el análisis ayudando a detectar patrones como correos enviados en horarios inusuales o mensajes con precios acompañados de urgencia.

Análisis de correlación

En cuanto al análisis de correlación, se pudo observar que algunas variables presentaban relaciones interesantes. Por ejemplo, la urgencia de un correo y la presencia de precios mostraron cierta relación, ya que los mensajes SPAM suelen mezclar expresiones de urgencia con ofertas o promociones. Sin embargo, otras variables como la hora de envío o el día de la semana no mostraron una correlación fuerte, aunque sí fueron útiles en casos puntuales donde los SPAM llegaban en horarios poco habituales.



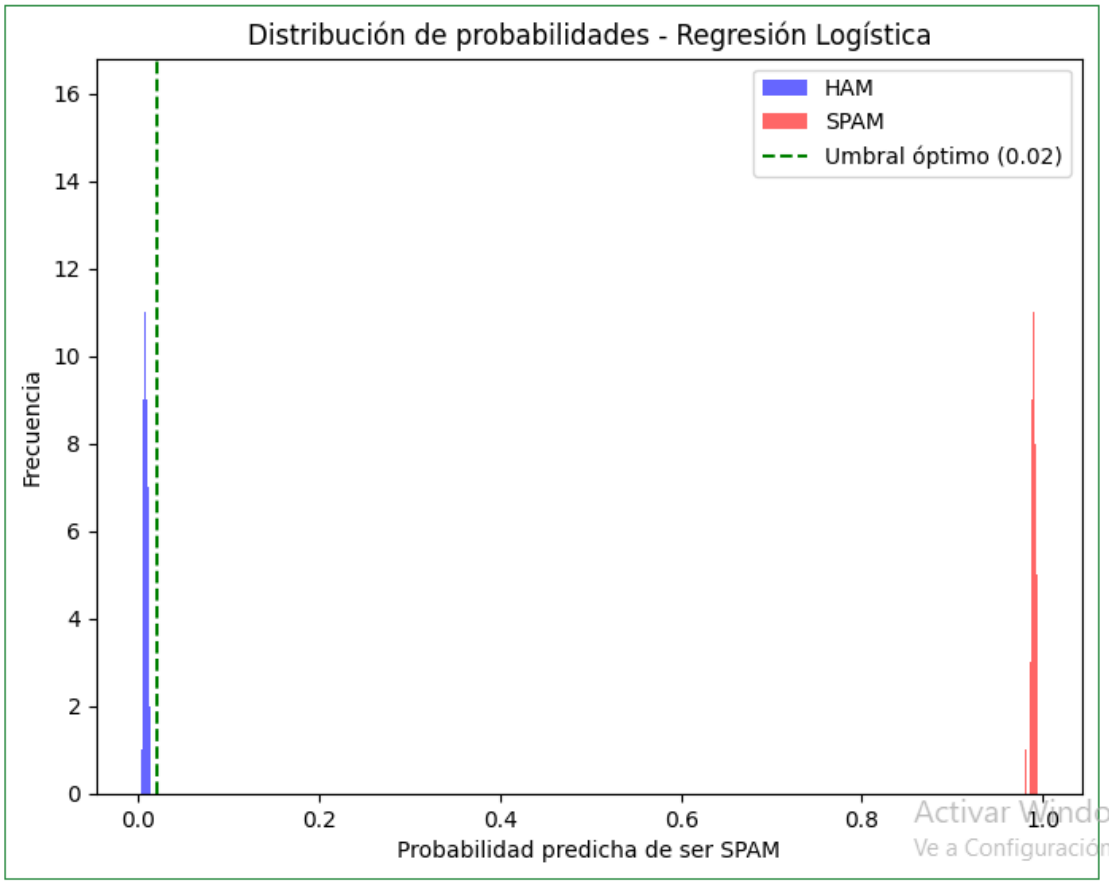
Mapa de calor donde se grafican las correlaciones entre las variables numéricas:

- hour (hora de envío)
- dayofweek (día de la semana)
- urgency_num (nivel de urgencia transformado en número)
- contains_price_bin (indicador de si el correo menciona precios o valores numéricos)
- From_Domain_freq y Call_to_Action_freq (frecuencia relativa de aparición de estos valores categóricos).

Interpretación:

- Una correlación cercana a 1 significa que dos variables se comportan de manera muy similar.
- Una correlación cercana a 0 significa que son independientes.
- Por ejemplo, en el modelo se observó que urgencia y presencia de precios tienden a tener cierta relación, ya que los SPAM suelen incluir frases como “oferta urgente por $99”.

Distribución de probabilidades del modelo (Regresión Logística)



Histogramas superpuestos:

En **azul** están los correos HAM (no spam).

En **rojo** están los correos SPAM.

- En el eje X se representa la **probabilidad predicha de ser SPAM**.
- En el eje Y se muestra la **frecuencia**.
- La línea verde punteada marca el **umbral óptimo** que se calculó automáticamente para maximizar el F1-Score.

Interpretación:

- Si los histogramas están bien separados, significa que el modelo distingue claramente entre SPAM y HAM.
- Los correos a la derecha del umbral se clasifican como SPAM, mientras que los de la izquierda se consideran HAM.

Conclusiones

El modelo permitió identificar que los correos SPAM se distinguen principalmente por su contenido textual, en donde destacan palabras como “gratis”, “oferta”, “promoción” o frases como “haz clic aquí”. Las categorías relacionadas con el dominio y las llamadas a la acción ayudaron a reforzar la detección, mientras que las variables numéricas tuvieron un papel complementario. Aunque no todas las características fueron empleadas en su forma original, la combinación final de las variables seleccionadas permitió obtener un modelo con un F1-Score competitivo y un umbral óptimo de clasificación ajustado automáticamente, lo cual refleja un buen desempeño para la tarea propuesta.