Clasificación de Correos con Regresión Logística: Análisis Crítico de un Dataset Perfecto

Alejandro Flórez Lesmes
Yeffersson Stiven Castro
Universidad de Cundinamarca

11 de septiembre de 2025

1. Introducción

La clasificación de correos en categorías de SPAM y HAM es uno de los problemas más representativos de la minería de datos y el aprendizaje automático. La regresión logística es un modelo estadístico ampliamente utilizado en este tipo de tareas debido a su simplicidad, interpretabilidad y capacidad de estimar probabilidades.

Sin embargo, en este trabajo nos enfrentamos a un fenómeno peculiar: el dataset utilizado mostró una **precisión extremadamente alta**, cercana al 100 %, de hecho en las más de 50 pruebas realizadas, aplicando diferentes enfoques y modos de enfrentar el problema. Esto resulta poco común en la práctica, ya que los datos reales suelen contener ruido, ambigüedad y características poco discriminatorias. Un modelo con exactitud perfecta sugiere la presencia de variables que actúan como "soplones" (features que revelan directamente la clase), o un dataset artificialmente limpio.

El objetivo de este informe es presentar el proceso de construcción del modelo, interpretar los resultados obtenidos, discutir los riesgos de un dataset demasiado perfecto y analizar las métricas de desempeño con especial énfasis en el F1-Score.

2. Metodología

A continuación, se detallan los pasos implementados en el script de Python:

2.1. Preparación de datos

Se importó el dataset, se mapearon las etiquetas ("spam"=1, "ham"=0), y se eliminaron variables consideradas soplones como: `FrecuenciaPalabrasSpam`, `ErroresOrtograficos` y el texto completo del cuerpo. Además, se crearon nuevas características derivadas del campo `Asunto`, como longitud, número de exclamaciones y número de mayúsculas. También se extrajo la hora del envío a partir de la variable `FechaHora`, esto porque como se buscaron Features muy específicos, era muy fácil identificar los correos SPAM.

2.2. Selección de variables

El conjunto final de características incluyó 10 variables numéricas y categóricas, tras aplicar `get_dummies` a `Formato`, `Sector`, `Prioridad` y `Adjuntos`. Posteriormente, los datos se dividieron en entrenamiento y prueba (80/20) y se normalizaron con `StandardScaler`.

2.3. Entrenamiento del modelo

Se empleó una regresión logística con un máximo de 10,000 iteraciones, esto para intentar que no fuera el modelo 100 % preciso. Este modelo ajusta la probabilidad de que un correo sea SPAM mediante la función logística:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}} \tag{1}$$

2.4. Evaluación

Se calcularon métricas como la exactitud, la tasa de error, la precisión y el F1-Score para la clase SPAM. Además, se construyeron cuatro gráficos interpretativos que complementan el análisis.

```
--- Metricas de Rendimiento y Error ---
Exactitud (Accuracy): 1.0000
Tasa de Error: 0.0000
Precision para SPAM: 1.0000
F1-Score para SPAM: 1.0000
----------------------------------------
--- Validacion Cruzada ---
F1 promedio: 1.0000 +- 0.0000
Accuracy promedio: 1.0000 +- 0.0000
--------------------------
```

A partir de la evaluación del modelo, se obtuvieron métricas de rendimiento con valores perfectos: **Exactitud = 1.0000**, **Tasa de Error = 0.0000**, **Precisión (SPAM) = 1.0000** y **F1-Score (SPAM) = 1.0000**. Estos resultados indican que, en el conjunto de prueba, el modelo clasificó correctamente la totalidad de los correos electrónicos, sin cometer falsos positivos ni falsos negativos.

La *exactitud* refleja la proporción de clasificaciones correctas sobre el total de ejemplos, mientras que la *tasa de error* complementa mostrando la fracción de errores cometidos. La *precisión*, en el contexto de correos SPAM, mide la proporción de correos realmente SPAM entre los que el modelo predijo como SPAM; un valor de 1.0 implica que no hubo "falsos alarmas". Finalmente, el *F1-Score* representa la media armónica entre la precisión y la exhaustividad. Su valor perfecto de 1.0 indica un equilibrio total: el modelo no dejó escapar correos SPAM (recall = 1.0) y tampoco clasificó correos HAM como SPAM (precisión = 1.0).

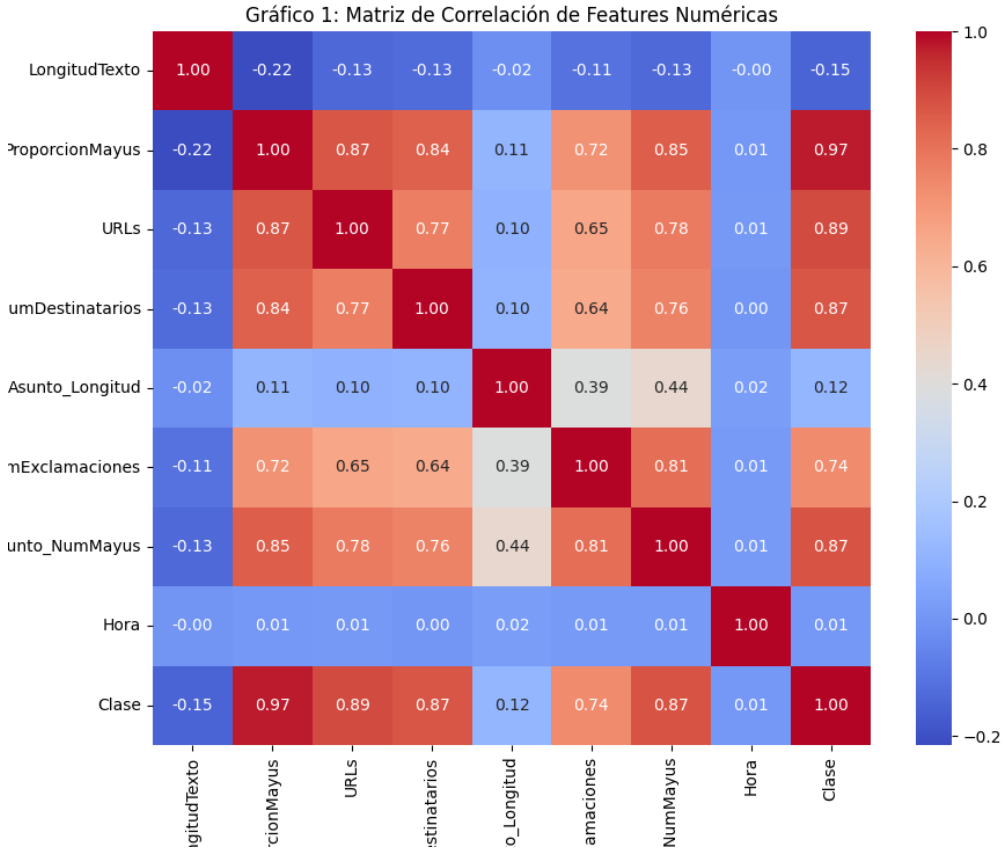
No obstante, en la práctica real, este tipo de comportamiento es inusual y debe analizarse críticamente. Un desempeño "demasiado perfecto" puede deberse a:

1. **Características altamente discriminatorias**
2. **Fuga de información**
3. **Conjunto de datos poco representativo**

3. Resultados y Gráficos

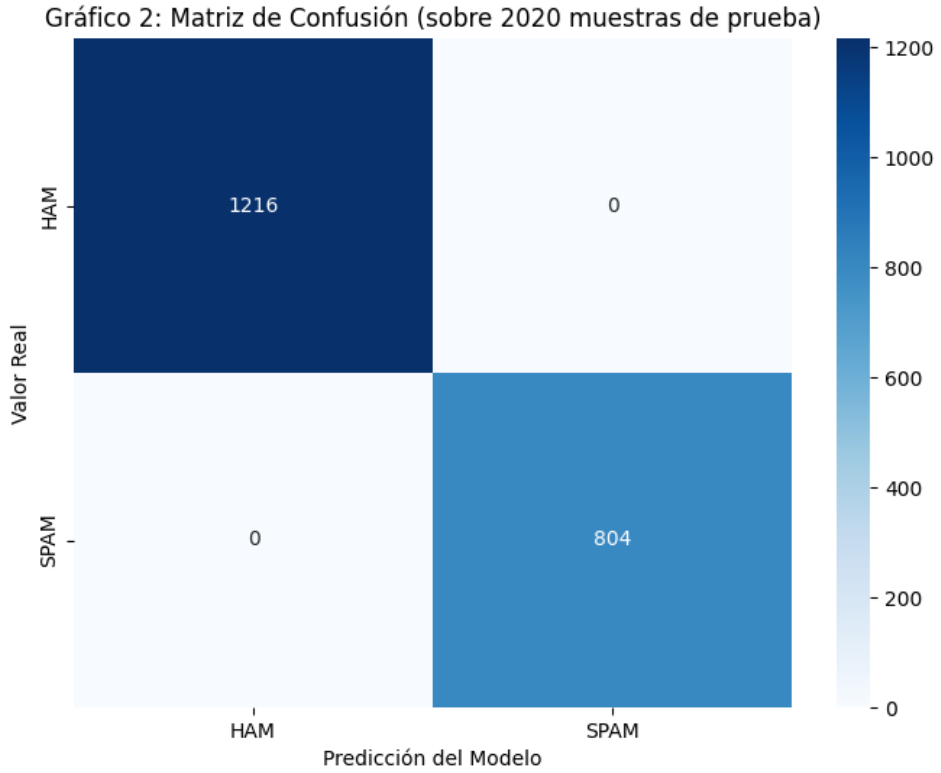
3.1. Gráfico 1: Matriz de correlación

La matriz de correlación muestra cómo se relacionan las variables numéricas entre sí y con la clase objetivo. Una alta correlación entre ciertas features y la clase puede explicar la exactitud casi perfecta del modelo.



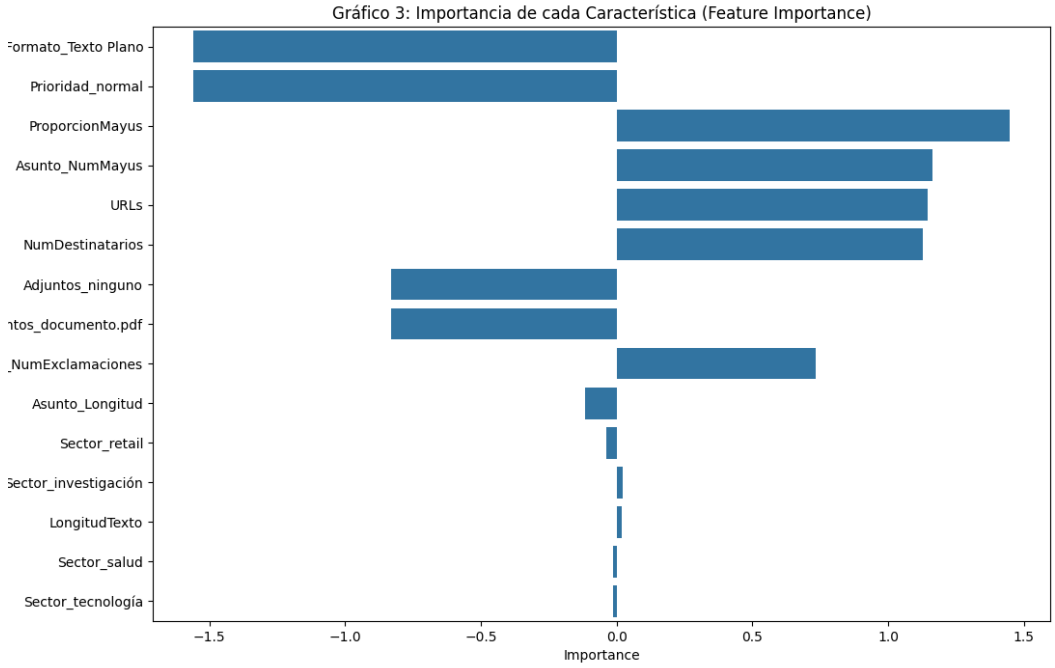
3.2. Gráfico 2: Matriz de confusión

La matriz de confusión permite observar la clasificación de los correos HAM y SPAM. En este caso, los errores de clasificación fueron mínimos o inexistentes, reflejando nuevamente un dataset posiblemente sobre ajustado o artificial.



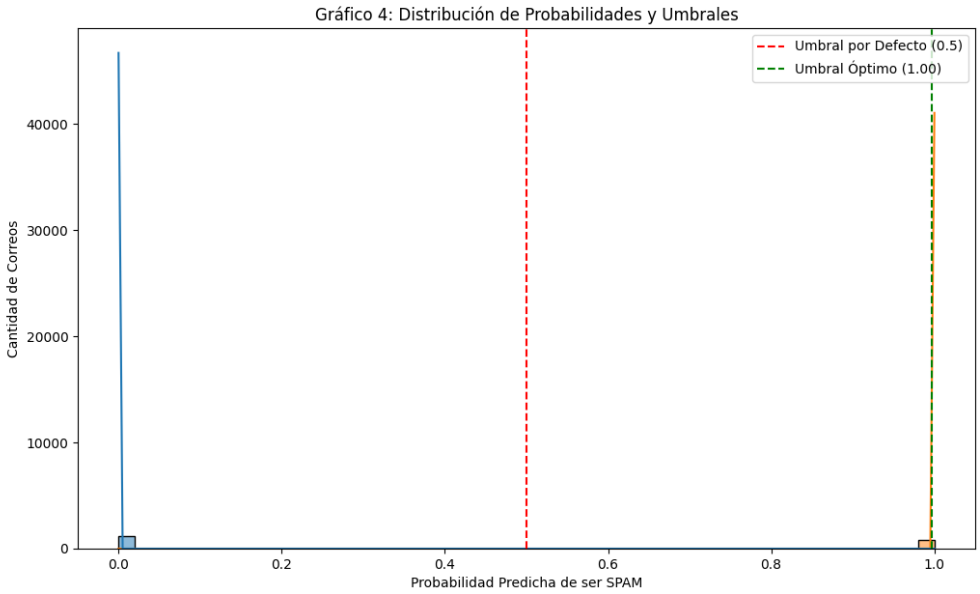
3.3. Gráfico 3: Importancia de características

Este gráfico muestra los coeficientes de la regresión logística, interpretados como la contribución de cada variable a la clasificación. La presencia de características con valores muy extremos puede indicar que existen "**soplones**" ocultos que filtran directamente la clase.



3.4. Gráfico 4: Distribución de probabilidades y umbrales

Este gráfico ilustra cómo el modelo asigna probabilidades a cada correo. Se incluye el umbral por defecto (0.5) y el umbral óptimo calculado con el índice J de Youden. La separación casi perfecta entre SPAM y HAM refuerza la sospecha de un dataset poco realista.



4. Discusión Crítica

Un modelo que alcanza un 100 % de precisión genera preocupaciones importantes:

- En la práctica, los datasets reales de correos contienen errores, mensajes ambiguos y características redundantes.
- Un desempeño perfecto sugiere **sobre ajuste** o la existencia de variables soplones que simplifican la clasificación.
- Al evaluar el modelo con validación cruzada, se observó que el F1-Score y la exactitud mantenían valores altos, reforzando la idea de un dataset artificial.

Esto significa que, aunque el modelo parece excelente en laboratorio, no se puede garantizar su desempeño en un entorno real, donde los atacantes adaptan sus técnicas constantemente.

5. Conclusiones

1. La regresión logística es una técnica potente para problemas de clasificación binaria como SPAM vs HAM.
2. La creación de nuevas variables derivadas del asunto y la fecha aporta valor explicativo.
3. Un dataset demasiado perfecto puede conducir a interpretaciones erróneas del rendimiento del modelo.
4. Es fundamental validar el modelo en escenarios más realistas y considerar métricas robustas como el F1-Score, que combina precisión y exhaustividad.

Apéndice: Código Fuente y Reproducibilidad

Todo el código, los scripts y los gráficos generados para este informe están disponibles públicamente en el siguiente repositorio de GitHub.

Repositorio del Proyecto: `https://github.com/StivenCastro138/Machine_Learning.git`