Clasificador de SPAM

Informe: Construcción de un Modelo de Detección de SPAM

Autor: Juan Fuentes

---

1. Introducción y Objetivo

El objetivo de este proyecto fue construir y evaluar un modelo de Machine Learning capaz de clasificar correos electrónicos como SPAM (correo no deseado) o HAM (correo normal). Para ello, se utilizó un dataset inicial de 1,000 registros y se empleó el algoritmo de Regresión Logística a través de la librería scikit-learn de Python.

El informe detalla el proceso de análisis de datos, la construcción del modelo, una serie de experimentos para evaluar su robustez y, finalmente, la creación de varios datasets sintéticos para futuras pruebas.

---

2. Análisis Inicial y Modelo Base

El primer paso consistió en analizar el dataset proporcionado (spam_ham_dataset.csv) para entender la relación entre sus 11 características predictoras y la variable objetivo (spam_label).

- **Análisis de Correlación:** Se generó una matriz de correlación para visualizar la relación entre las variables y se observó que no existían correlaciones lo suficientemente altas como para justificar la eliminación de alguna característica, por lo que se procedió a utilizar todas en el modelo inicial.
- **Construcción del Modelo:** Se entrenó un modelo de Regresión Logística con el 80% de los datos y se evaluó con el 20% restante. Los resultados fueron excepcionales:
- **F1-Score:** 1.00
- **Matriz de Confusión:** El modelo no cometió ningún error, clasificando correctamente todos los correos de la muestra de prueba.
- **Importancia de las Características:** Se analizó qué características eran más influyentes para el modelo. Los resultados mostraron que el número de

palabras en mayúscula, la cantidad de enlaces y el número de exclamaciones eran los predictores más potentes.

---

3. Serie de Experimentos y Pruebas de Estrés

Para comprender mejor el comportamiento y los límites del modelo, se realizaron tres pruebas iterativas:

Prueba 1: Modelo con las 5 Características Más Relevantes

Se entrenó un modelo más simple utilizando únicamente las 5 características más importantes.

- **Resultado:** El rendimiento se mantuvo perfecto (F1-Score de 1.00).
- **Conclusión:** Esto demostró que un subconjunto de características contenía información más que suficiente para lograr una clasificación precisa, dando a entender que un modelo mas simple es igual de eficaz.

Prueba 2: Modelo con Datos Normalizados

Se aplicó una normalización a todas las características para escalarlas a un rango más común (0 a 1). Esta es una práctica recomendada para asegurar que ninguna variable domine al modelo solo por tener una escala de valores mayor y así evitar los outliers.

- **Resultado:** El rendimiento se mantuvo perfecto (F1-Score de 1.00).
- **Conclusión:** El modelo es robusto y no es sensible a la escala de los datos de entrada, aunque la normalización puede ser una buena practica para otros proyectos.

Prueba 3: Modelo con las 2 Características Más Relevantes

Para llevar al modelo a su límite, se entrenó utilizando solo las dos características más fuertes (num_uppercase_words y num_links) con datos normalizados.

- **Resultado:** El rendimiento bajó ligeramente a un F1-Score de 0.9831. La matriz de confusión reveló 2 falsos negativos, es decir, dos correos SPAM que fueron clasificados como HAM.
- **Conclusión:** Este experimento fue clave para demostrar que, aunque dos características son muy potentes, no son suficientes para capturar todos

los patrones. Se necesita un conjunto más amplio de variables para manejar los casos más ambiguos y específicos y evitar errores.

---

4. Generación de un Dataset Sintético

Para facilitar pruebas futuras y entrenar el modelo con un volumen de datos mayor, se generó un dataset ficticio de 5,000 registros (fictitious_spam_ham_5000.csv). Este archivo fue creado simulando los patrones y distribuciones estadísticas observadas en el dataset original, para validar el rendimiento del algoritmo a mayor escala.

---

5. Conclusión General

El análisis y las pruebas iterativas demostraron que es posible construir un modelo de Regresión Logística muy preciso para la detección de SPAM con los datos proporcionados y se identificó que un conjunto de 5 características clave es suficiente para lograr una clasificación casi perfecta.