Nombres: Andres Julian Leon Montoya, Samuel Leonardo Tarazona Arciniegas

Materia: Introduccion a Machine Learning – 802

Docente: Alexander Espinosa Garcia

Actividad: Features propuestos y razones

longitud_mensaje: cantidad total de caracteres en el correo.

- Razón: correos SPAM suelen ser más largos o demasiado cortos con mensajes automatizados.

num_palabras_mayuscula: número de palabras escritas completamente en MAYÚSCULA.

- Razón: SPAM usa mayúsculas para llamar la atención (“OFERTA”, “GRATIS”).

num_signos_exclamacion: cantidad de “!” en el mensaje.

- Razón: SPAM abusa de signos de exclamación para persuadir.

num_links: número de enlaces dentro del correo.

- Razón: SPAM suele contener varios enlaces sospechosos.

num_palabras_dinero: cantidad de palabras relacionadas con dinero (“gratis”, “gana”, “premio”, “oferta”).

- Razón: los SPAM suelen enfocarse en temas económicos.

contiene_adjuntos: 0 (no), 1 (sí).

- Razón: muchos correos SPAM incluyen adjuntos maliciosos.

porcentaje_numeros: proporción de caracteres numéricos sobre el total.

- Razón: SPAM usa muchos números (descuentos, premios, códigos).

es_respuesta: 0 (no), 1 (sí).

- Razón: los HAM suelen ser respuestas en un hilo, los SPAM casi nunca.

num_palabras_raras: número de palabras poco frecuentes en lenguaje común.

- Razón: SPAM usa palabras poco comunes, a veces con errores ortográficos.

hora_envio: hora del día en que se envió el correo (0–23).

- Razón: algunos SPAM se envían en horarios extraños de manera masiva.

Tipos de predicciones que se pueden hacer con estos Features:

Tipos de predicciones que puedes hacer con estos features:

1. **Clasificación de correos (principal)**
   - Predecir si un correo nuevo es **SPAM (1)** o **HAM (0)** a partir de los 10 features.
   - Ejemplo: te llega un correo con 3 links, 7 palabras en mayúsculas, muchos signos de exclamación → el modelo predice **SPAM**.

2. **Probabilidad de SPAM**
   - En lugar de dar solo SPAM/HAM, algunos modelos (como *Regresión Logística, Random Forest, XGBoost*) permiten predecir la **probabilidad de que un correo sea SPAM**.
   - Ejemplo: "Este correo tiene un **82% de probabilidad de ser SPAM**".

3. **Análisis de importancia de variables**
   - Puedes entrenar un modelo (ejemplo: Árboles de decisión, Random Forest) y obtener qué **features son más importantes** para detectar SPAM.
   - Ejemplo: descubrir que num_links y num_palabras_dinero son más determinantes que hora_envio.

4. **Filtrado automático en producción (uso práctico)**
   - Si integras este modelo, puedes crear un **filtro automático de correos**:
     - Bandeja de entrada → HAM
     - Bandeja de spam → SPAM

5. **Predicciones exploratorias adicionales (no obligatorias, pero posibles):**
   - **Predicción de hora típica de envío**: Dado un conjunto de correos SPAM, el modelo podría aprender qué horas son más frecuentes.
   - **Segmentación por patrones**: Con *clustering* podrías agrupar tipos de SPAM (ej. "ofertas comerciales", "fraudes financieros", "publicidad de casinos").