**UNIVERSIDAD DE CUNDINAMARCA**

**FACULTAD DE INGENIERÍA**

**Justificación de Features**

**801**

**Francisco Antonio Vargas**

**Juan Manuel Pastor**

**Estudiantes**

**Machine Learning**

**Alexander Espinosa**

**Docente**

**Facatativá**

**30/08/2025**

**Justificación de Features**

A partir de la necesidad de una muestra para realizar el ejercicio propuesto en clase se generó por inteligencia artificial un archivo csv, el cual contiene 1000 instancias de correos con las siguientes 4 columnas de datos: La primera es un id o identificador, la segunda es el asunto del correo, la tercera el contenido del correo, y por último la dirección de correo del remitente.

Con este archivo se determinaron las siguientes features o características esenciales para considerar estos correos como spam o como ham, es decir, si pertenecen a los correos no deseados o a aquellos que si son legítimos.

**Features**

1. **Número de enlaces**

Primero se determinará el número de enlaces que posee el contenido de los correos, debido a que la mayoría de spam o correos no deseados, suelen estar llenos de enlaces a paginas fraudulentas, de phishing o publicidad. En ham los enlaces no son frecuentes, y en el caso de que se encuentren suelen ser a sitios confiables.

1. **Si menciona “oferta” o “gratis”**

En la mayoría de los casos las palabras oferta, promoción, etc, se encuentran en publicidad no deseada ya que este tipo de correos busca atraer la atención. En un correo deseado es muy improbable o nula la posibilidad de que se encuentre la palabra oferta o gratis, siendo este feature clave en la clasificación entre spam y ham.

1. **Cantidad de signos de exclamación**

Los corres spam al querer atraer la atención del publico suelen usar signos de exclamación en como muestra de urgencia o hasta de emoción, por ende, es un fuerte indicador de spam cuando se detectan varios en un mismo correo.

1. **Si el remitente es conocido**

En el caso de que algún correo cuente con varios signos que podrían señalarlo como spam, muchas veces la muestra mas clara es la dirección del corre del remitente debido a que los correos ham provienen de contactos habituales o dominios confiables.

1. **Cantidad de palabras**

Con la cantidad de palabras se pueden filtrar fácilmente los correos ham del spam, debido a que los spams suelen tener muy poco texto o por el contrario mucho pero repetitivo, así mismo los ham tienen una extensión intermedia y con un propósito claro.

1. **Cantidad de imágenes**

En el spam suelen hallarse múltiples imágenes o banners llamativos con el propósito de captar la atención de receptor, sin embargo, un correo ham no tiene la necesidad de hacer esto y suele ser texto plano pero informativo.

1. **Archivos adjuntos**

Dependiendo del tipo del archivo adjunto se puede sospechar de si es spam o no, ya que, en la mayoría de los casos si tiene un archivo adjunto suelen ser documentos en .zip, .exe, .scr, los cuales no son comunes, mientras que los ham por el contrario incluyen. docs, .pdf, etc. Lo que permite detectar malware disfrazado.

1. **Palabras completamente en mayúsculas**

Con el fin de atraer los spams usan la mayúscula sostenida en múltiples ocasiones, siendo así sencillos de reconocer puesto que en un correo ham el uso de las mayúsculas es mucho menor.

1. **Contiene signos de $ o %**

Spam usa mucho símbolos monetarios y porcentajes ya sea para ofertar o para engañar, mientras que en ham la frecuencia es menor, no obstante, si se pueden ver este tipo de signos en correos ham, pero junto con otros de los features mencionados esta revisión ayuda a comprobar la legitimidad de un correo.

1. **Uso de emoticones**

El spam puede usar emoticones para generar falsa confianza o llamar la atención. En ham, su uso es más común en contextos personales, pero no en correos laborales o formales, por lo que su presencia, especialmente en exceso, puede ser una señal de spam