

**Análisis de métodos de razonamiento e incertidumbre** (Grupo. 301)

Detección de Spam en correos con Naive Bayes

**Docente:**

Daniel Otero Fadul

**Alumnos:**

Ricardo Kaleb Flores Alfonso | A01198716

Iván Alejandro López Valenzuela | A01284875

David Alejandro Acuña Orozco | A00571187

Raúl Correa Ocañas | A01722401

Carol Jatziry Rendon Guerrero | A01425341

# **1. Problematización**

En un mundo inter y multi conectado, se ha vuelto relevante la optimización del tiempo que se invierte en las tareas del día a día, así como mantenerse seguro de los posibles ataques que existen para robar información. A diferencia de épocas anteriores, el correo ha migrado al medio digital de manera que se puedan recibir todas las notificaciones al instante.

Sin embargo, esto abre la posibilidad de recibir correos electrónicos “basura” o de remitentes desconocidos para obtener nuestra información de manera ingeniosa pero a la vez engañosa. De esta manera, se vuelve crucial el emplear métodos de detección de correos “basura” de manera automática, lo que podría reducir la cantidad de tiempo dedicado a la revisión de la bandeja de correos, así como brindar protección contra posibles ataques de *phishing* u otros relacionados con la recepción de correos spam.

# **2. Enfoque**

Para poder clasificar de manera automática estos correos se usará un clasificador de Bayes Ingenuo. Este modelo permite clasificar de manera rápida y con una baja complejidad computacional un gran número de muestras debido a que es un modelo supervisado que en este escenario permite estimar la probabilidad de que un mensaje pertenezca a una clase, ya sea spam o no spam, basándose en las palabras que contiene.

Es importante considerar que el modelo asume independencia entre cada una de las palabras. Sin embargo, el lenguaje tiene una estructura, por lo que la probabilidad de que aparezca cierta palabra después de otra sí tiene relación. Es por eso que el modelo se llama Bayes Ingenuo, debido a que para este caso la suposición de independencia no impide buenos resultados.

La clasificación se logra a través del cálculo de la probabilidad condicional de que un mensaje sea spam o no spam, dependiendo de las palabras que contiene. Estas probabilidades se calculan mediante la frecuencia en la que las palabras aparecen en los datos de entrenamiento.

**3. Propósito**

En el presente trabajo se busca evaluar el modelo de Bayes Ingenuo e implementar algunas mejoras de manera que este clasifique de mejor manera los correos spam, cuidando de igual manera la correcta clasificación de los correos no spam donde se evite la pérdida de comunicaciones importantes.

**4. Información**

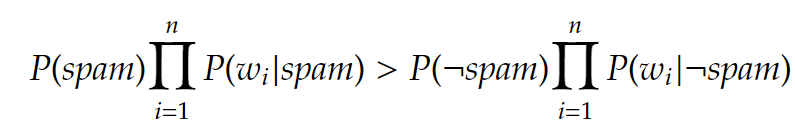
El conjunto de datos utilizado, titulado "SMS Spam Collection Dataset", proviene del repositorio de aprendizaje automático de la Universidad de California en Irvine (UCI) y está disponible en la plataforma de Kaggle. Este dataset consta de un archivo CSV que incluye 5,574 mensajes SMS en inglés, de los cuales 747 están etiquetados como spam, mientras que el resto se clasifica como "ham" (no spam). El archivo original contiene cinco columnas; sin embargo, el análisis se centra únicamente en las dos primeras: una etiqueta de clasificación (spam o ham) y el contenido del mensaje SMS. Se puede acceder al conjunto de datos original en [este enlace](https://archive.ics.uci.edu/dataset/228/sms+spam+collection).

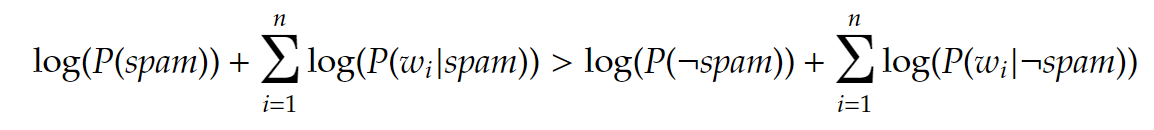
# **5. Razonamiento**

Se comenzó con la limpieza de los datos, donde se eliminaron columnas no relevantes y se renombraron las variables para mejor claridad. En seguida, se llevó a cabo un preprocesamiento de texto. Este paso es especialmente esencial para el correcto manejo de datos textuales. Algunas de las modificaciones principales consistieron en:

* La conversión de texto a minúsculas para dar consistencia.
* La eliminación de caracteres especiales mediante Regex, ya que estos pueden ser ruido para el modelo.
* La eliminación de stopwords, las cuales son palabras comunes que no aportan valor semántico significativo para la correcta clasificación.
* La aplicación de stemming y lematización, técnicas que reducen las palabras a sus raíces etimológicas o conjugaciones base. Esto reduce el procesamiento necesario para procesar la importancia de las palabras.

Después de este preprocesamiento, se aplica el método de “bag-of-words”, el cual consiste en recolectar la frecuencia de cada palabra presente en la muestra de textos. Con el conteo de palabras, se puede hacer un ajuste para convertir estas frecuencias en probabilidades de ocurrencia para cada palabra. Teniendo dos diccionarios de probabilidades, uno para palabras encontradas en mensajes spam y otro para palabras encontradas en mensajes ham, se puede obtener las probabilidades de cada etiqueta dado el conjunto de palabras del mensaje. Fundamentado en el teorema de Bayes, se determina si un mensaje es o no es spam de acuerdo al siguiente criterio:





Habiendo implementado estos métodos para llevar a cabo la clasificación de la información presente en el conjunto de datos, se obtuvieron como resultados relevantes la siguiente matriz de confusión y métricas.

|  | Spam predichos | No spam predichos |
| --- | --- | --- |
| Spam reales | 108 | 36 |
| No spam reales | 173 | 797 |

Accuracy: 0.812388

Precision: 0.384342

Recall: 0.750000

F1 Score: 0.508235

Estos resultados dan a entender que si bien existe un 81% de acierto en la clasificación general, al existir desbalance en las clases originales, se debe dar también importancia a lo que tienen que decir el resto de las métricas. En primer lugar, se identifica que de todos los mensajes clasificados como spam, solo el 38.43% realmente lo son. Aunque esta métrica parezca poco alentadora en el aspecto de la pérdida de comunicaciones importantes, en realidad influye dicho desbalance debido a que los no spam igual son clasificados correctamente en un 82.16%, aunque la necesidad de incrementar este porcentaje y que se minimicen los falsos positivos (spam) persiste dado el volumen de los datos para esta clase.

En cuanto a cómo se clasifican los mensajes spam reales y de manera que se evite su infiltración a la bandeja no spam, se tiene un 75% de desempeño favorable en este rubro, el cual es considerado aceptable en la prevención y protección contra este tipo de amenazas. De igual forma, el seguimiento de las recomendaciones de seguridad para evitar el robo y pérdida de información se mantiene indispensable porque como esta métrica lo refleja, el clasificador para este caso no es infalible.

El F1-Score de 0.5082 indica un equilibrio entre precision y recall en el que hay dificultades para lograr tanto una alta precisión como un buen recall, por lo que el modelo es adecuado para clasificar correctamente algunos mensajes de spam, pero no es completamente confiable. Dado el desbalance de clases, el modelo tiende a clasificar muchos mensajes no spam correctamente, pero su rendimiento se ve afectado por la cantidad de falsos positivos y falsos negativos. Mejorar el F1-Score puede implicar ajustar el umbral de clasificación o abordar el desbalance de clases ya que 0.5082 es relativamente bajo.

# **6. Conclusiones**

Tomando en cuenta lo anterior y en el entendimiento del propósito del trabajo, se puede concluir que el modelo implementado para la detección de mensajes SMS ha demostrado un desempeño satisfactorio, especialmente en entornos donde se requiere una clasificación rápida y de baja complejidad. Aunque presenta ciertas limitaciones en la identificación de correos no spam como spam, su efectividad en la clasificación de mensajes y su simplicidad lo convierten en una herramienta valiosa para esta problemática. Para mejorar su rendimiento en el futuro sería recomendable combinar este enfoque con técnicas más avanzadas para aumentar su precisión y capacidad de generalización, junto a su adaptación dadas las estrategias renovadas que puedan emerger por parte de los atacantes.

# **A. Anexos**

<https://drive.google.com/file/d/1CCGUzTrhEkaXyEhGiSJprRoAvvNhaVra/view?usp=drive_link>