UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL PARA CIENCIA E INGENIERÍA
Docente: Raul Ramos Pollan

Grupo: *MJ 10 - 12*

Integrantes:	**Cédula:**
Daniel Esteban Maya Portillo	11004540273
Sebastian Solorzano Betancur	1152467499
Manuela Gutiérrez Cano	1037657256

ENTREGA No. 1

1. Ante la problemática que se tiene de acuerdo a los correos electrónicos y mensajes no deseados denominados SPAM, se busca categorizar y predecir si el tipo de correo o mensaje pertenece a este grupo que presentan frecuencias de palabras como características, las cuales se utilizarán para el entrenamiento del método de aprendizaje supervisado.

2. El dataset que se tendrá en cuenta para este tipo de clasificación se ha tomado de la plataforma Kaggle (https://www.kaggle.com/datasets/colormap/spambase) que presenta una base de datos de 4601 de muestras y cuyos atributos son 58, de los cuales se consideran relevantes los siguientes:
    a. **Word_freq_address :** Porcentaje de palabras en el correo electrónico que coinciden con la dirección
    b. **charfreq#:** Porcentaje de caracteres en el correo electrónico que coinciden con el número.
    c. **capital_run_length_average:** Promedio de secuencias ininterrumpidas de letras mayúsculas.
    d. **capital_run_length_longest:** Longitud de la secuencia ininterrumpida más larga de letras mayúsculas.
    e. **capital_run_length_total:** Número total de letras mayúsculas en el correo electrónico.

    Hay que tener en cuenta que para que el dataset cumpla con los requisitos del proyecto se completa el número de muestras con datos faltantes, es decir, 400 datos faltantes en el dataset, y además se crean las variables categóricas haciendo uso de la librería pandas.

3. Como método de clasificación se plantea utilizar el método de Naive Bayes, el cual es muy popular para la categorización de texto, utilizando la frecuencia de palabras en cada documento en el correo electrónico para determinar si es SPAM o no lo es. Naive Bayes asume el efecto de una característica en particular como independiente de otras, ejemplo, las características mencionadas en el numeral 2 se consideran de

forma independiente, lo que simplifica la computación y además permite que el algoritmo se entrene con menos datos que los demás.
Para las métricas de desempeño del algoritmo se pretenden utilizar aquellas que permitan el mejor desempeño, las cuales son:

- **Accuracy :** Es la relación entre el número de predicciones correctas y el número total de predicciones.

- **Matriz de confusión:** Es un cuadro que registra el número de predicciones de un conjunto de datos que pertenecen a una categoría determinada La etiqueta de clase en un conjunto de datos binarios puede tomar dos posibles valores, que se denominan clase positiva y clase negativa. Como se ve en la figura 1, el número de instancias positivas y negativas que un clasificador predice correctamente se denomina Verdadero Positivo (VP) y Verdadero Negativo (VN) respectivamente. Las instancias mal clasificadas se conocen como Falsos Positivos (FP) y Falsos Negativos (FN).

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 1. Matriz de confusión

- **Precisión y recall:** Precisión permite medir la calidad del modelo de clasificación.

$$Precisión = \frac{VP}{VP + FP}$$

Recall es una métrica sobre la cantidad que el modelo es capaz de identificar.

$$Recall = \frac{VP}{VP + FN}$$

- **F1 score:** Es una métrica de la precisión de la prueba. Tiene en cuenta la precisión y el recall para calcular su puntuación

$$F1score = \frac{Precisión*Recall}{Precision + Recall}$$

4. Un primer criterio para evaluar el desempeño del algoritmo sería en primera instancia la precisión en la clasificación de correos electrónicos, buscándose que el accuracy sea mayor al 90%.