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	1004540273
Manuela Gutiérrez Cano	1037657256
Nilson Suarez Hernandez	1038124979

ENTREGA No. 2

1. Preprocesamiento de los datos

El conjunto de datos que se utiliza corresponde a un problema de clasificación de tipos de correo en spam y no spam o en ham y no ham, el cual tiene inicialmente 4601 muestras y 57 características y la variable objeto o de salida, con dos clases que corresponden a 0 y 1, en donde 0 corresponde a los correos que no son spam y 1 a los correos que son spam. Cada uno de estos tipos de correos posee una serie de características de acuerdo a su tipo, a cada correo se le evalúan las mismas características y así mismo tiene un valor distinto en cada una de ellas. Este problema pertenece a un problema supervisado, ya que se conoce de antemano la variable de salida que es la que se quiere predecir.

El conjunto de datos que se pretende obtener tiene las características que son: al menos 5000 muestras, el 10% de las características ha de ser de tipo categórico, el 5% de los datos en al menos 3 columnas deben ser datos faltantes, para ello se crea un dataframe con las muestras faltantes, con la finalidad de que este sea rellenado con datos vacíos o datos faltantes. Algunas de las características del conjunto de datos se codifican como datos o variables categóricas, es decir, en este caso las variables que se simularon para ser las categóricas fueron *char_freq_(, char_freq_[, char_freq_!, char_freq_#,* y *char_freq_$,* que se clasificaron en tres categorías, que fueron *Baja, Media y Alta* dependiendo de la característica de una muestra en específico, para esto se dividió en tres intervalos cada uno de los valores o datos de las muestras de éstas características.

Luego de haber realizado lo anterior, se dispuso a empezar con el preprocesamiento de los datos, el cual corresponde a dejar un conjunto de datos limpio con el fin de que sirva como entrada a los modelos de Machine Learning que se utilizarán posteriormente para resolver el problema de clasificación, esto es, un conjunto de datos sin valores faltantes, con una buena cantidad de muestras y sin variables de tipo categórico, porque los modelos sólo aceptan datos numéricos, cuantitativos de tipo discreto o continuo. El tipo de problema que se va abordar es de clasificación porque la variable objetivo es de tipo discreta.

Las variables de tipo categórica se codificaron por medio de la codificación One Hot Encoding, esto es que dichas variables quedaron representadas de acuerdo a su categoría

como un 1 si la variable está presente y un 0 si la variable está ausente, y para los datos faltantes se realizó una imputación de valores con la función *KNNImputer* de la librería de *Sklearn* de Python, se utilizó esta función, debido a que permite hacer una imputación de valores no aleatorios que tienen como principio de funcionamiento el algoritmo de *KNN* (k-Nearest-Neighbours), que corresponde a los *k* vecinos más cercanos del dato en particular.

2. Exploración de los datos

Para realizar la exploración de los datos y lograr comprender cómo es el comportamiento y la distribución de los datos y tener una visión cercana de cuáles de los algoritmos de Machine Learning se deben entrenar para conseguir predecir dicho comportamiento con muestras futuras, se visualizó que el conjunto de los datos preprocesado tiene 5000 muestras, 67 características y 1 variable de salida, 2 clases: 0 y 1 y un número total de 2988 muestras para la clase 0 y 2012 muestras para la clase 1. Esto evidencia que no hay un desbalance significativo entre la cantidad de muestras de una clase que de otra, ya que el problema del desbalance es serio, porque la clase con mayor número de muestras o mayoritaria puede sesgar los resultados y la clase minoritaria se ve mal representada y esto no conduce a realizar correctas predicciones.

También se realizó una exploración gráfica por medio de un gráfico de barras que muestra cómo es la distribución de las muestras por clase como se muestra en la figura 1, y se hizo un gráfico de la matriz de correlación entre variables que mide el grado de dependencia entre las mismas, en donde se observa que el grado de dependencia no es tan alto, tal y como se muestra en la figura 2, lo que permite ahondar más sobre este problema.

Figura 1. Número de muestras por clase

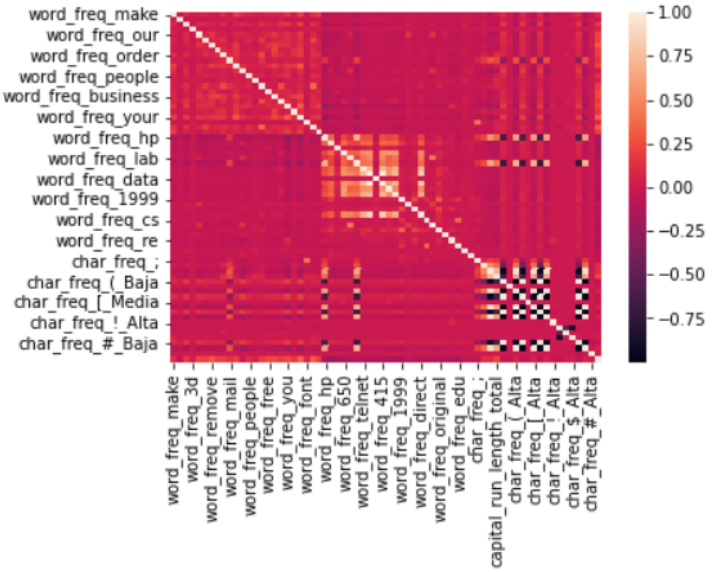

Figura 2. Matriz de correlación entre variables

Para este problema de clasificación se considerará el uso de metodologías de validación en el entrenamiento de los modelos, por supuesto, metodologías para problemas balanceados como la de validación cruzada (*K-fold)*, y también se les aplicará la normalización a los datos antes de entrenar los modelos.

3. Modelos para resolver el problema

Como se mencionó anteriormente, es importante entrenar varios modelos de Machine Learning con el fin de saber cuál es el que mejor se adapta o se ajusta a los datos del problema, puesto que hay modelos que se logran ajustar mejor que otros, dependiendo también de los hiper parámetros que tenga cada uno y de la cantidad de iteraciones que se realice para lograrlo.

Con el fin de realizar un adecuado entrenamiento y validación de los datos y que los modelos que se entrenen puedan generalizar con muestras nuevas, éstos se dividieron en conjunto de entrenamiento y conjunto de prueba, con un porcentaje inicial del 70% o 0.7 para el para el conjunto de entrenamiento y un 30% o 0.3 para el de prueba, por medio de la función *train_test_split* de *Sklearn.*

Un primer acercamiento al comportamiento de los datos se realizó a través del modelo de la *regresión logística* que se entrenó con un parámetro *solver* '*liblinear*' y un máximo de iteraciones de 10000, o *max_iter.* Con esto se encontró que el *accuracy* fue de aproximadamente el 89% o 0.89, lo que es relativamente bueno para el problema de clasificación que se está resolviendo, ya que se pretende que con el entrenamiento de un algoritmo o de un clasificador se alcance el 90% de accuracy o incluso más, no obstante esta es una primera suposición para un primer modelo evaluado, ya que como se ha mencionado anteriormente se busca obtener el modelo que mayor porcentaje de *accuracy* presente, esto para obtener una buena métrica de desempeño y a su vez también permita obtener una buena métrica de negocio.