Aprendizaje Automático
TP2: Algoritmos de Clasificación Supervisada

Profesores: J. Gambini - J. Santos

1. El 15 de abril de 1912, durante su viaje inaugural, el Titanic se hundió después de chocar con un iceberg, matando a miles de personas. Esta tragedia sensacional conmocionó a la comunidad internacional y condujo a mejores normas de seguridad aplicables a los buques. Una de las razones por las que el naufragio dio lugar a semejante cantidad de muertes fue que no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque hubo algún elemento de suerte involucrada en sobrevivir al hundimiento, algunos grupos de personas tenían más probabilidades de sobrevivir que otros, como las mujeres, los niños y la clase alta.

   El dataset **Titanic** proporciona información sobre el destino de los pasajeros en el viaje fatal del trasatlántico Titanic, que se resume de acuerdo con el nivel económico (clase), el sexo, la edad y la supervivencia.

   a) Dividir el conjunto de datos en dos partes, el conjunto de entrenamiento y el conjunto de prueba.

   b) Implementar el algoritmo ID3 para clasificar los datos y poder determinar si una persona sobrevivió o no, utilizando todas las variables y la entropía de Shannon para la función Ganancia.

   c) Implementar el algoritmo ID3 para clasificar los datos y determinar si una persona sobrevivió o no, utilizando todas las variables y el índice de Gini.

   d) Clasificar los datos para determinar si una persona sobrevivió o no, utilizando el método de Random Forest utilizando todas las variables.

   e) Construir la matriz de confusión para todos los métodos utilizando el conjunto de prueba. Compare los resultados.

   f) Realizar el gráfico de curvas de precisión del árbol en función de la cantidad de nodos para cada caso. Esto es, para cada método, graficar la precisión en función de la cantidad de nodos para el conjuto de entrenamiento y para el conjunto de prueba.

2. El archivo *reviews_sentiment.csv* contiene 257 registros con opiniones de usuarios sobre una aplicación. Variables:

- Review Title es el título del comentario.
- Review Text es el comentario.
- wordcount: cantidad de palabras utilizadas.
- Title sentiment: Valoración en positiva (asignar 1) o negativa (asignar 0) estimada y puede ser NaN.
- text sentiment: Valoración positiva o negativa, provista por la persona que dejó el comentario.
- sentimentValue: valor real entre -4 y 4 que indica si el comentario fue valorado como positivo o negativo.
- Star Raiting: estrellas que dieron los usuarios a la aplicación. Son valores discretos del 1 al 5.

a) Los comentarios valorados con 1 estrella, ¿qué cantidad promedio de palabras tienen?

b) Dividir el conjunto de datos en un conjunto de entrenamiento y otro de prueba.

c) Aplicar el algoritmo K-NN y K-NN con distancias pesadas para clasificar las opiniones, utilizando como variable objetivo la variable Stars Raitng y como variables explicativas las variables numéricas: wordcount, Title sentiment, sentimentValue y con $k = 5$.

d) Calcular la precisión del clasificador y la matriz de confusión.