

# Aprendizaje Automático

## TP2: Algoritmos de Clasificación Supervisada

Profesores: J. Gambini - J. Santos

1. Una persona disfruta de deportes en el río si se cumplen algunas condiciones sobre el clima. Teniendo la siguiente tabla de verdad:

Ejemplo	Cielo	Tempe.	Humedad	Viento	Agua	Pronóst	¿Disfruta?
1	sol	Cálida	Normal	Fuerte	Cálida	Estable	SI
2	sol	Cálida	Alta	Fuerte	Cálida	Estable	SI
3	Nublado	Frío	Alta	Fuerte	Cálida	Cambiante	NO
4	Sol	Cálida	Alta	Fuerte	Fría	Cambiante	SI

- a) Implementar un programa que construya el árbol de decisión para la función objetivo **Disfruta el Deporte** utilizando la función de Ganancia con la entropía de Shannon.
- b) Agregar el siguiente ejemplo de entrenamiento y encontrar nuevamente el árbol de decisión.

Ejemplo	Cielo	Tempe.	Humedad	Viento	Agua	Pronóst	Disfruta?
5	sol	Cálida	Normal	Débil	Cálida	Estable	NO

2. 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) Clasificar los datos para determinar si una persona sobrevivió o no, utilizando todas las variables y la entropía de Shannon para la función Ganancia.
  - c) Clasificar los datos para 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 para los dos ejercicios anteriores.
  - e) Construir la matriz de confusión para todos los métodos utilizando el conjunto de prueba.
  - f) Realizar el gráfico de curvas de la precisión del árbol en función de la cantidad de nodos para cada caso.
3. 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 Rating: 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 Rating 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.