Aprendizaje Automático TP2:
Algoritmos de Clasificación Supervisada

1. El conjunto de datos *german credit.csv* contiene información sobre personas que solicitaron créditos bancarios en bancos alemanes en el año 1994. Contiene 20 variables:

- **creditability:** si devolvió el crédito (1) o no (0)
- **account.balance:** toma valores 1,2,3,4 e indica que no tiene cuenta(1), que tiene una cuenta poco balanceada (2) o bien balanceada (4).
- **duration.of.credit..month.** (en meses).
- **payment.status.of.previous.credit:** toma valores de 0 a 4, 0 no pagó, 4 pagó todo.
- **purpose**: Toma valores de 0 a 10, indicando el objeto que el cliente desea comprar, por ejemplo 0 es un auto.
- **credit.amount** : variable numérica con el crédito solicitado.
- **Svalue.savings.stocks:** dinero ahorrado, toma valores de 1 a 5, 1 = nada, 2, ≤100, 3, (100, 500], 4 (500, 1000].
- **length.of.current.employment:** Toma valores de 1 a 5, desempleado (1), < 1 año (2), [1, 4) años (3), [4, 7) años (4), más de 7 añs (5).
- **instalment.per.cent:** toma valores 1 a 4, donde 1 indica que financia más del 35 % del crédito, 2: (25 %, 35 %), 3: [20 %, 25 %), 4: menos del 20 %.
- **sex...marital.status:** valores de 1 a 4, 1: Male, Divorced, 2: Male, Single, 3: Male, Married/Widowed, 4: Female.
- **guarantors:** toma valores de 1 a 3: Ninguno, garantía por un co-solicitante, garantía.
- **duration.in.current.address valores:** de 1 a 4.
- most.valuable.available.asset: 1 a 4 que son: None, Car, Life Insurance, Real Estate.
- **age..years:** edad variable numérica.
- **concurrent.credits:** valores de 1 a 3, en otro banco, en otra entidad financiera, no tiene.
- **type.of.apartment:** valores de 1 a 3 que son :Free, Rented, Owned.
- **no.of.credits.at.this.bank:** valores de 1 a 4.
- **occupation:** valores de 1 a 4, Unemployed, Unskilled Permanent Resident, Skilled, Executive.
- **no.of.dependents** : toma valores 1 y 2, más de 3 propiedades o menos de 3. telephone: toma valores 1 o 2, (sí o no)
- **foreign.worker**: 1, 2 si o no

a) Dividir el conjunto de datos aleatoriamente 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 devolverá el crédito o no, utilizando todas las variables y la entropía de Shannon para la función Ganancia.

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

d) Construir la matriz de confusión para cada método utilizando el conjunto de pruebas. Compare los resultados.

e) 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 conjunto 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 Tex**t 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 dejo 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 los algoritmos 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 diferentes valores de k.

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