# Trabajo Práctico 2 : NLP + Árboles de Decisión

El trabajo debe ser resuelto en una máquina virtual tipo júpiter en el sitio Kaggle, se debe entregar el enlace a la máquina correspondiente con los permisos de lectura y escritura necesarios para la evaluación.

Fecha máxima de entrega: Lunes 30 de mayo.

## Parte 1 - Análisis de sentimientos

En esta primera parte se trabajará sobre un conjunto de datos de comercio electrónico de ropa de mujer que contiene reseñas escritas por los clientes ( link ). El objetivo es realizar un análisis de sentimientos para clasificar las reseñas como positivas o negativas.

1. **Exploración, preprocesamiento y transformación de datos**

   a. Realizar Exploración de datos, describiendo las características de los mismos.
   b. Realizar las tareas de limpieza y transformación de datos que sean necesarias.
   c. Construir la variable objetivo considerando al atributo Rating de la siguiente forma:
      - Negativo: 1-2-3
      - Positivo: 4-5

2. **Generación y evaluación de modelos**

   a. Dividir el conjunto de datos en un 70-30, en donde el 70% de los datos se utilizarán para entrenar el modelo y el 30% restante para validarlo.

   b. Entrenar los siguientes algoritmos tal que a partir del texto en el campo “Review Text”, pueda clasificar correctamente la crítica como positiva o negativa.

      - Naive Bayes
      - Regresión logística
      - Árboles de decisión
      - Random Forest

   c. Evaluar todos los clasificadores utilizando las métricas, precisión, recall y F1-Score.
   d. Seleccionar el modelo con mejor desempeño y evaluarlo utilizando las 5 clases del atributo Rating en lugar de la clase binaria

3. **Conclusiones**

   Extraer conclusiones a partir de los análisis realizados en los puntos anteriores, y justificar cada conclusión.

## Parte 2 - Árboles de Decisión

En esta segunda parte se trabajará sobre un conjunto de datos de reservas de hotel ( link ). El objetivo es predecir cuál de ellas va a ser cancelada. Para resolver este problema se utilizarán árboles de decisión.

### 1. Exploración, preprocesamiento y transformación de datos

a. Describir los atributos realizando una breve explicación de qué representan y del tipo de variable (categórica, numérica u ordinal). En caso de que haya variables no numéricas, reportar los posibles valores que toman y cuán frecuentemente lo hacen.
b. Reportar si hay valores faltantes. ¿Cuántos son y en qué atributos se encuentran? En caso de haberlos, ¿es necesario y posible asignarles un valor?
c. ¿Qué variables se correlacionan más con la cancelación de la reserva? Para las cuatro más correlacionadas, mostrar un scatter plot en el que el eje x corresponde a la variable correlacionada, y el eje y a la cancelación.
d. Realizar las tareas de limpieza y transformación de datos que sean necesarias.

### 2. Generación y evaluación de modelos

En primer lugar, se deberá separar un 20% de los datos para usarlos como conjunto de evaluación (test set). El conjunto restante (80%) será el de entrenamiento.

a. Construir un árbol de decisión y optimizar sus hiperparámetros mediante kfold-Cross Validation para obtener la mejor performance.¿Cuántos folds utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
b. Graficar el árbol de decisión con mejor performance encontrado en el punto anterior.
c. Analizar el árbol de decisión seleccionado describiendo los atributos elegidos, y decisiones evaluadas.
d. Evaluar la performance del árbol en el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión. Comparar con la performance de entrenamiento.
e. Entrenar un modelo Random Forest y evaluar su performance sobre los conjuntos de entrenamiento y test. Comparar con el árbol de decisión del punto a)