## Práctica 2: Datos y exploración de datos

### Apartado 2.1: Datos

Realice los siguientes ejercicios usando el módulo Pandas y cualquier otro módulo adicional que considere. Para evaluar el efecto de las operaciones de preprocesado vamos a considerar un árbol de decisión y un clasificador de vecino más cercano. Evaluaremos el efecto de aplicar el preprocesado midiendo el rendimiento de los dos clasificadores usando la medida de error de clasificación.

1. Obtenga 5 ejemplos de ficheros de datos en formato CSV, ARFF u otro cualquiera de:
    - Weka datasets
    - UCI MLR
2. Evalúe el árbol de decisión y el vecino más cercano sobre los datos originales.
3. Estudie el efecto de la normalización (reescalar en el intervalo [0, 1]) y la estandarización ($\mu = 0, \sigma = 1$) sobre el error de clasificación usando el árbol de decisión y el vecino más cercano. Comente los resultados.
4. Estudie el efecto del análisis en componentes principales sobre el árbol de decisión y el vecino más cercano. Comente lo resultados.
5. Estudie el efecto del muestreo aleatorio del 10% de las instancias sin reemplazamiento sobre el árbol de decisión y el vecino más cercano. Comente lo resultados. Compare los resultados con un muestreo igual estratificado.
6. **Valores perdidos.** Existen diferentes métodos para imputar valores perdidos en la biblioteca *scikit learn* (https://scikit-learn.org/stable/modules/impute.html). Seleccione un conjunto de datos con valores perdidos y dos métodos de imputación. Estudie el efecto de los métodos de imputación sobre los dos clasificadores.
7. *** Selección de características.** Existen diferentes métodos de selección de características en la biblioteca *scikit learn* (https://scikit-learn.org/stable/modules/feature_selection.html). Seleccione un conjunto de datos con un número elevado de atributos y dos métodos de selección de características. Estudie el efecto de los métodos sobre los dos clasificadores.
8. **Discretización.** Existen diferentes métodos de discretización en la biblioteca *scikit learn* (https://scikit-learn.org/stable/modules/preprocessing.html#discretization). Seleccione un conjunto de datos y un método de discretización. Estudie el efecto del método sobre los dos clasificadores.

NOTAS:

- El apartado del manual de *scikit learn* destinado a preprocesamiento se puede acceder en https://scikit-learn.org/stable/modules/preprocessing.html
- Los ejercicios marcados con un ***** son opcionales.
- Se pueden repasar los conocimientos básicos del uso de clasificadores y su evaluación en la documentación de la práctica 2.