

Elements of Machine Learning 2023

Lista 01

26.enero.2023

En esta tarea vamos a explorar dos conjuntos de datos: (1) el conjunto de datos *Palmer Penguins*, que ya iniciamos en el aula, y (2) un conjunto de datos *Tips* de propinas en restaurantes.

El conjunto de datos *Palmer Penguins*, se puede cargar desde Python como

```
df = sns.load_dataset('penguins')
```

o también está disponible en el archivo **penguins.csv**.

El conjunto de datos *Tips*, se puede cargar desde Python como

```
df = sns.load_dataset('tips')
```

o también está disponible en el archivo **tips.csv**.

1. Completar el análisis exploratorio del conjunto de datos *Penguins*. Mencionar los siguientes aspectos:

- el tamaño del conjunto de datos,
- las variables que se incluyen y de qué tipo son,
- indicar si hay datos faltantes, y hacer algún tratamiento con ellos (e.g. removerlos),
- hacer gráficos o tablas de cada variable categórica para indicar su distribución,
- hacer histogramas o funciones de densidad de cada variable numérica,
- cuando considere conveniente, separar estos histogramas por categoría (`hue`),
- hacer *scatterplots*, *pairplots* o mapas de calor en 2D, para identificar relaciones entre variables,
- calcular algunos estadísticos descriptivos.

Pueden usar como material de apoyo <https://seaborn.pydata.org/tutorial/distributions.html>

2. Realizar un análisis exploratorio para el conjunto de datos *Tips*.

3. Redactar un informe con sus análisis de datos, destacando sus conclusiones o *insights* más importantes. Preparar una presentación (de sólo 2 diapositivas), con un resumen informativo.
