

MÓDULO 8: Modelización predictiva

Ejercicio 2: Caso Práctico

Nota preliminar: realizar este trabajo **por parejas**.

Se pide a las parejas elegir un juego de datos de clasificación o regresión de los disponibles en la literatura (no más de 5000 ejemplos para evitar largos procesos de entrenamiento) y realizar un “benchmark” de modelos, tanto de los vistos en clase como de los disponibles en la librería SKLEARN de Python. Se valorará especialmente la contribución y contenido del informe de la práctica en lo que respecta a:

- 1) Las dimensiones del conjunto de datos.
- 2) La utilización de diferentes modelos, no sólo lineales, y su explicación dentro del informe (utilización NO CIEGA de la librería).
- 3) La profundidad (número de parámetros) del ajuste de los modelos y la selección razonada del proceso de validación cruzada.
- 4) La utilización de diferentes scores predictivos para la comparación de los modelos, siempre de acuerdo a la naturaleza del problema planteado (clasificación/regresión, una/varias categorías).
- 5) La evaluación de la significancia estadística de las diferencias entre modelos en términos de performance usando e.g. test de contraste de hipótesis (Wilcoxon) o boxplots.
- 6) La representación de las regiones de decisión de los modelos empleados en el caso de emplear clasificadores.
- 7) La representación de la curva de aprendizaje para diferentes niveles de complejidad del modelo seleccionado y/o número de muestras de entrenamiento consideradas.
- 8) La utilización de conceptos no vistos durante el módulo (i.e. selección no lineal de características, selección de muestras, balanceo con pesos), junto con una explicación que demuestre la asimilación de los conceptos por los alumnos.

Fuentes para consultar fuentes de datos:

- [UCI repository](#)
- <https://github.com/caesar0301/awesome-public-datasets>

Fecha de entrega: a fijar según progreso en la asignatura