**POSTWORK**

**SESIÓN 07**

**Objetivo:**

Construir varios algoritmos de clasificación basados en Machine Learning, y mostrar cuál es el mejor con base en evidencia sólida.

**Si ya tienes un proyecto:**

* Evalúa si tu proyecto es de clasificación o de predicción. Si es de predicción te sugiero que, después de leer toda esta sección, te bases en la siguiente sección. Si tu proyecto es de clasificación, lo más probable es que hayas hecho previamente un proyecto con árboles de decisión y/o redes neuronales artificiales para clasificación.
* Casarte con un único método de clasificación es algo que ningún Data Scientist debería hacer: Es importante experimentar con varios algoritmos de clasificación y ver cuál te arroja mejores resultados. A veces es importante que pensemos por qué cierto algoritmo funcionó mejor que otro. Por ejemplo: ¿será que tus datos pueden agruparse con campanas de Gauss y Naïve Bayes funciona muy bien?
* ¿Recuerdas las bitácoras que hiciste en los postworks de las sesiones pasadas? Ahora es cuando van a brillar. Si hiciste una bitácora de experimentos para Redes Neuronales Artificiales y otra para Random Forests, ya tienes el 50% del postwork hecho. Si no… Pues hay que hacer esas dos, y hacer las de SVM y Naïve Bayes. Compara los resultados y conserva los mejores.
* Puede que en una ronda de experimentos las Redes Neuronales sean mejores, pero si corres los experimentos de nuevo, puede que ahora gane la máquina de soporte vectorial. ¿Por qué? Lo más probable es que quizás los datos de entrenamiento y prueba o las inicializaciones aleatorias hayan variado. Procura utilizar validación cruzada de K-fold y promedia cada resultado para elegir el mejor y no dejarlo al azar.

**Si no tienes un proyecto y deseas crear uno:**

* Puedes utilizar la base de datos de vinos o la base de datos de cáncer de mama para realizar una comparativa entre los cuatro clasificadores con los que hemos trabajado (Naïve Bayes, SVM, Redes Neuronales Artificiales, y Random Forests). Considera los puntos que he mencionado en la sección pasada de este Postwork.
* Te recuerdo que para cargar el dataset de cáncer de mama, puedes utilizar el siguiente código:

from sklearn.datasets import load\_breast\_cancer

dataset = load\_breast\_cancer()

x, y = dataset.data, dataset.target

* Para cargar y utilizar el dataset de vinos, puedes utilizar el siguiente código:

from sklearn.datasets import load\_wine

dataset = load\_wine()

x, y = dataset.data, dataset.target