# **Tarea 2:** *Sistemas de Clasificación y entrenamiento de modelos de caja blanca.*

**Docente:** Nicolás Abuhadba.

**Objetivos del Taller:**

- Aplicar métricas de clasificación y aplicar los principales algoritmos denominados como "Caja Blanca", *Regresión lineal, Regresión polinomial, Modelos regularizados y Regresión logística.*

| | |
|---|---|
| Tipo de actividad | Taller grupal (Máximo 2 integrantes) |
| Taller | Sistemas de Clasificación y entrenamiento de modelos de caja blanca. |
| Evaluación | Sumativa |
| Fecha máxima de entrega | 01-09-2024 |

**Entregable:**

- Elaborar un *Notebook* con el desarrollo del desafío, y luego subir a Classroom el *Notebook* (en Python en extensión. ipynb) en un archivo ".zip" (*Taller2_Apellido_Integrante1_Apellido Integrante 2.zip*). Recuerde el código debe ser legible y contener comentarios de los pasos realizados y una conclusión de cada desafío).

# Actividades:

## Desafío 1:

Después de completar las ultimas unidades de *Machine Learning Avanzado*, llega el momento en el que ustedes, un grupo de consultores expertos en *Data Science*, son contactados por una empresa de ventas por internet en Estados Unidos. Esta empresa recibe pedidos a través de correos electrónicos de ejecutivos. Dado que la dirección de correo para las ventas se encuentra ampliamente difundida en línea y configurada en diversos formularios de sitios web, la empresa creó hace algunos años un **clasificador de SPAM-HAM**. Este clasificador tiene como objetivo identificar correos no deseados (SPAM) y correos legítimos y autorizados (HAM) en inglés. Además, el clasificador ha demostrado ser altamente robusto, con una tasa de precisión superior al 98%.

El clasificador **tiene una precisión y una sensibilidad por sobre el 95%.** Sin embargo, desde hace un año aproximadamente, los ejecutivos de ventas se han quejado de que algunos días después de realizar ciertas campañas en redes sociales el clasificador falla y envía algunos HAM como SPAM y esos clientes nunca reciben respuesta de sus peticiones de compra.


De parte de la gerencia levantan a ustedes **la siguiente petición**:

1. Crear un nuevo clasificador que sea un 10% menos sensible, el cual se pondría en producción sólo por una semana después de iniciadas las campañas, para así no perder los correos HAM. Cuando no hay campañas, volverían a su modelo original. Los ejecutivos están dispuestos a ignorar los correos SPAM que igual le lleguen a su bandeja durante el tiempo de campaña.

2. Documentar las diferencias entre el clasificador actual que tienen en producción y el clasificador menos sensible: para ello se solicita como mínimo entregar un cuadro de diferencias que muestre el modelo usado (si fue regresión logística también u otro), la precisión, la sensibilidad y F1, como así también la traza de ambas curvas ROC/AUC para certificar que se cumplió con lo solicitado.

3. Para resolver este problema, se ha puesto a su disposición el modelo original del clasificador de SPAM-HAM en un archivo joblib y su correspondiente código fuente que generó este modelo en un Python notebook. El set de datos que se usa para entrenar proviene desde un conjunto de datos públicos de *Apache SpamAssassin* (https://homl.info/spamassassin). El mismo Python notebook posee el código necesario para la descarga del dataset. Como entregable, la empresa requiere el nuevo código fuente que genere ambos clasificadores y la documentación solicitada en el punto 2.

## Desafío 2:

El conjunto de datos "Diabetes" contiene información sobre el progreso de la enfermedad en pacientes con diabetes, específicamente la progresión de la enfermedad después de un año, en función de varias variables médicas. El objetivo suele ser predecir la progresión de la enfermedad en función de las características médicas proporcionadas en el conjunto de datos.

El conjunto de datos "Diabetes" tiene 442 muestras y 10 características o columnas. Estas características son las siguientes:

i. Age: Edad del paciente en años.
ii. Sex: Género del paciente (0 para masculino, 1 para femenino).
iii. BMI: Índice de masa corporal del paciente.
iv. BP: Presión arterial media del paciente.
v. S1, S2, S3, S4, S5, S6: Seis mediciones de suero sanguíneo relevantes para pacientes con diabetes.

El objetivo del conjunto de datos es la variable numérica que indica la progresión de la enfermedad después de un año.

https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset

**Preguntas de desarrollo:**

1. Utilizando el conjunto de datos "*Diabetes*", ¿cuál es la progresión de la enfermedad después de un año para un paciente de 45 años, con un índice de masa corporal (BMI) de 25 y una presión arterial media (BP) de 80? Utiliza un modelo de regresión lineal para hacer la predicción.

2. Aplicando una regresión polinomial de grado 2 al conjunto de datos "Diabetes", ¿cuál es la progresión de la enfermedad después de un año para un paciente de 55 años con un índice de masa corporal (BMI) de 30 y una presión arterial media (BP) de 90?

3. Utilizando un modelo de regresión logística, ¿qué probabilidad hay de que un paciente con diabetes sea de género femenino (Sex = 1) y tenga un índice de masa corporal (BMI) de 28? Elige un umbral de probabilidad del 0.5 para determinar si el paciente tiene una alta probabilidad de progresión de la enfermedad.

4. Aplicando el modelo de *regresión lasso* (L2) al conjunto de datos "Diabetes", ¿cuál es la progresión de la enfermedad después de un año para un paciente de 50 años con un índice de masa corporal (BMI) de 27 y una presión arterial media (BP) de 85? Utiliza un parámetro de regularización alpha = 0.5.