# Aprendizaje Automático I

# Práctica Curso 2023/24

## Parte A – Aprendizaje supervisado

## Instrucciones

- Esta práctica consta de un conjunto de ejercicios. Para cada uno de ellos, se deberá elaborar un cuaderno de Python (extensión *.ipynb*) con la correspondiente solución. Debéis utilizar celdas de *markdown* para las explicaciones y celdas de código.
- El código deberá estar comentado.
- La entrega consistirá en un archivo comprimido .zip cuyo nombre seguirá el siguiente formato, donde XX será el grupo asignado en Moodle a la pareja de prácticas:

  **PRÁCTICA_SUPERVISADO_23_24_GRUPO_XX.ZIP**

- Dentro del zip se incluirán los notebooks con el siguiente formato:
  - ejercicio_01.ipynb
  - ejercicio_02.ipynb
  - …
- Los cuadernos se entregarán con todas las salidas guardadas para las celdas de código.
- Se deberá poder ejecutar el cuaderno hasta la última celda, sin errores, en Google Colab. En caso de desarrollar la práctica en otro entorno, debéis aseguraros que ejecuta correctamente en este entorno.
- Para cada uno de los ejercicios se deberán aplicar los pasos necesarios para conseguir los mejores resultados posibles haciendo uso de las técnicas y herramientas vistas en la asignatura. A modo de ejemplo, si un dataset está desbalanceado, se deberá hacer uso de técnicas de *imbalance learning* para evaluar si de este modo se consiguen mejorar los resultados.
- Se valorará el uso de gráficas y la inclusión de explicaciones detalladas, así como del trabajo realizado de cara a conseguir los mejores resultados posibles.

## Ejercicio 1 – Detección de Malware

- Dataset: **01_dataset_malware.zip**

Este dataset contiene información sobre muestras maliciosas y benignas de aplicaciones para Android. Se deberán hacer las siguientes tareas:

- **Tarea 1a:** Realiza una descripción del dataset. Indica qué son las características que describen cada ejemplo y utiliza las herramientas que consideres necesarias para evaluar las características y su aporte a una tarea de clasificación.

- **Tarea 1b:** Realiza las tareas de pre-procesamiento necesarias para poder entrenar algoritmos de clasificación sobre este dataset. Evalúa la posibilidad de obtener subconjuntos de las características que sean relevantes y suficientes para la clasificación.

- **Tarea 1c:** Entrega y evalúa al menos 3 algoritmos de clasificación de los vistos en clase para un problema de clasificación binaria. Utiliza las métricas adecuadas para evaluar cada algoritmo mediante validación cruzada.

## Ejercicio 2 – Clasificación de textos

- Dataset: **02_text_classification.zip**

Este dataset contiene textos y una etiqueta que define el tópico del texto. Se deberán hacer las siguientes tareas:

- **Tarea 2a:** Realiza una descripción del dataset. Analiza los posibles problemas y realiza el preprocesamiento necesario.
- **Tarea 2b.** Realiza los pasos necesarios para generar una o varias representaciones de los textos mediante características que puedan ser utilizadas para entrenar algoritmos de clasificación. Si es necesario, considera una porción representativa más pequeña del dataset para poder entrenar un clasificador.
- **Tarea 2c:** Entrega y evalúa al menos 2 algoritmos de clasificación de los vistos en clase para clasificar el tópico de los textos. Utiliza las métricas adecuadas para evaluar cada algoritmo mediante conjuntos de entrenamiento y test.

## Ejercicio 3 – Clasificación de imágenes

- Dataset: **CIFAR_10 small**

Para cargar el dataset, puedes utilizar:

```
from sklearn import datasets
cifar_10 = datasets.fetch_openml('CIFAR_10_small')
```

Este dataset contiene imágenes de múltiples etiquetas. Se deberán hacer las siguientes tareas:

- **Tarea 3a:** Realiza una descripción del dataset. Realiza los pasos necesarios de preprocesamiento que sean necesarios
- **Tarea 3b:** Entrega y evalúa al menos 3 algoritmos de clasificación de los vistos en clase, incluyendo redes de neuronas de forma obligatoria, para clasificar las imágenes en sus categorías. Utiliza las métricas adecuadas para evaluar cada algoritmo mediante los conjuntos de entrenamiento y test que se proporcionan en el propio dataset.

## Ejercicio 4 – Predicción

- Dataset: **04_dataset_taxi_fare.zip**

Este dataset contiene información sobre el precio del recorrido de un taxi en Nueva York en función de diferentes variables. Se deberán hacer las siguientes tareas:

- **Tarea 4a:** Realiza una descripción del dataset. Indica qué son las características que describen cada ejemplo y utiliza las herramientas que consideres necesarias para evaluar las características y su aporte a una tarea de clasificación.

- **Tarea 4b:** Realiza las tareas de pre-procesamiento necesarias para poder entrenar algoritmos de clasificación sobre este dataset.

- **Tarea 4c:** Entrega y evalúa al menos 3 algoritmos para predicción de los vistos en clase. Utiliza las métricas adecuadas para evaluar estos algoritmos.