Práctica 1

1. Enunciado

El objetivo de esta práctica es construir un proyecto de ML, al que denominaremos **clasificador**[1], para predecir si una imagen tomada de masa mamaria y digitalizada se corresponde con un diagnóstico benigno (*malignant* = 0) o maligno (*malignant* = 1)

La imagen no está disponible. En su lugar hay 8 tablas donde se recogen diferentes características visuales extraídas con técnicas de visión artificial: *traintab01.csv*, ... , *traintab08.csv*.

El proyecto debe ser tal que se pueda poner en producción a partir de su entrega. Esto significa que, una vez entregado, el *cliente* puede probar nuevos ejemplos con un interfaz mínimo: simplemente proporcionando los ejemplos cumpliendo con el formato de las tablas e invocando un script de Python para generar un fichero de etiquetas estimadas.

Una vez cerrada la entrega, se realizará una competición entre todos los proyectos con un conjunto de tablas reservado.

2. Descripción del conjunto de datos

2.1. Ejemplos

Aunque aquí se da una descripción bastante completa de los ficheros, es muy importante dedicar un tiempo a inspeccionar como están presentados los datos antes de programar nada.

En la tabla de la derecha se muestra el número de columnas de las tablas de entrenamiento y competición.

El "2+" significa que la columna 0 es un identificador numérico y la columna 1 es un identificador de texto, y por tanto NO son características de la imagen.

Es decir, la tabla 01 contiene 512 características, la tabla 02 contiene 16 , etc.

ficheros de entrenamiento	ficheros de competición	núm. de columnas
traintab 01	testtab 01	2 + 512
traintab 02	testtab 02	2 + 16
traintab 03	testtab 03	2 + 150
traintab 04	testtab 04	2 + 16
traintab 05	testtab 05	2 + 24
traintab 06	testtab 06	2 + 256
traintab 07	testtab 07	2 + 108
traintab 08	testtab 08	2 + 8

Además, todas las tablas de entrenamiento tienen 2834 ejemplos (filas), mientras que todas las tablas de competición tienen 500.

2.2. Etiquetas

Sólo existe un fichero de etiquetas llamado *train_label.csv* que sirve para entrenar. Se trata de una tabla con 2834 ejemplos y dos columnas. La primera columna es el identificador numérico de la imagen. La segunda columna se denomina *malignant*.

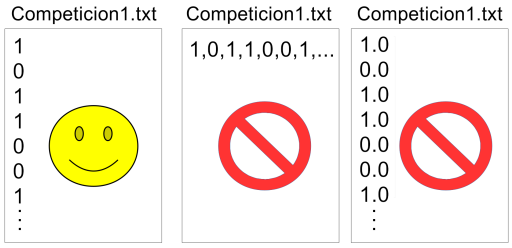
El fichero de etiquetas de la competición se liberará una vez que se publiquen los resultados.

[1] A la máquina que produce etiquetas a partir de entradas le denominaremos **modelo**

3. Condiciones de entrega

- El equipo debe estar formado por 2 alumnos.
- Todos tendrán acceso a los mismos datos, que consisten en ficheros CSV para entrenar y para competir.
- La entrega debe ser un archivo comprimido ZIP que contenga:
  - un fichero llamado **nombres.txt** con el nombre de los alumnos del grupo
  - el código utilizado para entrenar el clasificador
  - el código utilizado para cargar el clasificador y ejecutarlo sobre los ficheros de la competición.
  - Un fichero llamado **Competicion1.txt** donde se habrán guardado las etiquetas estimadas para los datos de la competición con el siguiente **formato OBLIGATORIO**

    !! Una etiqueta por línea, guardada como un entero, es decir sin decimales.



    Por ejemplo, si en el código las etiquetas generadas para el conjunto de datos de la competición se han almacenado en la variable `y_pred`, las siguientes instrucciones generan un fichero CSV que cumple los requisitos:

```
⋙ import numpy as np
⋙ np.savetxt('Competicion1.txt', y_pred, fmt='%i', delimiter=',')
```

  - una breve **memoria** (no más de 3 páginas + portada) explicando como se ha abordado el problema.
- Se puede utilizar el código proporcionado en clase; y si se utiliza código de terceros debe estar indicado con el comentario `#** codigo de terceros ! **`
- La fecha límite para subir el fichero ZIP aparece en la entrega del aula virtual.

4. Checklist

Se valorará cumplir **todos** los requisitos de entrega (esto no da puntos pero sí los quita). Comprueba todo con el siguiente checklist:

- ☑ Fichero *nombres.txt* con el nombre de los alumnos del grupo.
- ☑ Fichero *Competicion1.txt* con las etiquetas.
- ☑ Fichero *Competicion1.txt* formateado correctamente.
- ☑ Memoria en PDF
- ☑ Código fuente
- ☑ Todo empaquetado en un fichero ZIP

Además, asegúrate de que:

- El código está comentado.
- Has desarrollado un proceso correcto para entrenar.
- El proceso para inferir las etiquetas de la competición es correcto.