Práctica 1

1. Enunciado

El objetivo de esta práctica es construir un proyecto de ML, al que denominaremos **clasificador**[1], para predecir el tipo de cultivo en una imagen hiperespectral tomada desde el aire. Pero la imagen no está disponible; y en su lugar tenemos un array de 131 valores de intensidad de las diferentes longitudes de onda registradas.

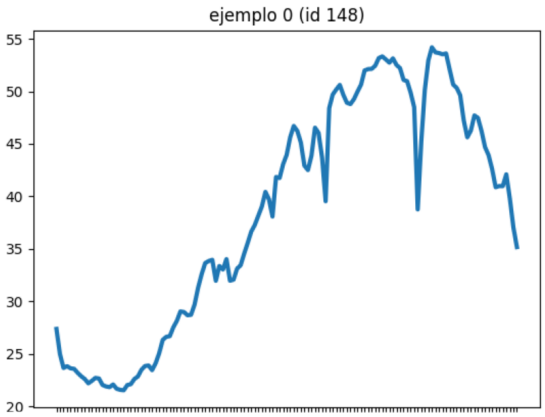
En total hay cinco tipos diferentes de cultivos, denominados 'corn', 'rice', 'cotton', 'soybean' y 'winter_wheat'. De cada uno de ellos hay un número difrente de ejemplos para aprender.

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.

2. Descripción del conjunto de datos

2.1. Ejemplos

Aunque aquí se da una descripción bastante completa de los datos, es muy importante dedicar un tiempo a inspeccionarlos y entenderlos antes de programar nada.



Cada uno de los ejemplos proporcionados es un barrido de 131 longitudes de onda de un cultivo de maíz, arroz, algodón, soja o trigo. Para cada longitud de onda se mide la intensidad de la radiación incidente en el sensor. Así, cada ejemplo genera una *huella* multiespectral como la que se puede ver en la figura de la derecha.

El fichero de entrenamiento (`X_train.csv`) tiene 6289 ejemplos. Cada ejemplo es una fila de 132 columnas, de las cuales la primera es un identificador del cultivo y las 131 restantes son las intensidades registradas.

El fichero de test (`X_test.csv`) tiene tiene el mismo formato pero solo hay 699 ejemplos.

2.2. Etiquetas

Sólo se proporciona un fichero de etiquetas llamado `Y_train.csv`, que son las etiquetas asociadas a cada ejemplo de entrenamiento. Cada fila de este fichero tiene dos columnas: el id del cultivo y el tipo de cultivo y, evidentemente, hay 6289 filas.

¡No existe `Y_test.csv`! **Tu tarea es, precisamente, generalo**.

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

3. Condiciones del modelo

- Sólo se pueden utilizar los modelos incluidos en `Scikit-Learn`.
- Se puede utilizar el código proporcionado en clase; y si se utiliza código de terceros debe estar indicado con el comentario `#** código de terceros ! **`.

4. Condiciones de la competición

- Habrá tres viernes consecutivos para subir un fichero `Y_test.csv` por equipo.
- La entrega se hace únicamente a través del aula virtual, en la actividad creada para ello.
- Después de cada intento se publicará un listado con el mejor resultado de cada equipo.
- El equipo ganador tendrá como premio un artículo con el logo del máster.
- Para calcular el resultado se tiene en cuenta el número de ejemplos de test en cada clase. En la figura de abajo se muestra un ejemplo de como se realiza.

Calculadora de puntos (números inventados)				
Crop	**num.ejemplos**	**puntos por acierto**	**aciertos**	**puntuación**
corn	10	0.1	5	0.5
soybean	20	0.05	14	0.7
cotton	4	0.25	1	0.25
rice	25	0.04	7	0.28
winter_wheat	10	0.1	6	0.6
			TOTAL	**2.33**

5. Condiciones de entrega

- El equipo debe estar formado por 2 alumnos, y basta con que uno haga la entrga, y sólo a través del aula virtual.
- 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 (distinto del anterior).
  - Un fichero llamado `Y_test.csv` donde se habrán guardado las etiquetas estimadas para los datos de la competición con el **formato OBLIGATORIO** que se muestra a la derecha.
  - una breve **memoria** (no más de 3 páginas + portada) con el nombre de los alumnos en la portada, explicando como se ha abordado el problema.

`Y_test.csv`

```
id;Crop
5346;corn
2123;soybean
5424;cotton
2802;rice
528;winter_wheat
3755;rice
```

¡¡ En cada línea aparece el id y el tipo de cultivo, separados sólo por un ';' !!

- La fecha límite para subir el fichero ZIP aparece en la entrega del aula virtual.
  Si se entrega después, la calificación será menor.
- La nota de la práctica NO depende de la posición final en la competición.
  La nota SÍ depende de cuanto trabajo propio se ha realizado.

6. 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 `Y_test.csv` con las etiquetas, formateado correctamente.
- ☑ Memoria en PDF con los nombres en la portada
- ☑ Código fuente para entrenar el clasificador.
- ☑ Código fuente para hacer inferencia a partir de ejemplos nuevos.
- ☑ Todo empaquetado en un fichero ZIP

Además, asegúrate de que:

- El código está comentado.
- Hay dos ficheros: uno para entrenar y otro para hacer inferencia.
  Esto es importante porque al *cliente* sólo se le entrega el segundo
  (¡¿para que quiere el cliente el primero?!)
- Has desarrollado un proceso correcto para entrenar.
- El proceso para inferir las etiquetas de la competición es correcto.

7. Recomendaciones

- Utiliza un IDE para programar y depurar; por ejemplo PyCharm, VS Code, Spyder, etc. Es muchísimo mejor que utilizar un cuaderno Jupyter.
- Prueba diferentes procesamientos de los datos e incluso trata de realizar una visualización en dos dimensiones para entenderlos mejor.
- Prueba diferentes modelos y realiza validaciones cruzadas.
- En la memoria, sé muy conciso. Basta con enumerar las cosas que has intentado. Puedes añadir imágenes; si son autoexplicativas y significativas ayudan a escribir mucho menos. Si además son de tamaño pequeño-mediano puedes poner muchas. Pero si NO son autoexplicativas, tienen el efecto contrario.
- En los comentarios del código puedes escribir todo lo que no puedes escribir en la memoria.