## 1. Objetivo del laboratorio

*Desarrollar de forma autónoma **un Notebook** que permitan explicar distintas hipótesis a partir de varios datasets de entrada, mediante la preparación y visualización de estos.*

## 2. Elementos a utilizar:

- Lenguaje Python
- Librería numérica *NumPy*, pandas, scikit-learn y gráfica *Matplotlib*
- Entorno Anaconda
- Editor Jupyter

## 3. Práctica 1 (Olimpiadas Tokio 2020)

### Objetivo (3 puntos)

Se quiere crear un modelo que permita saber si un atleta español va a tener medalla de oro, plata o bronce en los juegos olímpicos de Tokio que se celebrarán en 2020. Para ello, tendremos en cuenta el lugar que ocupan en un espacio n-dimensional donde n es el número de características de cada atleta.

Para ello usaremos el dataset "Juegos olímpicos" que se encuentra en Moodle. Elige el clasificador que más se adapte de entre los vistos en clase y usa scikit-learn junto con las librerías que necesites para resolver las siguientes cuestiones.

1) Haz todo el preprocesamiento para crear un set de entrenamiento y otro de validación qué permita clasificar atletas qué tengan sólo las características necesarias. Aparte de los totalmente necesarios, usaremos como atributos: Sexo, Edad, Altura y Peso. Explica qué has hecho y porqué. (0,5 puntos)
2) Prueba con distintas configuraciones de las dos métricas principales. La primera métrica corresponde al número de individuos que usarás para clasificar una nueva instancia y la segunda cómo vas a medir la cercanía de esa nueva instancia con el resto. (1 punto)
3) Elige la mejor configuración entre las anteriores. Para ello dibuja una tabla ver cómo evoluciona la clasificación. Dibuja los resultados que se obtienen con ambas configuraciones elegidas cómo las mejores. (1 punto)
4) Utiliza el clasificador para saber que medalla es más probable que ganen Bruno Hortelano (Athletycs), Carolina Marín (Badminton) o la selección femenina de baloncesto (Basketball). (0,5 puntos)

## 4. Práctica 2 (Clasificador de setas)

### Objetivo (3 puntos)

Estamos en otoño, una época en la que muchas personas aprovechan para salir de excursión al campo los fines de semana. Una de las actividades más propias de esta época es la recogida de setas. El problema es que muchas de esas personas que se aventuran a recogerlas no tienen las suficientes nociones para diferenciar una seta venenosa de una comestible. Es por ello qué queremos construir un clasificador que nos proporcione una serie de reglas de manera que cuando una persona vaya al campo y encuentre una seta pueda saber si esta es venenosa o no.

Para ello usaremos el dataset "Setas" que se encuentra en Moodle: mushroom.csv tendrá la información necesaria para entrenar y evaluar el modelo e info.txt la equivalencia a las etiquetas usadas en el dataset. La primera columna es la que indica si una seta es venenosa o no. Elige el clasificador que más se adapte de entre los vistos en clase y usa scikit-learn junto con las librerías que necesites para resolver las siguientes cuestiones.

1) Crea un clasificador en el que uses al menos dos criterios de división distintos. Calcula el error en cada uno de ellos y elige el qué mejor clasifique. (1 punto)
2) Dibuja el modelo elegido en el punto anterior. (0,5 puntos)
3) Selecciona tres reglas que sean las que generalicen lo menos posible e interprétalas. (0,5 puntos)

4) Usa tu clasificador para decidir si son venenos o no la “amanita muscaria” y la “amanaita cesarea”. (1 punto)

Amanita cesarea

Amanita muscaria

## 5. Práctica 3 (Clasificador de imágenes)

### Objetivo (2 puntos)

El etiquetado de imágenes es una tarea ardua. Es por ello y también debido a sus aplicaciones prácticas que los científicos llevan un tiempo intentando mejorar los métodos para clasificarlas automáticamente. En la oficina de correos de Pozuelo de Alarcón quieren poner en práctica un modelo que clasifique las cartas según el código postal escrito en ellas. Para ello vamos a crear un clasificador que leyendo un número escrito a mano pueda saber cuál es. Dicho clasificador funcionará mediante un set de entrenamiento donde se buscará un plano que divida las diferentes clases dispuesta en un espacio n-dimensional dependiendo de sus características.

Para ello usaremos el dataset “load_digits” que se encuentra en scikit-learn. Elige el clasificador que más se adapte de entre los vistos en clase y usa scikit-learn junto con las librerías que necesites para resolver las siguientes cuestiones.

1) Crea un clasificador que permita saber qué número es a partir de una imagen de este. Realiza al menos dos configuraciones y dibuja una tabla donde se muestre la precisión con la que clasifican. (1 punto)
2) Elige 5 números que no hayas usado ni para entrenar el modelo, ni para evaluarlo y clasifícalas. Usa para ello el modelo que mejor clasifique de los del punto anterior. Índica con que error ha funcionado el clasificador. (1 punto)

## 6. Práctica 4 (¿Lloverá o no lloverá?)

### Objetivo (2 puntos)

Los agricultores de una comarca están preocupados por las lluvias ocurridas en los últimos años, en épocas que no son habituales. Para ello han decidido crear un clasificador basado la probabilidad de que la temperatura mínima y máxima y las precipitaciones sean bajas, medias o altas. Se considera que la temperatura es baja si está por debajo de 10 grados, media si está entre 10 y 20 y alta si está por encima de 20. Se considera que ha llovido poco si la medida está por debajo de 1, normal si está entre 1 y 2, y que ha llovido mucho si está por encima de 2.

Para ello usaremos el dataset “lluvias” que se encuentra en Moodle. Elige el clasificador que más se adapte de entre los vistos en clase y usa scikit-learn junto con las librerías que necesites para resolver las siguientes cuestiones.

1) Realiza todo el preprocesamiento necesario para poder entrenar el clasificador. (Ojo: las temperaturas están en grados Celsius.) (1 punto)
2) Crea un clasificador e indica su error. Úsalo para saber si lloverá en los próximos 3 días. Obtener los datos de cualquier fuente de Internet indicando que día es. (1 punto)

## 7. Forma de entrega del laboratorio:

La entrega consistirá en un fichero comprimido RAR con nombre LAB03-GRUPOxx.RAR subido a la tarea **LAB3** que **contenga únicamente**

1. **Por cada práctica** un notebook de Jupyter (archivos con extensión **.ipynb**).
2. La **memoria del laboratorio** que se irá construyendo en el Notebook de manera que se explique todo lo que se hace.

**Las entregas que no se ajusten exactamente a esta norma NO SERÁN EVALUADAS.**

## 8. Rúbrica de la Práctica:

**1. IMPLEMENTACIÓN: Multiplica la nota del trabajo por 0/1**

Siendo una práctica de Data Mining, todos los aspectos de programación se dan por supuesto. La implementación será:

- Original: Código fuente no copiado de internet. Grupos con igual código fuente serán suspendidos
- Correcta: El programa funciona y ejecuta correctamente todo lo planteado en los apartados de cada práctica.
- Comentada: Inclusión (**obligatoria**) de comentarios.
- En las gráficas que se realicen proporciona todos los datos que creas necesarios.

**2. MEMORIA DEL LABORATORIO**

Obligatorio redacción clara y correcta ortográfica/gramaticalmente. Cada paso que se haga tiene que estar justificado.