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 (Asteroides peligrosos)

Objetivo (3 puntos)

La NASA quiere crear un modelo que permita saber si un asteroide es peligroso para el planeta Tierra o no. Para ello, tendremos en cuenta el lugar que ocupan en un espacio n-dimensional donde n es el número de características de cada asteroide.

Para ello usaremos el dataset "nasa.csv" 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. Muestra todos los resultados del algoritmo paso a paso.

1) Haz todo el preprocesamiento para crear un set de entrenamiento y otro de validación qué permita clasificar asteroides qué tengan sólo las características necesarias. Usaremos como atributos todos los del dataset excepto "Orbiting Body" y "Equinox". 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. ¿Qué decisiones has tomado? ¿Por qué? (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 ocurre con los asteroides de un dataset que se llame "nasa_clasificar.csv" que obtendremos del dataset proporcionado. (0,5 puntos)

4. Práctica 2 (Clasificador de crímenes)

Objetivo (3 puntos)

La ciudad de San Francisco es famosa entre otras cosas por la proliferación de empresas del ámbito tecnológico. Esto ha llevado a una profunda desigualdad entre sus habitantes, por lo que la proliferación de crímenes ha aumentado. Es por ello qué se quiere construir un clasificador que proporcione una serie de reglas de manera que se pueda saber que crímenes se van a cometer teniendo en cuenta el momento del día (noche de 0:00 a 7:59, mañana de 8:00 a 15:59 y tarde de 16:00 a 23:59), día de la semana y distrito. Muestra todos los resultados del algoritmo paso a paso.

Para ello usaremos el dataset "sf_crimenes.csv" 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) 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 y otras tres que especialicen los menos posible. Interprétalas. (0,5 puntos)
4) Usa tu clasificador para decidir qué tipos de crímenes ocurrirán una mañana de jueves en el distrito de Taraval. También para uno que ocurra la noche del sábado en el distrito de Park. Por último, un lunes por la mañana en el distrito Central. (1 punto)

5. Práctica 3 (Especies de monos)

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 aduana del aeropuerto de Madrid se intenta luchar contra el tráfico de animales exóticos. Para ello se va a crear un clasificador que realizando una foto a un animal (en este caso monos) pueda decidir si pertenece a una especie en peligro de extinción o no. 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. Muestra todos los resultados del algoritmo paso a paso.

Para ello usaremos el dataset “monos.zip” 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 especie de mono es a partir de una imagen. Realiza al menos dos configuraciones y dibuja una tabla donde se muestre la precisión con la que clasifican. (1 punto)
2) Elige 5 imágenes de diferentes especies 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 (Desguace Beni)

Objetivo (2 puntos)

Un desguace de Madrid quiere automatizar la entrada de coches para ser más óptimo a la hora de despiezarlos y separar las distintas partes de los coches que reciben. Para ello han decidido crear un clasificador basado en las diferentes características de los coches y la probabilidad que pertenezca a una clase de coche u otra.

Para ello usaremos el dataset “cars.csv” 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. Muestra todos los resultados del algoritmo paso a paso.

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 a qué clase corresponden al menos 5 coches que no hayas usado para entrenar el modelo. (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.