## 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 (Investigación en cáncer de mama)

### Objetivo (3 puntos)

Se dispone de un set de datos de cáncer de mama. En dicho dataset se contemplan datos sobre características de los tumores. Algunos de ellos son distintas medidas de la misma característica por lo que son muy dependientes entre sí y generan ruido. Por ejemplo, "radius_mean", "radius_se" y "radius_worst".

A partir de esta información, un equipo de investigación de oncología quiere crear un Decision Support System (https://es.wikipedia.org/wiki/Sistemas_de_soporte_a_decisiones) para entender mejor cuales son las características que mas influyen en el diagnóstico.

Para ello usaremos el dataset "cancer.csv". 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. Elimina aquellos datos que sean muy dependientes de otros y transforma en categóricos con 3 valores los que miden el área, el diámetro y la compacidad (0,5 puntos)
2) Crea los distintos clasificadores en el que utilices al menos dos criterios de división distintos o medidas de desorden. Calcula el error en cada uno de ellos y elige el qué mejor clasifique. (0,5 puntos)
3) Dibuja los modelos elegidos en el punto anterior. (0,5 puntos)
4) Selecciona dos reglas que sean las que generalicen lo menos posible y otras dos que especialicen lo menos posible. Interprétalas. Si la estructura obtenida para sacar las reglas es demasiado grande repite el proceso para un 10% de los datos. (0,5 puntos)
5) Usa tu clasificador para clasificar a 5 individuos que no se hayan usado en los pasos anteriores. Dichos individuos deberán presentar diferentes situaciones. (1 punto)

## 4. Práctica 2 (TCGA)

### Objetivo (2 puntos)

TCGA (The Cancer Genome Atlas) es un proyecto colaborativo dirigido por el Instituto Nacional del Cáncer (NCI) y el Instituto Nacional de Investigación del Genoma Humano (NHGRI) en los Estados Unidos. Su objetivo principal es caracterizar exhaustivamente las alteraciones genómicas en el cáncer mediante el análisis de grandes conjuntos de datos de pacientes con cáncer. Se pretende crear un clasificador de cáncer que tenga en cuenta las probabilidades de pertenecer a una clase dependiendo de las distintas variables y sus valores. 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 con datos categóricos (en el caso de transformaciones de datos continuos se usarán los cuartiles creados por el diagrama de tallos y hojas). Muestra las distintas tablas de distribución. (1 punto)

2) Crea el clasificador e indica su error. Úsalo para saber a qué clase corresponden al menos 10 clientes que no hayas usado para entrenar los modelos. (1 punto)

## 5. Práctica 3 (Detección de malware)

### Objetivo (2 puntos)

En la Universidad Francisco de Vitoria se quieren trazas de comunicación de la red para encontrar distintos malwares. 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 "Malware". 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, realiza al menos tres configuraciones y dibuja una tabla donde se muestre la precisión con la que clasifican. ¿Cómo funcionaría si no usamos kernels? ¿Y al usar distintos kernels? (1 punto)
2) Elige 10 imágenes 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 o acierto ha funcionado el clasificador. (1 punto)

## 6. Práctica 4 (Diagnóstico de cáncer con genes)

### Objetivo (3 puntos)

El fichero "genes.csv" contiene información de dos genes y la posibilidad de tener cáncer o no. Con todo ello se quiere crear un modelo que permita hacer un diagnóstico para un nuevo paciente teniendo en cuenta su similitud en un campo de n-dimensiones

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, otro de validación y uno de test qué permita hacer un diagnóstico lo mas preciso posible aplicando las estrategias pertinentes. (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 la mejor configuración y los distintos hiperparametros del punto anterior para ver su evolución. (1 punto)
4) Utiliza el clasificador para saber que ocurre con los datos de un del dataset de test "pacientes_test.csv" que obtendremos del dataset proporcionado. (0,5 puntos)

## 7. Forma de entrega del laboratorio:

La entrega consistirá en un enlace al github donde esté el jupyter notebook con la resolución de los 4 ejercícios

## 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 original. 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.