PREPARACIÓN DE UN *DATASET* EN UN PROYECTO DE ML

El preprocesamiento y la limpieza de datos son tareas importantes que se deben llevar a cabo para que un conjunto de datos se pueda usar en el entrenamiento de modelos. Los datos reales a menudo son incompletos y en ocasiones pueden presentar valores anormales "*outliers*". El pre-procesado de datos se puede llevar a cabo en mediante el uso de programas similares a Microsoft Excel y por lo tanto no requiere de habilidades específicas de programación, aunque siempre será más rápido y eficiente realizarlo mediante código.

Para la realización de esta práctica es necesario contar con un conjunto de datos en bruto que queramos emplear con un objetivo, ya sea para la predicción de una variable continua (modelo de regresión) o para la predicción de una variable categórica (modelo de clasificación). Este conjunto de datos puede ser propio, obtenido desde internet o uno de los propuestos a continuación. Os planteamos cuatro *datasets* disponibles de forma gratuita en la plataforma Kaggle (solo requiere introducir una dirección de email para el registro):

https://www.kaggle.com/datasets/krcsoft/knn-data

https://www.kaggle.com/datasets/vijayaadithyanvg/breast-cancer-prediction

https://www.kaggle.com/datasets/vpkprasanna/melanoma-tumor-size-prediction-machinehack

https://www.kaggle.com/datasets/pranavraikokte/braintumorfeaturesextracted

Para la realización de la práctica es recomendable haber visto ya las clases del módulo 2, al menos la clase "Datos en el entorno de la Física Médica", especialmente la parte de modelos predictivos.

Es preferible el uso de un *dataset* propio, ya que el conocimiento de las características (origen, significado etc) puede resultar de importancia a la hora del pre-procesado. En ningún caso será necesario compartir el conjunto de datos elegido.

A continuación responda las siguientes preguntas:

1. Describa brevemente el *dataset* elegido, indicando el objetivo, número de observaciones y número de características disponibles.

2. Describa las características del *dataset*, el tipo (categóricas o continuas) y si están completas. En caso de no estar completas proponga, de forma razonada en cada caso, un método para solventar este problema.

3. Características continuas. Analice en cada caso la distribución de las características y los posibles valores atípicos (*outliers*). Proponga un método de normalización si corresponde.

4. Características categóricas. Analice en cada caso la distribución de las clases. Proponga un cambio en la presentación si corresponde.

5. En el caso de ser un modelo de clasificación, analizar si se trata de un modelo balanceado en la distribución de clases. En caso de estar desbalanceado proponga un método para solventar este problema.

6. Selección de características. Analice el número de características disponibles y si resulta un número adecuado para el entrenamiento del modelo propuesto. Justifique en cualquier caso la selección y/o el descarte de características.