PRACTICE 1: Reducción Dimensionalidad (Parte II)

En esta práctica haremos un estudio de reducción de dimensionalidad. Partiremos de la base de datos de la Parte I previamente procesada (limpia y preparada) y separado un conjunto de entrenamiento del resto de los conjuntos (validación y test, en el caso de mantener el conjunto de validación).

Aunque en esta práctica no se pide, deberíamos realizar pruebas con los distintos conjuntos de variables y con modelos diferentes. Por cada prueba se obtendrían indicadores del rendimiento del modelo predictivo resultante del entrenamiento al aplicarlo sobre el conjunto de test. Posibles indicadores serían: *AUC-ROC*, *accuracy*, *sensitivity* y *specificity*. De esta forma podríamos seleccionar el conjunto de variables y modelo que mayor rendimiento alcanzase.

En nuestro caso vamos a sustituir la modelización por la visualización de los datos en 2D o bien histogramas cuando no sea posible la representación 2D.

En esta práctica se deben estudiar los siguientes métodos de reducción de dimensión:

1. **Aproximación con test estadísticos.** En el conjunto de datos existen variables continuas y variables categóricas, para el caso de las variables continuas usaremos el estadístico Chi-Square (que asume distribuciones normales) y para el caso de las variables categóricas utilizaremos el test Mann-Whitney U que es la versión no paramétrica del test de Student.

   https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/

   En función de los p-valores, indica con qué variables te quedarías para un primer estudio. Representa las dos variables más significativas (scatter plot). Para diferenciar las clases, representamos cada clase de un color distinto.

2. **Información mutua.** Realizad el mismo estudio que en el apartado anterior. Ahora no es necesario distinguir entre variables continuas y variables categóricas transformadas.

3. **Análisis Discriminante Lineal.** En este caso solo vamos a trabajar con las variables continuas (no hay opción con las categóricas) y solo podremos obtener un resultado, el correspondiente a la proyección del único autovector cuyo autovalor es distinto de cero. ¿Cuál sería el vector de proyección W?.
   En este caso, representa el histograma resultante de la nueva variable que se obtiene.
4. **Análisis Componentes Principales.** Como en el caso anterior trabajaremos solo con las variables continuas y aquí sí que es posible tener más variables transformadas, las correspondientes a las proyecciones sobre las componentes principales. Mostrad ordenadamente, de mayor a menor explicabilidad, las variables originales. ¿Cuál sería la matriz de proyección W?
   Representa el diagrama de *scatter plot* para las dos primeras componentes principales.