PRÁCTICA 2

A partir de la base de datos de la práctica anterior, vamos a estudiar la importancia (o no) de reducir la dimensionalidad.

Como en el conjunto de datos, la variable dependiente es binaria, vamos a comprobar la posible mejora en la predicción de los resultados aplicando regresión logística como modelo predictivo.

Para ello, inicialmente se deben crear dos conjuntos de datos: entrenamiento y test, con una proporción 70:30.

A continuación, se aplicarán las técnicas de selección de variables para seleccionar las características más significativas para discernir entre enfermo y control.

Hemos visto información mutua y métodos estadísticos como chi2. A priori, no sabemos cuál es el número de variables. Para los modelos estadísticos podríamos seleccionar los más significativos para un umbral dado. Vamos a aplicar un método *forward* incrementando de una en una el número de variables, observaremos en cada caso cómo varía el resultado.

Para el caso de extracción de variables (PCA y LDA), iremos viendo la evolución a medida que vamos añadiendo más componentes.

Se construirá una tabla resumen similar a la aquí expuesta

REGRESIÓN LOGÍSTICA	Accuracy	Precision	Sensitivity	Specificity
Datos Completos – Sin reducción				
MI – 2 variables				
MI – 3 variables				
…..				
Chi2 – 2 variables				
Chi2 – 3 variables				
…..				
PCA (2 componentes)				
PCA (3 componentes)				
PCA (4 componentes)				
….				
LDA (2 componentes)				
LDA (3 componentes)				
LDA (4 componentes)				