## Informe Actividad 3

En esta actividad, se ha estudiado el dataset “Dataset expresión genes.csv”, que incluía información sobre la expresión génica en diferentes pacientes con cáncer, al igual que información sobre otras variables sociodemográficas.

Primero se ha aplicado un PCA sobre los datos de expresión génica, reduciendo la dimensionalidad de los datos mediante la identificación de una serie de Componentes Principales (PCs por sus siglas en inglés) que explican la mayor parte de la variabilidad de los datos originales. La contribución de cada variable a cada PC se puede encontrar en la Tabla 2 y, si se analiza a la Tabla 3, la primera PC es la que más variabilidad explica, la cual decae de forma progresiva a medida que se avanza en el orden de las PCs. De todas ellas, solo se han seleccionado las 5 primeras ya que con ellas se explica más del 70% de la varianza debida a la expresión génica del dataset.

Luego se ha construido un modelo de regresión logística para intentar predecir la metástasis en los pacientes a través de las variables estudiadas. Primero se ha realizado una regularización ya que, aún habiendo reducido la dimensionalidad mediante el PCA, seguíamos teniendo demasiadas variables. El método elegido fue ElasticNet, que ha permitido seleccionar 19 variables además de los 5 PCs. Tras construir 2 modelos de regresión se decidió hacer un 3º únicamente con las PCs ya que las variables sociodemográficas tenían o bien una altísima colinealidad o bien no eran significativas (todos sus ORs cruzaban el 1). Los resultados del 3º modelo se pueden ver en la Tabla 4 expresados como terciles de las ORs. Como se puede apreciar, los ORs significativos son los correspondientes al 3º tercil de las PC 3 y 4, siendo respectivamente 6.26 y 0.16. Esto significa que los pacientes con mayor expresión de los genes significativos para la PC3 (reflejados en la Tabla 3c) tendrán 6.26 veces más probabilidades de tener metástasis con respecto a los del primer tercil. Por contraparte, para la PC4, los pacientes con mayor expresión en los genes significativos de la Tabla 3d tendrán metástasis 0.16 veces por cada vez que la tenga un paciente de t1. Los primeros genes son posibles oncogenes y los segundos posibles genes supresores de tumores y constituyendo todos ellos posibles dianas terapéuticas, bien para su respectiva represión o potenciación.