# Análisis Avanzado de Datos

Taller 1

9 de marzo de 2023

**Instrucciones**

- La calificación se dará sobre 100 puntos y el trabajo se desarrolla preferiblemente en parejas (ver consideraciones).
- Una persona de la pareja debe enviar el informe al correo andresn.lopez@urosario.edu.co o la URL de su repositorio publico junto a la página web de su trabajo (ver consideraciones) **el día 22 de marzo antes de la media noche**. Por cada minuto posterior a la fecha y hora de entrega se restará un punto a la calificación obtenida.

**Consideraciones** Estas consideraciones se verán reflejadas positivamente en la nota en forma de bonificación en caso de presentar una nota por debajo de la nota máxima:

- El formato de entrega del trabajo puede ser un archivo word guardado en formato pdf, pero se recomienda a los estudiantes a realizar un cuaderno en rmd: vea un ejemplo: https://anlopezl.github.io/AED/NB1_E.html
- Se invita a los estudiantes a versionar su trabajo en git y trabajar en equipo usando la plataforma. Si se trabaja individualmente, este punto no será considerado en la bonificación.
- El estudiante puede publicar su trabajo mediante github pages y enviar la URL correspondiente.

**Problema**. El conjunto de datos `taller1.txt` contiene la información del perfíl genómico de un conjunto de 1200 líneas celulares. Para estas se busca determinar cuáles de los 5000 genes (ubicados en cada columna) son de relevancia para la predicción de la variable respuesta (efectividad del tratamiento anticancer, medida como variable continua). Responda las siguientes preguntas:

(1) ¿Hay multicolinealidad en los datos? Explique sucintamente.

(2) Separe aleatoriamente (pero guarde la semilla) su conjunto de datos en dos partes:

- Entrenamiento: 1000 líneas celulares
- Prueba: 200 líneas celulares.

(3) **Usando los 1000 datos de entrenamiento**, determine los valores de $\lambda_r$ y $\lambda_l$ de regesión ridge y lasso, respectivamente, que minimicen el error cuadrático medio (ECM) mediante validación externa. Utilice el método de validación externa que considere más apropiado.

(4) Ajuste la regresión ridge y lasso con los valores estimados de $\lambda_r$ y $\lambda_l$ obtenidos en (3) **usando los 1000 datos de entrenamiento**.

(5) Para los modelos ajustados en (4) determine el más apropiado para propósitos de predicción. Considere únicamente el ECM **en los 200 datos de prueba** para su decisión.

(6) Ajuste el modelo seleccionado en (5) **para los 1200 datos**. Note que en este punto ya tiene un $\lambda$ estimado y un modelo seleccionado.

(7) Grafique las trazas de los coeficientes en función de la penalización para el modelo ajustado en (6).

(8) En un párrafo resuma los resultados obtenidos dado el objetivo inicial del estudio.

Note que existen múltiples viñetas y complementos con detalles en el ajuste de modelos penalizados mediante el uso de R y el paquete glmnet (por ejemplo https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf