Project 1 – *n*-fold cross-validation

This assignment builds upon the R/RStudio class and expands the n-fold cross-validation example.

1. for the assignment use the second dataset called TCGA_breast_cancer_ERpositive_vs_ERnegative_PAM50.tsv that shows ER assignment for each sample (Positive vs. Negative).
2. compute 5-fold and 10-fold cross-validation estimates of prediction accuracies of ER using all genes by utilizing logistic regression and compare with NNC (2x2 table).
3. modify the the R markdown document template to report your computation and results in a table format.
4. comment on the quality of results
5. In the second part of the assignment use Project1fs.R to process a large data set by first removing all genes with sd $<$ 1 and subsequently use Feature selection to pick top 50 genes vs top 100 genes for cross-validation based on the t-test statistic.
6. For extra credit – please replace centroid based classifier with one utilizing logistic or lasso regression similarly to the first part of the assignment and report on any difficulties.

For the assignment use Project1.Rmd file which has a number "FIXME:" labels indicating where your intervention is required. There is a companion Project1.R where you can test and debug your code before adding it to Project1.Rmd.

For extra points use lasso regression on the large dataset instead of logistic regression.

The assignment is due on – February 12, 2026 midnight.

The submission should be zip compressed file named "project1-[*your UC username*].zip" (e.g. "project1-lastnfi.zip") which includes project1.Rmd, project1.docx and any supporting R files. The zip file should be uploaded Canvas The assignment entry in Canvas will be created shortly.