**Assignment 3 Breast Cancer Stage Classification**

Breast cancer (BRCA) is the most common cancer in women. One important task to improve the survival rate of BRCA patients is identifying the cancer stage and applying different treatment strategies. We can train a model to classify cancer stages using RNA-seq of patient samples.

Tasks:

1. Prepare a dataset using TCGA-BRCA RNA-Seq data as features and cancer stages as labels. (Hint: you can find the processed RNA-Seq data and patient phenotype data from UCSC Xena)
2. Applying data processing methods. (Normalization, Training-Test split, etc.)
3. Applying three different classification estimators and optimizing the parameters through cross-validation.
4. Comparing three estimators by evaluating the performance on the test dataset.
5. Applying feature selection to improve performance.

You are encouraged to do object-oriented programming style and comment your code detailly. You should upload your program files in a zip and a separate pdf report of fewer than three pages.

Dataset:

<https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Breast%20Cancer%20(BRCA)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443>