BIOMEDICAL DATA SCIENCE HOMEWORK #3

DUE DATE 10/25/2018

- Please turn your homework code through a GitHub Repository (see below).
- No late homework will be accepted (you will not be able to submit the repository link to Canvas after the due date). Do not change the repository after the due date; your homework will not be graded.
- 20 points total

Use the accompanied dataset for this homework. Read the dataset description below carefully and make sure you understand the dataset features and values.

> *Dataset description: Dataset description: This dataset (Colon Cancer) contains expression levels of 2000 genes taken in 62 different samples. For each sample, it is indicated whether it came from a tumor biopsy or not (0/1). Note that the first column in the file corresponds to the label of the instance. See the Genes.txt file for description of genes and tissues.*

a) Load the dataset in an iPython notebook [2 point].

b) Feature selection is an important machine-learning task that allows us to select the most important features in a given dataset. Scikit-learn provides multiple methods for choosing the best features. Use the Recursive Feature Elimination method (REF) with cross-validation here, and show a plot to demonstrate the performance versus number of selected features [11 points].

c) Use the holdout method for testing using only the selected features. Report the performance. [5 points].

d) Create a GitHub repository and share your code via GitHub with the instructor by submitting the link on Canvas [2 points].