SMDA 2018/19 – Exercise 3, Lecture L6 - 28/11/2018

Exercise 3: Analysis of Prostate Cancer dataset – variable subset selection
Please, execute the following tasks and provide answers to the proposed questions.

1. Open your kernel SMDA_EX2(L5)_ProstateCancer_Surname in Kaggle

2. Generate a copy called SMDA_EX3(L6)_SubsetSelection_Surname by the Fork button

3. Starting from the ols models achieved in the last steps, perform best-subset selection.

- Generate one model for each combination of the 8 variables available
- For each model compute the RSS on training and test set, the number of variables and the R^2 of the model
- Save these numbers in suitable data structures

4. Generate a chart having the subset size in the x-axis and the RSS for the training set of all models generated at step 3 in the y-axis

5. Generate a chart having the subset size in the x-axis and the R^2 of all models generated at step 3 in the y-axis

6. Generate a chart having the subset size in the x-axis and the RSS for the test set of all models generated at step 3 in the y-axis

7. Perform forward selection

- Start from the empty model
- Add at each step the variable that minimizes the RSS (other performance measures can be used)

8. Generate a chart having the subset size in the x-axis and the RSS for the test set of the models generated at step 7 in the y-axis

9. Perform backward selection

- Start from the full model
- Remove at each step the variable that minimizes the RSS (other performance measures can be used)

10. Generate a chart having the subset size in the x-axis and the RSS for the test set of the models generated at step 9 in the y-axis