**Project 2**

Data refer to 569 patients from a cancer study on breast cancer. Variables are computed from a digitized image of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus:   
  
a) radius (mean of distances from center to points on the perimeter)   
b) texture (standard deviation of gray-scale values)   
c) perimeter   
d) area   
e) smoothness (local variation in radius lengths)   
f) compactness (perimeter^2 / area - 1.0)   
g) concavity (severity of concave portions of the contour)   
h) concave points (number of concave portions of the contour)   
i) symmetry   
j) fractal dimension ("coastline approximation" - 1)

There are also two more variables, namely the ID (not useful for this project) and the  Diagnosis (M = malignant, B = benign)  for the tumor.

The project has 3 parts

**Part I**

Use at least 2 different methods in order to create a rule to predict whether the tumor is malignant or benign. Explain the methods, how good is their prediction and examine whether one needs all the variables for this classification. You need to explain in sufficient detail all the procedures and measures used.

**Part III**

Forget now the column with the diagnosis. Apply at least 2 methods of clustering in order to assign the patients into clusters. Explain which method do you prefer, how good is your clustering, how many cluster you need and which are the characteristics of them,  and which variables you really need for this. When you end up with just one clustering result, check how the derived clustering relates to the diagnosis made.

**PART III**

Examine whether the usage of PC improve  your findings in both Part I and II. In other words, create a number of  principal components and then use them instead of the initial variables to see whether you can create a classification rule as in part I or you can improve clustering done in PART II

You have to write a detailed enough report to explain to your boss about that, explaining the model you used together with sufficient technical details on what you have done.   Use tables and plots that really contribute to your story. Explain what other information you may need. It is important that the report is self-contained and if somebody reads it could follow the arguments even with limited statistical knowledge.

You can use whatever package you prefer.  The code (or sufficient details how the fitting was done if the fitting was done without script) needs to be attached but NOT be part of the report. The report must be uploaded in moodle.

Deadline  Sunday 19/6/2016, 23.59.  After that date  you will have -1 point for every day of delay.