Lab 9 task–Machine Learning Methods for Classification

The dataset "PimaIndiansDiabetes" available in R package "**mlbench**" originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 768 women from a population near Phoenix, Arizona, USA. The outcome variable "diabetes" was the diabetes test result, either positive or negative. There were 8 predictors: pregnant, OGTT(Oral Glucose Tolerance Test), blood pressure, skin thickness, insulin, BMI(Body Mass Index), pedigree diabetes function and age. Details are given below.

```
str(PimaIndiansDiabetes)
```

```
'data.frame':     768 obs. of  9 variables:
  $ pregnant: num  6 1 8 1 0 5 3 10 2 8 ...
$ glucose : num  148 85 183 89 137 116 78 115 197 125 ...
$ pressure: num  72 66 64 66 40 74 50 0 70 96 ...
$ triceps : num  35 29 0 23 35 0 32 0 45 0 ...
$ insulin : num  0 0 0 94 168 0 88 0 543 0 ...
$ mass    : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
$ pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
$ age     : num  50 31 32 21 33 30 26 29 53 54 ...
$ diabetes: Factor w/ 2 levels "neg","pos": 2 1 2 1 2 1 2 1 2 2 ...
```

The goal of this lab task is to perform classification using all 8 predictors to determine whether women in a test data set can be accurately assigned to either the positive or negative class. In general, you are required to do the following steps:

1. Get familiar with the dataset and prepare data by
    - inspecting the summary statistics and structure of the dataset
    - converting the test outcome to a binary variable Y (1=positive, 0=negative)
    - selecting predictors and Y for classification
2. Splitting data into training (80%) and test (20%) sets
3. Apply logistic regression (LR), kNN and support vector machine (SVM) for classification.
    - For logistic regression, you are required to identify significant predictors at level 0.05.
    - For kNN, you are required to choose the best value of k.
    - For SVM, you are required to use a linear kernel function and a nonlinear kernel (such as radial kernel), and tune parameters in SVM to find the best SVM.
4. Comment on the comparison of performance between LR, kNN (best choice) and SVM (best choice) classifiers in Part 3 in terms of classification accuracy and misclassification rates based on confusion matrices.