Lab 9 task (Assignment 2) –Machine Learning Methods for Classification

The dataset "rotterdam" available in R package "**survival**" contains information on 2982 primary breast cancers patients whose clinical and pathological records after surgery were extracted from the Rotterdam tumor bank and included in 15 columns. Details are given below. The outcome variable **recur** indicates whether a patient experienced a recurrence (relapse) of breast cancer with 1=relapse and 0=no relapse. We are interested in classifying those alive patients (**death=0**) into two classes (relapse and no relapse) based on 6 clinical and pathological predictors: **age**, **size**, **grade**, **nodes**, **pgr** and **er**.

```
str(rotterdam)
```

```
'data.frame':  2982 obs. of  15 variables:

 $ year  : int  1992 1984 1983 1985 1983 1983 1993 1988 1988 1988 ...
 $ age   : int  74 79 44 70 75 52 40 53 60 52 ...
 $ meno  : int  1 1 0 1 1 0 0 1 1 0 ...
 $ size  : Factor w/ 3 levels "<=20","20-50",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ grade : int  3 3 2 3 3 3 2 2 2 3 ...
 $ nodes : int  0 0 0 0 0 0 0 0 0 5 ...
 $ pgr   : int  35 36 138 0 260 139 13 1 627 316 ...
 $ er    : int  291 611 0 12 409 303 4 4 151 976 ...
 $ hormon: int  0 0 0 0 0 0 0 0 0 0 ...
 $ chemo : int  0 0 0 0 0 0 0 0 0 0 ...
 $ rtime : num  1799 2828 6012 2624 4915 ...
 $ recur : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dtime : num  1799 2828 6012 2624 4915 ...
 $ death : int  0 0 0 0 0 0 0 0 0 0
```

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 outcome variable recur into a factor variable using function factor()
   - create a new dataframe which only contains the relevant predictors and outcome variable used in the 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 and explain their effects on the risk of cancer relapse.
   - 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 or sigmoid 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 (including the false positive rate and false negative rate).