# Assignment 2 : Statistical Learning

Risk classification is a key factor in many companies such as banks and insurance companies. In general companies using costumers informations as a training dataset to classifiy the risk of the new costumers based on some algorithm such as the k-nearest neighbor algorithm or decistion tree.

The aim of this assignment is to apply the k-nearest neighbor algorithm to classify the customers as either "good risk" or "bad loss" by suing the real-world `ClassifyRisk` dataset. This dataset is available in the R package `liver`. You could find more information about this dataset at the following link on page 6: manual of the liver package. This dataset has 6 variables in which the variable *risk* is the trarget variable along with 5 predictors, *mortgage*, *number of loans*, *age*, *marital status*, and *income*. In the target variable, customers are classified as either "Good Risk" or "Bad Risk".

Your task is to answer to the following questoins and creating a report as a **R-markdown** ( .Rmd file including R code). Please upload your **R-markdown** file with the **HTML** file on Canvas at the latest on **Wednesday 14 October 2020 at 23:59** (Amsterdam time). The total number of points assigned is 100.

### 1- Importing and understading the dataset (15 points)

Import the `ClassifyRisk` dataset which is available in the `liver` package. Report a summary of the dataset by using appropriate R functions. What is the number of variables? Which type of variables are they? Are there any missing values? What would be your strategy to deal with the missing data?

### 2- Partitioning the dataset (10 points)

Partition the dataset randomly into a training set (70%) and a test set (30%) by use twofold cross-validation technique.

### 3- Validate the partition (15 points)

The composition of the training and test data sets should be as similar as possible. Validate the partition by using appropriate hypothesis testing.

### 4- Appling the KNN algorithm (20 points)

Find the k-nearest neighbor predictions for the test set using k = 2 and k = 90. Explain how you preparing the variables for the KNN algorithm. Which variables need to be normalized? Use an appropriate normalization function. Which variables need dummy variable?

### 5- Evaluate the accuracy of the predictions (10 points)

Evaluate the accuracy of the predictions using the Mean Square Error (MSE). You could use the `mse` function from the R package **liver**.

**6- Optimal value of k for the kNN algorithm (15 points)**

In the previous questions, to find the k-nearest neighbor for the test set, we set k = 2 and k = 90. But why 2 or 90? Find out the optimal value of `k`.

**7- Writing a report (15 points)**

Create a report as an R-markdown which should includes the R code and the result of the code beside your interpretation.

**8- Bonus question (30 points)**

We want to apply the kNN algorithm for analyzing the *bank* dataset which is available in the **R** package **liver**. You could find more information about the *bank* dataset at the following link on pages 4-5: manual of the liver package. In the weekly exercises of week 5, we applied the kNN algorithm for this dataset by using all the 19 predictors in this dataset. But we know, we should use only those predictors that have a relationship with the target variable. So, your task is first to find out which of those predictors in the dataset have a relationship with the target variable `deposit`. Then apply the kNN algorithm by using only those predictors that have a relationship with the target variable. Finally, report the optimal value for the `k`.