# Tutorial 9 - Revision

Consider a random sample of 1,000 people. Their health information is recorded for a study about factors that may affect the diabetes status of people. The data set is given in the file `diabetes-dataset-1k.csv`. The table below gives the description for some variables in this study.

| Variable | Description |
|---|---|
| `age` | person's age (in years) |
| `hypertension` | 0 = No, 1 = Yes |
| `heart_disease` | 0 = No, 1 = Yes |
| `bmi` | the Body Mass Index |
| `HbA1c_level` | Hemoglobin A1c level (%) |
| `blood_glucose_level` | blood glucose level (mg/dL) |
| `diabetes` | 0 = No, 1 = Yes |

1. Load the data set into R and name it as **data**. Run the command **set.seed(1101)**.

   Then, write code to randomly split **data** into two subsets, one set with 800 rows, to be named as **train.set** and other set with the rest of rows, to be named as **test.set**.

2. Let **m** denote a vector of possible values for the argument `minsplit` which ranges from 1 up to 50 inclusively.

   For each value in **m**, write code to

   (i) fit a decision tree using **train.set**;

   (ii) use the fitted tree to predict the diabetes status of people in **test.set**;

   (iii) derive the precision and the accuracy for the prediction in (ii).

3. Consider `minsplit = 50`, write code to fit a decision tree for **train.set**, to be named as **DT**. Write code to plot the ROC curve of the tree **DT** in black color based on its prediction for **test.set**. Derive and report the value of AUC.

4. A naive Bayes classifier using **train.set**. Write code to fit the classifier, to be named as **NB**.

5. Write code to plot the ROC curve of the classifier **NB** in blue color based on its prediction for **test.set**. Derive and report the value of AUC.

6. Let $\delta$ denote the threshold used for the ROC curve in Question 5. Write code to plot a figure that shows how the TPR and the FPR change when the threshold $\delta$ changes.

7. Propose a value for the threshold $\delta$ that you think it's good.

8. With the proposed value of $\delta$ in Question 7, write code to calculate the accuracy of the classifier **NB** when it is used to predict the diabetes status for **test.set**.

9. Write code to fit a logistic regression model using **train.set**, to be named as **LR**, by including all the given input features.

10. Using the model **LR**, compute and report the odds ratio of having diabetes between a person that has hypertension and a person without hypertension, given that they both have the same values for the other features.

11. Write code to plot the ROC curve of the model **LR** in red color based on its prediction for **test.set**. Derive and report the value of AUC.

12. Standardize all the quantitative input features in **train.set** and **test.set**, then write code to fit a KNN classifier with $k = 3$ using **train.set** to predict the winning probabilities for **test.set**.

13. From the winning probabilities derived in 12, write code to find the predicted probability of having diabetes for each person in **test.set**.

14. With the proposed value of $\delta$ in Question 7, write code to calculate the accuracy of the 5-NN classifier above when it is used to predict the diabetes status for **test.set**. Report the accuracy.