Project (Group of one-three)

R Assignment Due March, 30th (9 am)

1. Select a dataset from the following source.

   http://vincentarelbundock.github.io/Rdatasets/datasets.html (source 1)

   https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table (source 2)

   https://www.kaggle.com/datasets (source 3)

   http://www.statsci.org/data/multiple.html (source 4)

   http://ww2.amstat.org/publications/jse/jse_data_archive.htm (source 5)

   http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets?CGISESSID=10713f6d891653ddcbb7ddbdd9cffb79 (source 6)

   https://r-dir.com/reference/datasets.html (source 7)

   Note: If you would like to use the other source, please contact me before using it.

Requirement for the selected dataset

- The dataset you select cannot be the same as the dataset selected earlier by your classmate.
- The response must be a binary variable. Note: If your response has more than two levels, you may choose to combine several levels together.
- There must be at least 5 predictors with at least 1 categorical predictor and at least 1 quantitative predictor.
- You need to post details of your selected dataset via LINE by Thursday, February 23rd at 9 am including team members, Data name, the source (URL), number of predictors, number of categorical predictors, number of quantitative predictors and number of records.

2. Utilizing data visualization along with the following three techniques with your dataset.

(i) kNN
(ii) Naïve Bayes
(iii)Classification tree

3. Objective: Compare the performances of the three methods and give the recommendation which method is the most suitable for your selected dataset.

4. Necessary steps

   1. Provide necessary information of the dataset, e.g. source, variables, and so on. Also, specify the objective of your report and the reason why you are interested in this dataset.
   2. Perform data visualization, kNN, classification tree and naive Bayes.
   3. Summarize information and knowledge you gain from this study.

   Note: You need to add R codes for all three techniques and data visualization above.
   Layout for each plot or technique: R-code → Results → Discussion

5. On Thursday 30/03, you will have to submit
   a) a hard copy of your project (in class at 9 am) and a file for your project via email (by the class time),
   b) the original data (in CSV format) and CSV files you use for kNN, naïve Bayes, and classification tree (by the class time). If you use the same data file for different methods, please explain it in your email.
   Sending everything through email: pannapa.cha@mahidol.edu.

Recommendations

Naïve Bayes and Classification Tree

1. The original dataset can be used directly.
2. Need to factor categorical predictors that use numbers to represent categories.

kNN

1. The csv file for kNN should not contain any categorical predictors coded by text since the code may not run properly.

Note: I recommend you to separate files for different methods since this will be easier for you to run codes successfully.
