Statistical Modelling & Machine Learning Final Exam
(9:00 - 10:15AM 12/9/2021, Thursday)

- **Instruction:**
  - Solve all 3 problems.
  - There should be only `R` code for your final model/method in your `R` code file.
  - Upload a `.R` file with your code and answers for the problems on I-campus.

1.[15pts] Consider '`SMQ1.csv`' data file and load the file on your `R` using `read.csv('SMQ1.csv')`. Suppose that we want to predict `wage` of workers based on 6 input variables as follows:

- `year`: Year that wage information was recorded.
- `age`: Age of worker.
- `maritl`: A factor with levels '`1. Never Married`', '`2. Married`', '`3. Widowed`', '`4. Divorced`', and '`5. Separated`' indicating marital status.
- `race`: A factor with levels '`1. White`', '`2. Black`', '`3. Asian`' and '`4. Other`' indicating race.
- `education`: A factor with levels '`1. < HS Grad`', '`2. HS Grad`, '`3. Some College`', '`4. College Grad`', and '`5. Advanced Degree`' indicating education level.
- `jobclass`: A factor with levels '`1. Industrial` and '`2. Information` indicating type of job.

However, there are missing values for all variables including `wage` variable.

(1) Investigate whether there is a specific pattern of missing values in `wage` variable. Give the statistical evidence for your answer.

(2) From part (1), if there is a specific pattern of missing values in `wage` variable, describe the characteristics of workers with a missing `wage` value.

(3) In our prediction problem of `wage`, explain how we should handle the missing values of `wage`. Also, explain what effect or problem we can expect from your method to handle the missing values of `wage`.

2.[15pts] Consider `SMQ2train.csv` file for training set and `SMQ2test.csv` file for test set. In the dataset, there are a binary `Y` variable and 100 continuous input variables (`X1 - X100`).

(1) Visualize effectively two groups of **training data points** on a 2-dimensional plot (i.e., the plot should show almost two distinctive groups; You should use different colors or symbols for the two groups).

(2) Based on the result of part (1), build your best predictive model and compute the misclassification error rate for the test set.

Instruction for Q2:

- Use the test set only for calculating the misclassification error rate (Do not use the test set for model selection or decision of tuning parameter values).

3.[20pts] Consider '`SMQ3train.csv`' for training set and '`SMQ3test.csv`' file for test set. The datasets have information of credit card transactions. Suppose that the credit card company wants to detect fraudulent credit card transactions based on transaction information. In the dataset, there are an output variable `Class` and 30 input variables as follows:

- `Time`: Number of seconds elapsed between this transaction and the first transaction in the dataset.
- `V1 - V28`: Variables transformed from transaction and credit card holder information.
- `Amount`: Transaction amount.
- `Class`: 0: Normal transaction, 1: Fraudulent transaction.

Build your best model using the '`SMQ3train.csv`' (training set) data file, and then apply it to '`SMQ3test.csv`' (test set) data file. Report $\mathbf{F-measure}$ for the test set (i.e., try to attain the highest test F-measure value).

Instruction for Q3:

- When you build your model, use `set.seed(1)` to obtain constant results.
- Use the test set only for calculating the F-measure (Do not use test set for model selection or decision of tuning parameter values).
- Both your model building process and F-measure will be evaluated.