STAT3010/6075 Statistical Methods in Insurance

Assignment 2

- This assignment is worth 10% of the overall mark for STAT3010/6075.
- The deadline for submission is **16.00 on Thursday 5 May 2022**.
- Standard University policies and procedures will be followed for late submission, extensions and academic integrity (see the Module Outline for details).
- Submission is via Blackboard. You must submit a report of at most six pages (in pdf format), containing your answers, **and** a separate R script, containing the code that you used to obtain your results.
  - You should submit your report via TurnitinUK on Blackboard (see Module Outline for details) in a file called `report-`*ID*`.pdf`, where *ID* is your student ID number, for example `report-12345678.pdf`. In the Assignments folder, click on View/Complete to submit your report. Please enter this file name as the Assignment Title.
  - You should not include R code used in your analysis in your report, but you must submit a separate R script via Blackboard containing your code called `code-`*ID*`.R`, for example `code-12345678.R`. Please rename and use the R template `code-yyy.R` provided. In the Assignments folder, click on Assignment 2 code submission to submit your code. Please enter this file name as the Link Title.
  - Please start your R script with the command `set.seed(`*ID*`)`, for example `set.seed(12345678)`.
- The page limit is strict and is easily sufficient to receive full credit. If your report is more than six pages of A4, only the first six pages will be marked.

Recall from Assignment 1 that a health insurance company is developing a model to assess the risk of its policy holders having diabetes based on the following data from the file `diabetes.csv`:

`Diabetes`	Binary variable indicating diabetes diagnosis, either positive (*pos*) or negative (*neg*)
`Age`	Age of individual, recorded in years
`BMI`	Body mass index (weight in kg/(height in m)^2)
`Glucose`	Plasma glucose concentration
`Pressure`	Diastolic blood pressure (mm Hg)
`Pregnant`	Number of times pregnant

Take the first 450 observations as the training data set and the remaining 274 observations as the test data set.

1. Calculate the diabetes rate in the test and training data sets, and hence calculate the classification rate of the naïve classifier. Comment on the usefulness of this classifier for identifying cases of diabetes.

   [4 marks]

2. Fit a logistic regression model to predict `Diabetes` from `Age`, `BMI`, `Glucose`, `Pressure` and `Pregnant` using the training data set and calculate its classification rate using the test data set.

   [4 marks]

3. Fit ridge regression models with $\lambda = 0.1, 0.2, 0.3$ and 0.4 to predict `Diabetes` from `Age`, `BMI`, `Glucose`, `Pressure` and `Pregnant` using the training data set and calculate their classification rates using the test data set.

   [8 marks]

4. Fit logistic regression models using LASSO with $\lambda = 0.01, 0.02, 0.03$ and 0.04 to predict `Diabetes` from `Age`, `BMI`, `Glucose`, `Pressure` and `Pregnant` using the training data set and calculate their classification rates using the test data set.

   [8 marks]

5. Calculate the classification rates on the test data set for the K-nearest neighbours classifiers with $K = 1$ to 15 to predict `Diabetes` from `Age`, `BMI`, `Glucose`, `Pressure` and `Pregnant` trained on the training data set.

   [8 marks]

6. Produce a classification tree to predict `Diabetes` from `Age`, `BMI`, `Glucose`, `Pressure` and `Pregnant` grown on the training data set.

   [4 marks]

7. The `R` function `predict` can be used on a classification `tree` to classify new observations contained in a dataframe `unseen`: `predict(tree, unseen, type="class")`. Use this function to calculate the classification rate for the tree produced in part 6.

   [4 marks]

8. Which of the above classifiers would you recommend the company uses? Justify your answer. Start by selecting a value for λ for the ridge regression model and logistic regression model using LASSO, and a value for K for the K-nearest neighbours classifier.

   [10 marks]