FIT3152 Data analytics: Assignment 2

This assignment is worth 20% of your final marks in FIT3152.

Due:	Friday 21st May 2021 11:55pm GMT+10
How to submit:	Submit your written report as a single pdf with R code pasted in as machine-readable text as an appendix, or as an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and print off as a pdf and submit. Whichever method you choose, you will submit a single pdf, and your R code will be machine readable text. Use the naming convention: Firstname.Lastname.studentID.pdf. Upload the file to Moodle. Do not zip. Do not submit your data file.

Objective:

The objective of this assignment is to gain familiarity with classification models using R.

You will be using a modified version of the Kaggle competition data: Predict next-day rain in Australia. https://www.kaggle.com/jsphyg/weather-dataset-rattle-package, but predicting whether or not the following day will be cloudy. The data contains a number of meteorological observations as attributes, and the class attribute "CloudTomorrow". Details of the decision attributes follow the assignment description.

You are expected to use R for your analysis, and may use any R package. Clear your workspace, set the number of significant digits to a sensible value, and use 'WAUS' as the default data frame name for the whole data set. Read your data into R using the following code:

```
rm(list = ls())
WAUS <- read.csv("CloudPredict2021.csv")
L <- as.data.frame(c(1:49))
set.seed(88888888) # Your Student ID is the random seed
L <- L[sample(nrow(L), 10, replace = FALSE),] # sample 10 locations
WAUS <- WAUS[(WAUS$Location %in% L),]
WAUS <- WAUS[sample(nrow(WAUS), 2000, replace = FALSE),] # sample 2000 rows
```

We want to obtain a model that may be used to predict whether it is going to be cloudy tomorrow for 10 locations in Australia.

Assignment questions:

1. Explore the data: What is the proportion of cloudy days to clear days.? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-valued attributes. Is there anything noteworthy in the data? Are there any attributes you need to consider omitting from your analysis? (1 Mark)

2. Document any pre-processing required to make the data set suitable for the model fitting that follows. (1 Mark)

3. Divide your data into a 70% training and 30% test set by adapting the following code (written for the iris data). Use your student ID as the random seed.

```
set.seed(XXXXXXXX) #Student ID as random seed
train.row = sample(1:nrow(iris), 0.7*nrow(iris))
iris.train = iris[train.row,]
iris.test = iris[-train.row,]
```

4. Implement a classification model using each of the following techniques. For this question you may use each of the R functions at their default settings, or with minor adjustments to set factors etc. (5 Marks)

   • Decision Tree • Naïve Bayes • Bagging • Boosting • Random Forest

5. Using the test data, classify each of the test cases as 'cloudy tomorrow' or 'not cloudy tomorrow'. Create a confusion matrix and report the accuracy of each model. (1 Mark)

6. Using the test data, calculate the confidence of predicting 'cloudy tomorrow' for each case and construct an ROC curve for each classifier. You should be able to plot all the curves on the same axis. Use a different colour for each classifier. Calculate the AUC for each classifier. (1 Mark)

7. Create a table comparing the results in Parts 5 and 6 for all classifiers. Is there a single "best" classifier? (1 Mark)

8. Examining each of the models, determine the most important variables in predicting whether or not it <u>will be cloudy</u> tomorrow. Which variables could be omitted from the data with very little effect on performance? Give reasons. (2 Marks)

9. Starting with one or some of the classifiers you created in Part 4, create a classifier that is simple enough for a person to be able to classify whether it will be cloudy or not tomorrow by hand. Describe your model, either with a diagram or written explanation. How well does your model perform, and how does it compare to those in Part 4? What factors were important in your decision and why you chose the attributes you used. (2 Marks)

10. Create the best tree-based classifier you can. You may do this by adjusting the parameters, and/or cross-validation of the basic models in Part 4, or using an alternative tree-based learning algorithm. Show that your model is better than the others using appropriate measures. Describe how you created your improved model, and why you chose that model. What factors were important in your decision and why you chose the attributes you used. (2 Marks)

11. Using the insights from your analysis so far, implement an Artificial Neural Network classifier and report its performance. Comment on attributes used and your data pre-processing required. How does this classifier compare with the others? Can you give any reasons? (2 Marks)

12. Write a brief report (suggested length 6 pages) summarizing your results in parts 1 – 10. Use commenting (# ----) in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown. (2 Marks)

Description of the data:

Attributes 1:3, Day, Month, Year of the observation

Attribute 4, Location: the location of the observation

Attribute 5, MinTemp: the daily minimum temperature in degrees celsius

Attribute 6, MaxTemp: the daily maximum temperature in degrees celsius

Attribute 7, Rainfall: the rainfall recorded for the day in mm

Attribute 8, Evaporation: the evaporation (mm) in the 24 hours to 9am

Attribute 9, Sunshine: hours of bright sunshine over the day.

Attribute 10, WindGust: direction of the strongest wind gust over the day.

Attribute 11, WindGustSpeed: speed (km/h) of the strongest wind gust over the day.

Attribute 12, WindDir9am: direction of the wind at 9am

Attribute 13, WindDir3pm: direction of the wind at 3pm

Attribute 14, WindSpeed9am: speed (km/hr) averaged over 10 minutes prior to 9am

Attribute 15, WindSpeed3pm: speed (km/hr) averaged over 10 minutes prior to 3pm

Attribute 16, Humidity9am: humidity (percent) at 9am

Attribute 17, Humidity3pm: humidity (percent) at 3pm

Attribute 18, Pressure9am: atmospheric pressure (hpa) reduced to mean sea level at 9am

Attribute 19, Pressure3pm: atmospheric pressure (hpa) reduced to mean sea level at 3pm

Attribute 20, Temp9am: temperature (degrees C) at 9am

Attribute 21, Temp3pm: temperature (degrees C) at 3pm

Attribute 22, RainToday: boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

Attribute 23, CloudyTomorrow: the target variable. If the average fraction of sky obscured by cloud at 9am and 3pm is greater than 4 "oktas", the day is classified as cloudy. Otherwise the day is classified as not cloudy. (Oktas record how many eigths of the sky are obscured by cloud.)