FIT3152 Data analytics: Assignment 2

This assignment is worth 10% of your final marks in FIT3152.

Due: Sunday May 26th 2019

Note: Students are expected to work individually on this assignment.

How to submit: Submit your written report as a pdf file (.pdf). and R working as an R
                    script (.R), *or*
                    Submit your report comprising both written answers and script as an R
                    Markdown file in HTML format (.html).

Use the naming convention: Firstname.Lastname.studentID.{pdf, R, .html} Upload the one or
two files to Moodle. Do not zip. Do not submit the 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 rain tomorrow
in Australia. https://www.kaggle.com/jsphyg/weather-dataset-rattle-package The data
contains a number of meteorological observations as attributes, and the class attribute "Rain
Tomorrow". Details of the decision attributes follow the assignment description.

You are expected to use R for your analysis, and may use any R package. Set your R working
directory to 'desktop', clear the 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("WAUS2019.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 rain tomorrow
for 10 locations in Australia.*

Assignment questions:

1.   Explore the data: What is the proportion of rainy days to fine 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 may need to consider omitting from your analysis for any reason?
        Document any pre-processing of the data required based on your exploration.

2.   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(XXXXXXX) #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,]
```

3. Create 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.

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

4. Using the test data, classify each of the test cases as 'will rain tomorrow' or 'will not rain tomorrow'. Create a confusion matrix and report the accuracy of each model.

5. Using the test data, calculate the confidence of predicting 'will rain 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.

6. Create a table comparing the results in parts 4 and 5 for all classifiers. Is there a single "best" classifier?

7. Examining each of the models, determine the most important variables in predicting whether or not it will rain tomorrow. Which variables could be omitted from the data with very little effect on performance? Give reasons.

8. By experimenting with parameter settings for at least one of the classifiers, create the best classifier you can – that is, one with an accuracy greater than the models you originally created in Part 3. Demonstrate this improved accuracy using ROC, AUC, or other accuracy measures. Report the parameter settings and assumptions made in designing this classifier.

9. 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?

10. Write a brief report (4 pages max) summarizing your results in parts 1 – 9. Use commenting (# ----) in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown.

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, Cloud9am: fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

Attribute 21, Cloud3pm: fraction of sky obscured by cloud at 3pm.

Attribute 22, Temp9am: temperature (degrees C) at 9am

Attribute 23, Temp3pm: temperature (degrees C) at 3pm

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

Attribute 25, RainTomorrow: the target variable. Did it rain tomorrow?