MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2022: Assignment 2

Your task	● The objective of this assignment is to gain familiarity with classification models using R. ● This is an individual assignment.
Value	● This assignment is worth **20%** of your total marks for the unit. ● It has 20 marks in total.
Suggested Length	● 4 – 6 A4 pages (for your report) + extra pages as appendix (for your code) ● Font size 11 or 12pt, single spacing
Due Date	**11.55pm Friday 20th May 2022**
Submission	● PDF file only. Naming convention: *FirstnameSecondnameID.pdf* ● Via Moodle Assignment Submission. ● Turnitin will be used for similarity checking of all submissions.
Late Penalties	● 10% (2 mark) deduction per calendar day for up to one week. ● Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Instructions and data

The objective of this assignment is to gain familiarity with classification models using R.
We want to obtain a model that may be used to predict whether tomorrow will be warmer than today for 10 locations in Australia.

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 meteorological observations as attributes, and the class attribute “Warmer Tomorrow”.

There are two options for compiling your report:
(1) You can submit a single pdf with R code pasted in as machine-readable text as an appendix, or
(2) 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.

Regardless of which method you choose, you will submit a single pdf, and your R code will be machine readable text. We need to conform to this format as the university now requires all student submission to be processed by plagiarism detection software.

Submit your report as a single PDF with the file name ***FirstnameSecondnameID.pdf*** on Moodle.

Creating your data set

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 and create your individual data using the following code:

```
rm(list = ls())
WAUS <- read.csv("WarmerTomorrow2022.csv")
L <- as.data.frame(c(1:49))
set.seed(XXXXXXXX) # 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
```

Questions

1. Explore the data: What is the proportion of days when it is warmer than the previous day compared to those where it is cooler? 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 if suitable. **(5 Marks)**

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

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

6. Using the test data, calculate the confidence of predicting 'warmer 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 it will be warmer tomorrow or not. Which variables could be omitted from the data with very little effect on performance? Give reasons. **(2 Marks)**

9. Starting with one 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 warmer tomorrow or not 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? State 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? State why you chose the attributes you used. **(3 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 – 11. Use commenting in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown. **(1 Mark)**

Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but please document these in your report and include in your R code.

Description of the data

Attributes 1-3, Day, Month, Year	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	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, WindGustDir	Direction of 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, WarmerTomorrow	The target variable. Will tomorrow be warmer than today?