ASSESSMENT TASK 2 (PROBLEM SOLVING)

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task. Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work. Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions:

https://www.deakin.edu.au/students/study- support/referencing/academic-integrity

---

Using aggregation functions for data analysis

The provided zip file contains the data file [Forest_*Forest_2021.txt*] and the R code [*AggWaFit718.R*]

to use with the following tasks, include these in your R working directory.

Total Marks 100, Weighting 20%

Forest Fires Data Set

In order to predict the burned area of forest fires ("UCI Machine Learning Repository: Forest Fires Data Set", 2017), in the northeast region of Portugal ("Montesinho.Com - Nature Tourism In Montesinho Natural Park", 2017), analysis of the meteorological and other data is required (see details at "Forest Fires Dataset", 2017).

The first four variables of the original dataset are categorical, the remaining nine variables are numerical.

In this assignment you will be using a modified version of the original dataset, with numerical variables only, **Fires_2021.txt**. You can always refer to the original data for deeper understanding of the data.

Attribute (Numerical Variable) Information:

X1: FFMC - FFMC index from the FWI system: 18.7 to 96.20 (Happe, 2017)

X2: DMC - DMC index from the FWI system: 1.1 to 291.3 (Happe, 2017)

X3: DC - DC index from the FWI system: 7.9 to 860.6 (Happe, 2017)

X4: ISI - ISI index from the FWI system: 0.0 to 56.10 (Happe, 2017)

X5: temp - temperature in Celsius degrees: 2.2 to 33.30

X6: RH - relative humidity in %: 15.0 to 100

X7: wind - wind speed in km/h: 0.40 to 9.40

X8: rain - outside rain in mm/m2 : 0.0 to 6.4

Y: area - the burned area of the forest (in ha): 0.00 to 1090.84

For more information, read (Cortez, 2007)

Assignment tasks

T1. Understand the data **[10 marks]**

(i) Download the txt file (Fires_2021.txt) from CloudDeakin and save it to your R working directory.

(ii) Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.table("Fires_2021.txt"))
```

The variable of interest is **X1**: FFMC - FFMC index from the FWI system: 18.7 to 96.20 (Happe, 2017).

To investigate **X1**, generate a subset of 300 with numerical data e.g. using:

```
my.data <- the.data[sample(1:517,300) c(1:9)]
```

This would give you a new dataset with 300 rows and 9 columns.

Using scatter plots and histograms, report on the general relationship between each of the variables and your variable of interest **X1**. (You should build 8 scatter plots and 9 histograms. Include the plot(s) and 1 or 2 sentences for each of the eight variables and the variable of interest **X1**. You can also investigate possible associations between the other eight variables).

T2. Transform the data **[20 marks]**

(i) Choose **any four** variables from **X2, X3,X4,X5,X6,X7, X8** and **Y.** Make appropriate transformations so that the values can be aggregated in order to predict the *variable of interest* **X1** (FFMC index).

Assign your *transformed* data along with your *transformed* variable of interest to an array (it should be 300 rows and 5 columns). Save it to a txt file titled "name-transformed.txt".

```
write.table(your.data,"name-transformed.txt",)
```

(ii) Briefly explain the general relationship between each of your transformed variables and your variable of interest **X1** (FFMC index). (2-3 sentences each)

T3. Build models and investigate the importance of each variable. **[20 marks]**

(i) Download the AggWaFit.R file (from CloudDeakin) to your working directory and load into the R workspace using,

```
source("AggWaFit718.R")
```

(ii) Use the fitting functions to learn the parameters for

a. A weighted arithmetic mean,

b. Weighted power means with $p = 0.5$, and $p = 2$,

c. An ordered weighted averaging function, and

d. A Choquet integral.

(iii) Include two tables in your report - one on the error measures, and one summarising the weights/parameters that were learned for your data.

(iv) Compare and interpret the data in your tables. Be sure to comment on

a. How good the model is,

b. The importance of each of the variables (the four variables that you have selected),

c. Any interaction between any of those variables (are they complementary or redundant?) and

d. Better models favour higher or lower inputs. (1-3 paragraphs)

T4. Use your model for prediction. **[20 marks]**

(i) Using your best fitting model, predict the area for the following input

X1=95.9**;** **X2**=158; **X3**=633.6; **X4**=11.3; **X5**=27.5; **X6**=29; **X7**=4.5; **X8**=0.0, **Y**=43.42.

You should use the same pre-processing as in Task 2.
Give your result and comment on whether you think it is reasonable.
To be able to compare you are given the measured value of **X1** for this set of measurement. (1-2 sentences)

(ii) Comment generally on the ideal conditions (in terms of your chosen four variables) under which a large FFMC index will result (list all your assumptions) . (1-2 sentences).

T5.* Comparing with a linear regression model **[20 marks] (This task is for those aiming for HD)**

Linear regression is used to predict the value of an outcome variable **X1** based on one or more input predictor variables **X**i, i=2,…,n.

The equation is

$$X_1 = \beta_0 + \beta_1 Y + +\beta_2 X_2 + \cdots \beta_n X_n + \varepsilon.$$

[*Use the same pre-process as T2*]

The built-in function lm() is used to fit linear models in R.

Build your linear model using the same dataset in task **T2** and describe the summary statistics for your model using the function summary().

(i) Compare the performance of the linear model you got with your best fitting model in T4.

Visualise the predicted Y values of both models on the 300 data and compare them with the true **Y** values.

(ii) Give your comment on the differences between the linear model and your best fitting model. (2-4 sentences).

T6. Summarising your data analysis in a 3-minutes presentation **[10 marks]**

Using a simple and accessible platform such as YouTube or PowerPoint, create a 3 minutes presentation that summarises your data analysis procedures, findings, implications, and the limitations of the model you used.

For **referencing**, follow the Harvard style:

https://www.deakin.edu.au/students/studying/study-support/referencing/harvard

SUBMISSION:

Submit to the **SIT718 CloudDeakin Dropbox**. Your final submission must include the following **FOUR** files:

1. A PDF report, "**name-report.pdf**", covering all of the items in above

(where “name” is replaced with your name -you can use your surname or first name).

The total report must be **up to 8 pages** (for everything) including a cover page which contains your full name and student ID. Penalties for oversized assignments apply: the pages after page 8, these pages will not be marked. Thus, strictly keep to page limit of 8 pages.

Minimal font size 11pt, figures should be of appropriate size, easy to read, well-annotated, with captions.

Marks will be deducted for figures which are too small, not annotated and without captions.

Check your text for spelling and grammatical mistakes. Marks will be deducted for spelling and grammatical errors.

2. A data file named "**name-transformed.txt**" - just to help us distinguish them!).

3. Presentation recordings or slides with audio (a link to YouTube/Dropbox is acceptable)

4. The R code file (that you have written to produce your results) named "**name-code.R**" (where “name” is replaced with your surname or first name). A data file named "**name-transformed.txt**" (just to help us distinguish

them!).

References

"UCI Machine Learning Repository: Forest Fires Data Set". *Archive.ics.uci.edu*. N.p., 2017, http://archive.ics.uci.edu/ml/datasets/Forest+Fires, 29 Apr. 2017.

"Forest Fires Dataset". *Dsi.uminho.pt*. N.p., 2017, www.dsi.uminho.pt/~pcortez/forestfires, 29 Apr. 2017.

Cortez, P. and Morais, A.D.J.R., 2007. A data mining approach to predict forest fires using meteorological data, http://www3.dsi.uminho.pt/pcortez/fires.pdf.

Happe, Harry. "Meteomalaga". *Malagaweather.com*. N.p., 2017. Web. 29 Apr. 2017.

"Montesinho.Com - Nature Tourism In Montesinho Natural Park". *montesinho.com*. N.p., 2017. https://www.montesinho.com/en, 29 Apr. 2017.