RMIT University
Computer Science & IT, School of Science

# COSC 2789 — Practical Data Science

## Assignment 2: Data Modelling and Presentation

Due: 23:59, Friday (13th, September). (week 4)

This assignment is worth 40% of your overall mark.

## Assignment Teams

**This assignment should be carried out in groups of *two*.**

It is up to you to form a team.

Once you have formed your team, you should register using the form at:

`https://goo.gl/forms/KQixgtipMcTX5Wxa2`.

**Important: you must register your team within 1 week** at the latest.

If you have strong reasons for needing to complete the assignment individually, you may apply to do so by sending an email to the lecturer, explaining your reasons. However, bear in mind that the requirements and available marks will be the same as for pairs, so you are strongly advised to work in a team.

In addition, please submit what percentage each partner made to the assignment (a contribution sheet will be made available for you to fill in), and submit this sheet in your submission. The contributions of your group should add up to 100%. If the contribution percentages are not 50-50, the partner with less than 50% will have their marks reduced. Let student A has contribution X%, and student B has contribution Y%, and $X > Y$. The group is given a group mark of M. Student A will get M for assignment 1, but student B will get $\frac{M}{\frac{X}{Y}}$.

## Introduction

This assignment focuses on *data modelling*, a core step in the data science process. You will need to develop and implement appropriate steps, in IPython, to complete the corresponding tasks.

This assignment is intended to give you practical experience with the typical 5th and 6th steps of the data science process: *data modelling*, and *presentation and automation*.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis – it is your responsibility to stay informed with regards to any announcements or changes. Login

through `https://rmit.instructure.com/`.

# Where to Develop Your Code

You are encouraged to develop and test your code in two environments: **Jupyter Notebook on Lab PCs** and **Teaching Servers**.

### Jupyter Notebook on Lab PCs

On Lab Computer, you can find Jupyter Notebook via:

Start → All Programs → Anaconda2 (64-bit) → Jupyter Notebook

Then,

- Select New → Python 2
- The new created '*.ipynd' is created at the following location:
  - C:\Users\sXXXXXXX
  - where sXXXXXXX should be replaced with a string consisting of the letter "s" followed by your student number.

# Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. For further information, please see the *Academic Integrity* information at `http://www1.rmit.edu.au/academicintegrity`.

# General Requirements

This section contains information about the general requirements that your assignment must meet. *Please read all requirements carefully before you start.*

- You *must* do all modelling in IPython.

- You *must* include a plain text file called "readme.txt" with your submission. This file should include your name(s) (if you are a group of two) and student ID(s), and instructions for how to execute your submitted script files. This is important as *automation* is part of the 6th step of data science process, and will be assessed strictly.
- Parts of this assignment will include a written report, this *must* be in *PDF* format.
- Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is `gryphon`, then that is exactly the file name you should submit; `Gryphon`, `GRYPHON`, `griffin`, and anything else but `gryphon` will be rejected.

## Task 1: Data Retrieving (3%)

This assignment will focus on data modelling, and you can choose to focus on one approach: ***Classification*** or ***Clustering***.

For this assignment, you need to select **one** suitable dataset, from the following options:

1. Find and then analyse your own data set, in a domain that is of interest to you. If you choose this option, you will need to:
    - include a detailed description of the data in your report in Task 4, and describe each attribute of it, including the type, the range of possible values, whether it contains any missing values/errors
    - submit a copy of the dataset, to allow the assessment of your modelling result.
2. Select one data set from the UCI Repository: `http://archive.ics.uci.edu/ml/`. Choose one dataset from either the *Classification* or *Clustering* task.

NOTE: Please do not use any dataset that we used in the teaching (including lectures, tuteLabs and Assignment 1) of this course.

Being a careful data scientist, you know that it is vital to set **the goal of the project**, then **thoroughly pre-process** any available data (each attribute) before starting to analyse and model it. In your report in Task 4, You need to clearly state the goal of your project, and the design/steps of pre-processing your data.

Please ensure you understand the data you selected, including the meaning of each attribute. For datasets from the UCI repository, you can obtain this information from the corresponding Web page under the sections *Data Set Information* and *Attribute Information*.

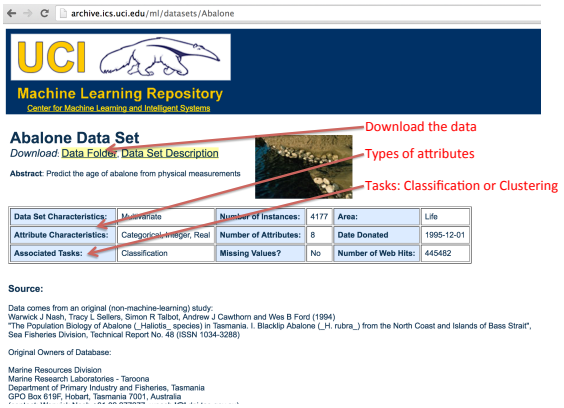


Figure 1: Example of the Abalone data set with instructions: where to download the data, the types of the attributes in the data, and the suitable task(s) of this data.

## Task 2: Data Exploration (5%)

Explore the selected data, carrying out the following tasks:

- Explore each column (or at least 10 columns if there are more than 10 columns), using appropriate descriptive statistics and graphs (if appropriate), e.g. the distribution of a numerical attribute, the proportion of each value of a categorical attribute. For each explored column, please think carefully and report in your report in Task 4): 1) the way you used to explore a column (e.g. the graph); 2) what you can observe from the way you used to explore it.

  (Please format each graph carefully, and use it in your final report. You need to include appropriate labels on the $x$-axis and $y$-axis, a title, and a legend. The fonts should be sized for good readability. Components of the graphs should be coloured appropriately, if applicable.)

- Explore the relationship between all pairs of attributes (or at least 10 pairs of attributes, if there are more in the data), and show their relationship in an appropriate graph. You may choose which pairs of columns to focus on, but you need to generate a visualisation graph for each pair of attributes. Each of the attribute pair should address a **plausible hypothesis** for the data concerned. In your report, for each plot (pair of attributes), state the hypothesis that you are investigating. Then, briefly discuss any interesting relationships (or lack of relationships) that you can observe from your visualisation.

## Task 3: Data Modelling (10%)

Model the data by treating it as **either** a *Classification* or *Clustering* Task, depending on which dataset you previously selected.

You must choose **two** models within the particular Task category (i.e. two Classification models, or two Clustering models), and carry out the following steps for *each* model:

- Select the appropriate model (e.g. $DecisionTree$ for classification) from *sklearn*.
- If you choose to do a *Classification* Task,
  - Split the data into *training* set and the *test* set. Specifically, please split the data at the following ratio:
    - 50% for training and 50% for testing;
    - 60% for training and 40% for testing;
    - 80% for training and 20% for testing;
  - For each of the training/testing split, perform the following steps:
    - Train the model by selecting appropriate values for each parameter in the model.
      - You need to *show* how do you choose this value, and *justify* why you choose it (for example, $k$ in the $KNearestNeighbor$ model).
    - Test the accuracy of the model on the *test* set, and report the performance of the model in the following terms:
      - Confusion Matrix
      - Classification Error Rate
      - Precision
      - Recall
      - F1-Score
- If you choose to do a *Clustering* Task,
  - Train the model by selecting appropriate values for each parameter in the model.
    - *Show* how do you choose this value, and *justify* why you choose it (for example, $k$ in the $k$-means model).
  - **Determine the optimal number of clusters**
  - Evaluate the performance of the clustering model by:
    - Checking the clustering results against the true observation labels
    - Constructing a "confusion matrix" to analyse the meaning of each cluster by looking at the majority of observations in the cluster. (You can do this by using a pen and a piece of paper, as we did in Practical Exercise 3 in Tute/Lab 06 (week7); if you prefer, you can also explore how to do this step directly in IPython.)

After you have built two Classification models, or two Clustering models, on your data, the next step is to **compare** the models. You need to include the results of this comparison, including a recommendation of which model should be used, in your report (see next section).

## Task 4: Report (12%)

Write your report and save it in a file called `report.pdf`, and it must be in PDF format, and must be **at most 12 (in single column format) pages (including figures and references) with a font size between 10 and 12 points** Penalties will apply if the report does not satisfy the requirement. Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs.

Your report must have the following structure:

- A cover page, including
  - Title
  - Author (your name(s))
  - Affiliations
  - Contact details
  - Date of report
- Table of Content
- An abstract/executive summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- References

Please revisit p. 51-57 in the Week1 lecture slides.

## Task 5: Presentation (10%)

You will be required to do a presentation for your assignment 2 in the last session of the course:

- The presentation should
  - explain the goal of the project.

  - briefly describe your chosen data set.
  - describe the data preparation steps.
  - state the hypotheses/questions that you were investigating.
  - explain what the modelling steps are, and what the results are.
  - show the final conclusion and recommendation.
- The presentations are a maximum of 5 minutes per group, and we suggest each group to have at most 5 slides.

# What to Submit, When, and How

The assignment is due at

23:59, Friday (13th, September). (week 4).

Assignments submitted after this time will be subject to standard late submission penalties. There are four files you need to submit:

- Notebook file containing your python commands, 'Assignment2.ipynb'.

# For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:

1. Main menu → Kernel → Restart & Run All
2. Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.

- Your `report.pdf` file **at most 12 (in single column format) pages (including figures and references) with a font size between 10 and 12 points**.
- Your presentation slides.
- The "readme.txt": includes your names and student IDs, and instructions for how to execute your submitted script files.

They must be submitted as ONE single zip file, named as your student numbers (for example, 1234567_7321283.zip if your student ID are s1234567 and s7321283). The zip file must be submitted in Canvas:

*Assignments/Assignment 2.*

Please do NOT submit other unnecessary files.