Practical Data Science with Python
COSC 2670/2738
Assignment 2

Assessment Type	individual
Due Date	23:59 on the 21st of May, 2023
Marks	35

Assignment Teams (Please read this carefully before attempting)

In the submission (your PDF file) you will be required to certify that the submitted solution *represents your own work only* by agreeing to the following statement:

> *I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes":*

A sample format for this requirement is provided, and please find it in `Canvas −> Assignments −> Assignment2`.

Introduction

This assignment focuses on *data modelling*, a core step in the data science process. You will need to develop and implement appropriate steps 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 **The Anaconda versions suggested in the course Canvas Announcement**.

Jupyter Notebook on Lab PCs

On Lab Computer, you can find Jupyter Notebook via:

$$\text{Start} \rightarrow \text{All Programs} \rightarrow \text{Jupyter Notebook - Anaconda}$$

Then,

- Select New $\rightarrow$ Python 3

- 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.

Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods

- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source

- Copyright material from the internet or databases

- Collusion between students

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity.

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* include a plain text file called "readme.txt" with your submission. This file should include your name and student ID, 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: Retrieving and Preparing the Data (5%)

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

For this assignment, you need to select **one** dataset from the following options, and then work on it:

1. User Identification From Walking Activity Data Set More details can be found from the following UCI webpage about this dataset: `https://archive.ics.uci.edu/ml/datasets/User+Identification+From+Walking+Activity`

2. User Knowledge Modeling Data Set More details can be found from the following UCI webpage about this dataset: `https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling`

3. Turkish Music Emotion Dataset Data Set More details can be found from the following UCI webpage about this dataset: `https://archive.ics.uci.edu/ml/datasets/Turkish+Music+Emotion+Dataset`

4. Online Shoppers Purchasing Intention Dataset More details can be found from the following UCI webpage about this dataset: `https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset`

Please note that: if the original dataset(s) is(are) not in csv format, please convert them into csv format manually via a text editor or MS spreadsheet.

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.

Task 2: Data Exploration (10%)

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). 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 the relationship in an appropriate graphs. 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.

Please note you do not need to put all the graphs in your report, and you only need to include the representative ones and/or those showing significant information.

Task 3: Data Modelling (10%)

Model the data by treating it as **either** a *Classification* or *Clustering* Task, depending on your choice.

You must use **two** different models (i.e. two Classification models, or two Clustering models), and when building each model, it must include the following steps:

- Select the appropriate features

- Select the appropriate model (e.g. *DecisionTree* for classification) from *sklearn*.

- If you choose to do a *Classification* Task,

    - Train and evaluate the model appropriately.
    - Train the model by selecting the appropriate values for each parameter in the model. You need to show how you choose this values, and justify why you choose it.

- 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, and justify*
    - 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; 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 (10%)

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 for everything (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
    - Statement of the solution representing your own work as required
    - Title
    - Author Information
    - Affiliations
    - Contact details
    - Date of report

- Table of Content

- An abstract/executive summary

- Introduction

- Methodology

- Results

- Discussion

- Conclusion

- References

Please revisit the relevant slides in Week1 lecture if needed.

What to Submit, When, and How

The assignment is due at

23:59 on the 21st of May, 2023.

Assignments submitted after this time will be subject to standard late submission penalties.

The following files should be submitted:

- 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**.

- The "readme.txt": includes your name and student ID, 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.zip if your student ID are s1234567). The zip file must be submitted in Canvas:

Assignments/Assignment 2.

Please do NOT submit other unnecessary files.