RMIT University

Practical Data Science with Python

Assignment 2: Data Modelling

Due: 23:59 on the 8th of October 2023

This assignment is worth 35% of your overall mark.

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. These tasks must be completed individually.

The "Practical Data Science (with Python)" Canvas contains further announcements and a discussion board for this assignment. Please 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/`.

Academic Integrity

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 actions. Plagiarism includes, e.g., submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also considered as plagiarism. All plagiarisms will be penalised; there are no exceptions and no excuses. For further information, please see the RMIT Academic Integrity Policy information at `https://policies.rmit.edu.au/document/view.php?id=168`.

Turnitin will be used for Plagiarism check for this assignment in Canvas.

General Requirements

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

- You *must* do the assignment in **Jupyter Notebook** (available in Anaconda).
- You *must* use the appropriate Python functions (or methods) to complete the tasks and you must figure out (and determine) what functions should be used by yourself. You may refer to relevant Python (or Python Library) official documentations as needed. You may use any Python Library and any function as appropriate.
- 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 not be accepted.

Task 1: Data Preparation and Analysis (10 marks)

In this assignment, you will use the dataset `A2data.csv`, which is available in Canvas under the `Assignments/Assignment 2` section of the course Canvas shell. The dataset is related to attributes and quality of wine.

The inputs include physicochemical tests (e.g. PH values) and the output is wine quality, which scales between 0 (very bad) and 10 (very excellent). There are 12 variables (i.e., attributes or columns) and 4781 instances (i.e., entries or rows). For description of attributes, check the file `Readme-A2data.txt`.

Task 1.1. Load the CSV data from the file (using appropriate Python/Pandas functions). Take a **random** sample (i.e., subset) of **600** instances from the data set, and ensure that these 600 instances don't have any missing values. Write the random sample into a CSV file and name the file as **A2RandomSample.csv**. Use this file (instead of the provided A2data.csv file) for all subsequent tasks in this Assignment, unless otherwise specified.

Task 1.2. Explore the relationship between two variables: **alcohol** and **density**.

- Show the relationship in an appropriate graph (i.e., chart). Describe any interesting relationships (or lack of relationships) that you observe from the visualisation.
- Build a linear model (i.e., Simple Linear Regression) for the two variables, with alcohol being dependent variable and density as independent variable. Present the linear model in the Report and interpret the coefficients of the model.

Task 1.3. Explore the relationship between two variables: **quality** and **alcohol**.

- Create the side-by-side boxplot for alcohol grouped by quality level.
- Summarise your findings based on the boxplot.

Task 2: Classification (10 marks)

Use the random sample data file **A2RandomSample.csv** and complete the following classification (sub)tasks. In Task 2, you need to classify the wine quality based on the other variables.

Task 2.1. Select a model, either k-NN (k-Nearest Neighbours) or Decision Tree. Train and evaluate the model appropriately. Use at least 3 metrics for evaluation.

Task 2.2. Study the impact of at least one key parameter of the model. Describe your findings. Choose the best value(s) for the parameter(s) and justify your choice.

Task 2.3. With the above optimal parameter(s), train and test the model on different training/test data splits: 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20. What is the best train/test split? Why?

Task 3: Clustering (10 marks)

Use the random sample data file **A2RandomSample.csv** and complete the following clustering (sub)tasks. In Task 3, you need to conduct clustering upon the random sample dataset. Don't use the output variable **quality** when building the model or tuning the parameters (i.e., Task 3.1 or Task 3.2).

Task 3.1. Select a model, either k-means or DBSCAN. Build and evaluate the model. Tune the key parameter(s) of the model and justify your choice of the value(s).

Task 3.2. Determine the optimal number of clusters, and justify.

Task 3.3. Analyse the meaning (i.e. predicted quality level) of each cluster by checking the clustering results against the true quality levels (i.e., the variable **quality**). Construct and explain the confusion matrix of the results.

Task 4: Report / Presentation (5 marks)

Write your report and save it in a file called `A2Report.pdf`. 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. Moreover, the quality of the report will be considered, e.g. clarity, grammar mistakes, the flow of the presentation.

Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs. Use any referencing style recommended by RMIT Library (https://www.lib.rmit.edu.au/easy-cite/).

Use the **template** provided in Canvas to structure and format your report.

What to Submit, When, and How

The assignment is due at

23:59 on the 8th of October 2023.

Resubmission without penalty is allowed until the above deadline.

You need to submit the following files:

- **Jupyter Notebook file** containing your Python commands, named `assignment2.ipynb`. **Please use the provided solution template to organise your code**: *assignment2_TEMPLATE.ipynb* (remember to rename the file)

# For the Notebook file, 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 random sample data file named `A2RandomSample.csv`.

- Your report file `A2Report.pdf`.

They must be submitted as **ONE single zip file, named as follows:**

A2_YourStudentNumber.zip

For example, A2_1234567.zip if your student ID is s1234567. The zip file must be submitted in **Canvas**:

Assignments/Assignment 2.

Please do NOT submit other unnecessary files.