RMIT Vietnam University
School of Science, Engineering and Technology

COSC2999 - Practical Data Science with Python

Assignment 2: Data Modelling

Due: 17:00, Sunday the 22nd, December 2024, Week 8
This assignment is worth 45% 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, in Python (Jupyter Notebook), 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.

The "Practical Data Science with Python" 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.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: Jupyter Notebook (or Jupyter Lab) on Lab PCs or your laptop.

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. More information on Academic Integrity is available at https://www.rmit.edu.vn/students/my-studies/assessment-and-results/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 do the analysis in Python Jupyter Notebook (Jupyter Lab).
- 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.
- AI should not be used to generate results and write your assignment. If you use AI tools in developing your solution, please specify (by providing names of the tools or libraries, prompts, and relevant techniques) in the report. Teaching staff reserves the right to run more tests, inspect your code manually, and arrange a face-to-face meeting for a discussion and demo of your solution if needed.

Task 1: Data Preparation and Goal Statement (5%)

This assignment will focus on data modelling, and you need to focus on two approaches: **Classification** and **Regression**. You need to select one dataset from the following options, retrieve and prepare the data for the next tasks in this assignment.

1. Estimation of Obesity Levels Based On Eating Habits and Physical Condition. Details about this dataset can be found from this link to UCI webpage.
2. Air Quality. Details can be found from this link to UCI webpage.

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 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:

2.1. Explore at least 10 columns using appropriate descriptive statistics and graphs (when appropriate). For each explored column, please think carefully and report in your report in Task 4): 1) Why would you select the column? 2) The method(s) you used to explore the column (e.g. the graph); 3) What you can observe from exploring 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.

2.2. Explore the relationship between pairs of attributes (use at least 10 pairs of attributes) and show the 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 pairs 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 (20%)

Model the data by treating it as Classification **and** Regression tasks (both tasks can be applied for each of the given data sets).

You must use **two different models** for each approach (i.e. two classification models and two regression models), and when building each model, it must include the following steps:

- Select appropriate features.
- Select the appropriate model (e.g. *DecisionTree* for classification).
- Train and evaluate the model appropriately.
- Train and evaluate the model by selecting the appropriate values for each parameter in the model. You need to show how you choose these values and justify why you choose them.

After you have built two Classification models and two Regression models on the selected

data, the next step is to **compare** the performance of the models. You need to include the results of this comparison, including a recommendation of which model should be used, in your report (see Task 4).

Task 4: Report (10%)

4.1. Report on the tasks (8%)

Write your report and save it in a file called report.pdf, and it must be in PDF format, and must be **at most 15 (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 requirements. Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your solutions.

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
- Declaration of using AI in the assignment
- Reference.

4.2. Self-reflection (2%)

In this task, you will need to provide a short presentation (max 300 words) of your reflection on what you have learnt from this course on Practical Data Science with Python. You should focus on answering the following questions:

- Have you gained much improved understanding of key concepts and major techniques in practical data science? What is your reflection on the journey (considering now that you have completed your two assignments)? What have you learned from working on exercises and doing both assignments 1 and 2?
- What can be done better by the teaching staff for your studying the course?

Please include this presentation of self-reflection in the PDF file, after the report on the tasks above.

What to Submit, When, and How

The assignment is due at **17:00, Sunday the 22nd, December 2024** (in Week 8).
Assignments submitted after this time will be subject to standard late submission penalties.
You need to submit the following files:

- 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).
- Your **report.pdf** file is at most 15 pages (in single column format, including figures and references) with a font size between 10 and 12 points.
- A "readme.txt" file (if needed) 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 number (for example, 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.