RMIT Vietnam University
School of Science, Engineering and Technology

COSC2999

Practical Data Science with Python

Assignment 3: Group project

Due: 08:00 PM, January 16th, 2025 (week 12)
This assignment is worth 30% of your overall mark.

Introduction

This assignment covers core steps in the data science process. You will need to develop and implement appropriate steps, in Ipython (Jupyter Notebook), to complete the corresponding tasks. This assignment is intended to give you practical experience with the typical 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 regard to any announcements or changes.

This assignment is teamwork, each team with **at most 3 students**. It is up to you to form a team. Once you have formed your team, you should register your team on Canvas.

Important: you must register your team on Canvas. Anyone without a team **by 30th December 2025** will be randomly assigned to a team. If you have strong reasons for needing to complete the assignment with less than 3 members, 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 a team of 3. In addition, please submit what percentage each member contributed to the assignment and include this in your report. The contributions of your group should add up to 100%. The ones with too little contribution (e.g. less than 15%) will have their marks reduced. You may need a team leader to manage the teamwork.

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

Task 0: Choosing your project topic (1%)

This assignment covers the core steps of the data science process. You need to identify the data science problem that you want your project to solve. The data science problem must be solvable using Classification, Regression or Clustering techniques. Please choose carefully as you must list measurable project goals, tangible deliverables and work on the project with a full data pipeline and model deployment to solve that problem.

You need to select ONE of the three options below for this asignment.

1. **Problem type 1: Focusing on Data Modeling**.

For this option, you will focus on modeling the data using classification, regression or clustering approaches. You need to select **at least two approaches** among the three (classification, regression and clustering)**, one of which must be a clustering task**. For example, your choice can be c*lassification* and *clustering.*

You need to select one of the following datasets then work on it:

1.1. Incident management process enriched event log data set. More details can be found from the following UCI webpage about this dataset:
https://archive.ics.uci.edu/dataset/498/incident+management+process+enriched+event+log

1.2. Online Shoppers Purchasing Intention Dataset Data Set. 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#

2. **Problem type 2: Building a recommendation system.**

For this problem, you will work on this dataset: Anonymous Microsoft Web Data Dataset. Details can be found from the following UCI webpage
https://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

You need to implement at least two approaches for building a recommender system, such as *content-based* recommendation and *collaborative filtering-based* recommendation. For each of the approaches, you need to use some of the data modeling techniques such as classification, regression and/or clustering.

3. **Problem type 3 (optional)**

You can propose **another data set** to work on tasks for Problem type 1 OR type 2. However, the data set must be at least at the level of complexity (in terms of size and data types) with the data sets given above and must be with the same tasks. You need to send an email with a detailed description of the data set and the tasks that you will work on for the project. You need to get written permission from the teaching staff before working on your proposed project.

Task 1: Retrieving and Preparing the Data (5%)

Being a careful data scientist, you know that it is vital to set **the goal for the project**, then **thoroughly pre-process** any available data (each attribute) before starting to analyse and model it. In this step, you need to deal with potential issues in the data (such as impossible values, missing values, duplication, etc.) and explore 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: Feature Engineering (5%)

Use suitable Python functions to extract potential features and/or perform feature transformation for model input. Conduct appropriate analysis to evaluate feature importance (e.g. correlation analysis), then use suitable method(s) to select the final features for the model. The feature choices must be explained via analysis.

Note: These steps must be performed consistently for training, validation, and test sets.

Task 3: Data Modeling (10%)

Model the data by treating it as either a c*lustering, classification and/*or *regression* task, depending on your choice.

For **Problem type 1,** you must use at least **two different models** for each task (i.e. two classification models and two clustering models).

For **Problem type 2**: The two data modeling techniques (in the two recommender systems) can be any of the three data modeling approaches (classification, clustering, or regression).

For **both** types, build a supervised learning model (classification or regression) on clustered data to predict **target**, incorporating cluster labels as features. Compare performance to baseline models. Please note, the **target** variable should be chosen by you, based on your selected dataset and the research questions you defined.

When building each model, you must include the following steps:

- Select appropriate features.
- Select the appropriate model (e.g. *DecisionTree* for classification) from *sklearn*.
- 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.
- Discuss any problem you may observe or discover, such as data leakage, bias.

After you have built the models for each task from 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 and why, in your report (see Task 4).

Other Evaluation Criteria: Innovative Model (bonus 2%)

Out of the four selected models, there should be at least one innovative model (the other three models can be simple models). A simple model using only one algorithm for model training with some parameter tuning is not considered as an innovative model. For example, using a K-NN classifier from *scikit-learn* without any modification will be considered a simple model and won't have any point.

If you use a model from any research work, you must cite the reference correctly. An *example* of an innovative model is as below:

+ 1 point: a linear stacking of multiple algorithms or an ensemble model.
+ 2 points: a complex ensemble model or a complex combination of multiple algorithms. You can **propose a new model** (new algorithm) here.

Give a short explanation about the classification results obtained from the innovative model.

Task 4: Report (4%)

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 12.** Penalties will apply if the report does not satisfy the requirements. Remember to clearly cite any sources (including books, research papers, course notes, source code, 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
- Reference

Task 5: Presentation (5%)

You will be required to make a presentation in the last session of the course. The presentation should include, but not limit to:

– briefly describe your chosen problem and dataset(s).
– describe the data preparation steps.
– state the hypotheses/questions that you were investigating.
– explain what the modelling steps are, and what the results are.
– demo of the model deployment.
– show the conclusion and recommendation.

You need to prepare 10-12 slides for the in-class presentation and demonstration.

The presentation should be at a maximum of 15 minutes per group, including 3-5 mintues for demo and 5 minutes for Q&A. Each group member must present at least 2 slides in the presentation. Your presentation slides must be included in the submission before the presentation date.

5.1. Slide and presentation (2 points)

- The slides must follow RMIT University template.
- The slides and presentation must clearly present the research question(s), the used methods for solving the problem(s), the findings (results), and recommendations.
- The presentation is scheduled on Saturday, January 17th, 2025 (week 12) during our regular class time).

5.2. Demo (1 point): The code runs without error, showing the results as presented in the report.

5.3. Q&A (2 points): Students answer the questions by the lecturer and other students clearly and convincingly.

What to Submit, When, and How

Each team needs to make **one submission** on Canvas.

The assignment is due at **8:00 PM, Friday the 16th, January 2025** (in week 12). Assignments submitted after this time will be subject to standard late submission penalties.

You need to submit the following files:

- A notebook file containing your python commands, 'Assignment3.ipynb'. For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells).
- Your **report.pdf** file at most 15 (in single column format) pages (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.
- A presentation file (slides, in pptx or PDF format) for your presentation.

All the files should be zipped together, and they must be submitted as ONE single zip file, named as your team number (for example, 1.zip if your team ID is 1). The zip file must be submitted in Canvas: Assignments/Assignment 2. Please do NOT submit other unnecessary files.

Important information

Academic Dishonesty: We expect full professionalism and ethical conduct. Plagiarism is a serious offense. Sophisticated *plagiarism detection* may be used to check against other submissions in the class as well as resources available on the web. We will pursue the strongest consequences available according to the **University Academic Integrity policy**. In a nutshell, **never look at solutions done by others** (e.g., classmates, websites or **AI tools**).

Silent Policy: A silent policy will take effect **24 hours** before this assignment is due. This means that no question about this assignment will be answered, whether it is asked on the newsgroup, by email, or in person.

--- The End ---