CDS6214 Data Science Fundamentals

# Project (40%)

The aim of this project is to propose a problem that can be solved by undergoing the data science process. You are to be involved in the entire process from figuring out the problem and question you intend to ask, understand, collect data, perform cleaning/pre-processing, explore the data, build data-driven models and visualize the data meaningfully. The data science process also involves the ability to pitch your problem and to convey the message convincingly to the target audience.

**Deadline:** **11.59pm, 29th September 2024,** Sunday of Week 7.

## General Instructions

1. This is a group project to be completed in a group of a max of 4 students within the SAME tutorial section. Please refer to your tutor on instructions to register as a group for Project.

2. Based on your tutorial section, your group can only work on dataset within the domain that has been assigned.

| Tutorial Section | Domain |
|---|---|
| TT1L | Medical and Science |

3. With the assigned domain, propose a *problem* to solve. You should try to uncover insights from your data and do some literature review to gain necessary domain knowledge of the proposed problem. Under that main problem, ask four questions that you would like to answer (note: go for a variety of different types of questions).

## Deliverables

Submission and evaluated components will be in the form of:

1. *Written Report (PDF)*
   - Note: Your report should not be more than 15 pages.
   - The Cover Page, Table of Content and References are not counted as part of the 15 pages. Please use the template cover page.
   - For References, please use APA style. Excluding figures and tables, the content of your report to be written within Normal margin with Times New Roman, font size 12.
   - ***Submit only PDF file format.***

2. Supporting *materials (e.g. datasets, codes etc.) – softcopy.*

- You are REQUIRED to use **Python** for code development and data visualization. Make sure that all the codes are clearly documented and consolidated in a Python Notebook.
- Include the dataset in the appropriate format as part of your submission. The **dataset link** should be included on the Cover Page of the report.

*3.* Presentation

- Record your group's presentation, upload it to YouTube and kept as unlisted. The maximum duration is 10 minutes per group. All members in the group must do presentation and introduce themselves during the presentation. The **YouTube link** should be included on the Cover Page of the report.
- Submit your presentation slides in pptx format.

Softcopy submissions are to be done via **eBwise** before the deadline.
**Do NOT submit any zip file.**
Late submission is acceptable with penalty of 10 marks per hour. Zero marks will be awarded for submission after 5 hours.

Be aware that plagiarism is a serious offence. Cite all your references! This includes, but not limited to:

- Materials taken from websites, articles,
- Research papers, books,
- Images, videos (YouTube etc.) and other media.

**<u>Penalties</u>**

- 10 marks will be deducted for each hour late after the deadline.
- 0 mark will be awarded for this Project if the content of this Project is plagiarised from any sources
- 0 mark will be awarded for this Project if the group submit the Project 5 hours late.
- 3 marks will be deducted for the video that exceeds 10 minutes
- 3 marks will be deducted for exceeding 15 pages of the report excluding cover page and references
- 3 marks will be deducted for submitting a slide deck that exceeds 10 slides.
- 3 marks deducted for not having a Cover Page.
- 20 marks will be deducted for working with data in different domain other than being assigned per tutorial section

## Evaluation Mark Breakdown

| **Deliverable** | | **Marks (/40)** | **Totals** |
|---|---|---|---|
| Report | - Problem description<br>- Motivation<br>- Impact to communities / society / nation<br>- Domain knowledge<br>- Description of solution pipeline for each DS process component.<br>o (Questions, Data collection, Data pre-processing, EDA, Data mining/data modelling – analysis / assessments of solutions, Data visualization)<br>o Snapshots of codes and output can be included to support description.<br>- Limitation of current work and future work<br>- References | 10 | |
| Code | - Data pre-processing/cleaning<br>- Data Analysis/statistical study<br>- Feature selection/engineering<br>- Modelling<br>- Evaluation | 10 | |
| Visualization | - Data<br>- Results<br>- Models' performance | 8 | |
| Presentation | - Content organization<br>- Slides quality<br>- Audio quality<br>- Flow of presentation | 10 | |
| Others | Supplementary materials (datasets, reference etc) | 2 | |
| | **TOTAL** | | 40 |

Note: Not all group members will get the same marks. The **contribution** (in %) you filled in the cover page will be used to weight the marks among the group members. Please ensure all members agree to the submitted percentage.