TDS2101 Data Science Fundamentals

(Trimester 2, 2021/2022)

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.

General Instructions

1. This is a group project to be completed in a group of 2-3 students.
2. Each group will work on two datasets one from a and one from b:
    a. Dataset 1:

    Select with your group one of the datasets found in the UNICEF website (https://data.unicef.org/resources/resource-type/datasets/). This dataset should be tacled by all team members.

    b. Dataset 2:

    Select one of the datasets below:

    - https://www.kaggle.com/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring?select=fifa21+raw+data+v2.csv
    - https://www.kaggle.com/rishidamarla/colleges-and-universities-in-the-us
    - https://www.kaggle.com/ayushggarg/covid19-vaccine-adverse-reactions

    Please submit your group members' names and indicate your choice of dataset (in order of preference) when registering your group. This should be done by the end of Week 8 (latest) via Google Form to be provided.

3. You are allowed to use relevant data from other sources to supplement the given data.
4. With the assigned dataset, propose a ***problem*** to solve and to uncover insights from your data. Under that main problem, design a number of questions that you would like to answer (**a question from each category of questions studied in the course**).

5. Submission and evaluated components will be in the form of:
    a. Written Report – softcopy
        - Part A: Proposal write-up: Motivation, Problem description, Questions
        - Part B: Overall report
    b. Supporting materials (e.g. additional datasets, codes etc.) – softcopy. Note: You are REQUIRED to use **Python** for code development and data visualization.
    c. Presentation:
        - Part A: **Pitch** Session – 5mins max per group. Presentation will be held in Week 11 (TBA).
        - Part B: **Deliver** Session – 15mins max per group. (10mins presentation + 5mins Q&A).

Softcopy submissions are to be done via **MMLS**, unless specified otherwise.

Deadlines

Part A: Your first checkpoint will be in **Week 11.** At this point, a write-up (maximum of 2 pages) should be prepared to highlight the following points:

- The problem that you intend to solve using data science, and the motivations behind it
- The question(s) that you formulate, with the aim to answer them via the data science process.
- Any supplementary dataset(s), and its justification for use.
- The potential benefit/impact of the problem to be solved on a specific target sector or customer/business, or society/nation.

You will conduct a 5-minute pitch (the points above provide some guide, but it is free-form), to convince the audience that your problem is suitable and impactful to be solved using the Data Science process.

Part B: Your final submission is expected in **Week 15 (last working day)**. The final report should be handed in, together with Python implementation demonstrating how the outcome or results were achieved, and the slide deck.

All groups will be required to do a final presentation on the entire problem, steps taken in the data science pipeline, and the findings and analysis after concluding the task. This presentation is to be held on a date after the submission (to be announced).

In your presentation, ensure that your main points are conveyed in a convincing manner. Be prepared to answer questions relating to what you have done, your choice of methods, and the outcomes.

> 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.

Timeline

Week	4	5	6	7	8	9	10	11	12	13	14	15*
				Project released		Team submission		Proposal			Final Submission	Final Presentation
				Part A: Proposal			Part B: DS Pipeline in action					

Guide

Part A. Understanding how a problem can be solved with the Data Science process (12%)

1. Formulate your Problem based on the chosen starting dataset. It can be an overarching theme in a very broad way, or it can also be something interesting and narrow that piques the interest of the audience. The motivation behind why you formulate it as such, is crucial.

2. Formulate a few Questions to the proposed problem. Aim for a variety of types (Descriptive, Exploratory, Predictive, etc.). While basic Questions could be easy to answer, more complex Questions are more challenging but also more insightful and can give provide you with an edge over other projects working on the same data.

3. Try to look at your Problem from the perspective of its impact towards a particular target sector, or customer/business, or even to the level of society and nation. This will give you a foundation as to why the problem you are tackling is meaningful or beneficial.

4. The art of "pitching" an idea or problem requires the ability to communicate *clearly*, *concisely* and *convincingly* to the target audience so that they can see that your problem is worth pursuing or investigating. Imagine someone were to be funding you or giving you a job based on what you are pitching, how would you do it?

Part B. Walking through the Data Science pipeline (28%)

Picking up from what was proposed in Part A, you need to then put the ideas and plans into action. Here are some points to guide you along:

1. Understand the context of the Questions that you are going to answer. You can refine your Questions from Part A if it is needed with new developments to your ideas, or if new data enters into the picture, etc.
    - At this point of time, you should know how different types of Questions can be investigated or analysed for its outcome through different ways. E.g. A exploratory question is likely to be answered by examining trends or relationships or patterns, but a predictive questions would likely need a model to be built and validated. Basically, understand what is needed.
2. Describe the contents of the dataset(s) from various perspectives.
    - Also provide some preliminary insights into the condition of the data and what it offers.
    - If you have other supplementary datasets, describe them as well, also disclosing clearly where the source is and how it was collected.
3. Examine the quality of the data and perform necessary data cleaning steps.
    - What are the data cleaning activities/tasks that you have performed? Describe in detail how they were done and justify their need.
    - Deciding on the right collection and ordering of pre-processing steps is essential. Note that how this is done also depends a lot on your project needs (Problem, Questions)
    - If you are using more than one datasets, separate cleaning procedures may be required. Merging of data also constitutes a pre-processing step to prepare the merged data for later steps.
4. Explore the data to understand its descriptive statistics and uncover possible anomalies or relationships (influencing factors).
    - Understanding the dataset(s) in depth requires exploration of the data, both quantitatively and qualitatively.
    - At this juncture, potential relationships or influencing factors can be uncovered.
    - If necessary, it is fine to show some graphical plots or visualizations at this stage, if it helps to illustrate the current state of the data and if there exists any interesting relationships or trends within the data.
5. Employ data mining (e.g. Clustering, Association Rule Mining) OR predictive modelling techniques (e.g. Decision Trees, Linear Regression, Logistic Regression etc.) according to the suitability to your Problem/Question to gain insights into the potential of the data in churning out patterns or making simple predictions.
    - Are you able to mine some interesting patterns, or build a model to predict future behaviour based on the data you have obtained?

- Note: It is not a necessity to use machine learning/AI techniques here since it falls outside the scope of this course. However, there is also no restriction if wish to.

6. Use compelling visualizations to support a consistent narrative.
   - Show with visuals how your Question(s) can be answered.
   - With visuals, you should also take the opportunity to discuss and analyse further to generate actionable insights that will establish important observations and conclusions that are impactful. These visuals also would be very useful to construct your final slide deck.
   - Note: It is acceptable to have visualizations done in earlier stages of the pipeline such as EDA, data mining or predictive modelling, when deemed appropriate.
7. Mention and discuss some of the challenges or restrictions faced in this project.
   - What was challenging when dealing with the data or the Problem that you had proposed?
   - Provide some thoughts and future directions of the way forward to improve on what you have done.

Evaluation Mark Breakdown

Part	Deliverable	Marks (/40)	Totals
A 30% of total	Report - Content (3) o Motivation, Problem, Questions - Organization, Clarity and Language (1)	4	
	Evaluation on the Proposed Problem to be solved - Impact (1) - Feasibility of proposal (2)	3	
	Presentation - Clarity of main idea/problem to be solved (2) - Pitch (1) - Q&A (2)	5	
			12
B 70% of total	Report - Content – all required components (6) - Organization and Language (1) - References (1)	8	
	Requirement of DS process components - (Questions, Data collection, Data pre-processing, EDA, Data mining/data modelling, Data visualization)	3	
	Pipeline solutions - Description and implementation of solution for each DS process component (5) - Analysis and assessment of chosen solutions (2) - Challenges encountered and future proposals (1)	8	
	Presentation - Flow and verbal clarity of presentation (3) - Visual content & impact (3) - Q&A (2)	8	
	Supplemental materials (datasets, codes, etc.)	1	
			28
	TOTAL		0

Rubrics for Part A and B will be released reasonably earlier than their respective submission deadlines.