Project Phase #3

Due: ~~11/28/22~~ 12/2/22 @ 11:59pm

Content Covered
Data Product and Project Wrap-Up

Project Overview

The course project forms the hands-on practical learning component of the course, and will have students putting into practice each step of the data science pipeline (depicted in Figure 1, adapted from [1]). The project will be broken into 3 phases, with Phase 3 covering step 8 of the data science pipeline shown below. The project is expected to be motivated by issue(s) in an application domain of your interest, and addressing these issues using data gathered from the domain.

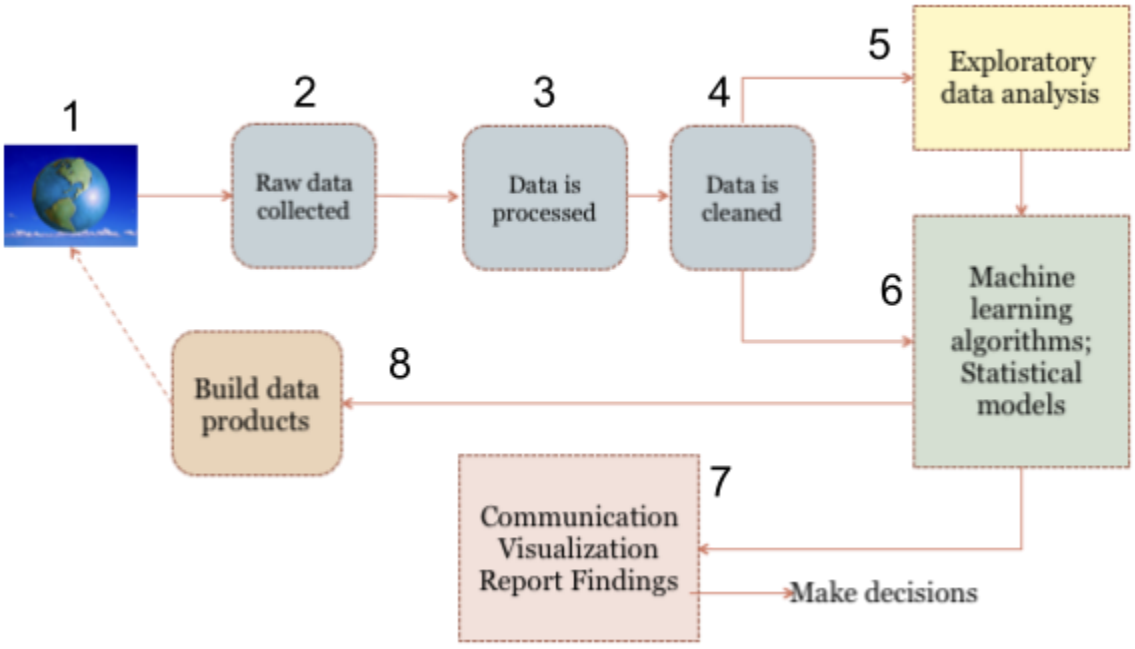

Figure 1: The Data Science Pipeline

Learning Outcomes for Phase 3:

1. Build a data product around the exploration. Could be as simple as a web tool to input different data sets to the model built and explore the outcomes with different parameters.

Description:

Now that you have formulated a problem statement and processed, cleaned and analyzed your data (Phase 1), and built and tuned models to gain intelligence from this data (Phase 2), you must wrap it all up with a final data product. In Phase 3, you will be building a data product from your Phase 2 models which would allow a user to interact with your models to gain insight into the data/problem statement you set out to solve. This could be as simple as allowing a user to input their own dataset for automatic analysis, or something more complicated tailored to your particular problem domain. The deliverables for Phase 3 will include your entire project code and documentation, as well as a short demo video showing your completed product.

General Project Requirements

1. **Work Environment:** Recommended language for the project is Python. You can use any Python environment of your choice: Jupyter, IPython, etc.
2. **Programming:** Prepare yourself to program by learning from the course textbooks and online resources.
3. **Academic Integrity: You will get an automatic F for the course if you violate the academic integrity policy.**
4. **Project Phases:** This project will span three separate phases, each building on the last. Each phase has its own due date, and must be completed before you can move onto the next phase. Late submission will not be accepted.
    a. During Phase 1 you will be forming a problem statement, getting your data, and doing initial EDA. During Phase 1 you may change your problem, or the data you choose to use. Once Phase 1 is complete you will no longer be allowed to change, which makes it critical that you carefully complete Phase 1.
5. **Teams:** For the duration of the project you may work in groups of one or two only. Project discussion should only occur between you and your teammate, or you and course staff. Each team member must complete the each part of the project and submit their own notebooks, and writeups. A google form will be made available via Piazza to register your team and be assigned to one of the course TAs. **You must complete this step before asking for project guidance.**
6. **487 vs 587:** In certain instances 587 students will be required to complete additional work, and in general their projects will be held to higher standards. Instances of additional work will be clearly identified in the deliverables section.

Deliverables **[50 marks total]**

This phase involves building a data product that automates and parameterizes the process you established in Phases 1 and 2. This could be a web-application, mobile app, or enterprise application to allow users /clients to interact with it. Deliverables will include the code for your entire project, including the data product, complete with documentation, as well as a demo video showing off the product.

1. **Code [25 marks]:** In your submission you must include complete code for Phases 1-3, fully documented, and with clear and succinct instructions on how to run your code with different datasets. The report should contain any relevant notes on how your models are tuned/evaluated (ie: p value is 0.0006 etc), and any recommendations related to your problem statement based on your analysis.
    a. **[5 marks]** for fully documented code and working instructions to demo/use your finished product
    b. **[10 marks]** for relevant notes on how you specifically used the models from phase 2 (which models did your product end up using, tuning of any relevant parameters, etc)
    c. **[10 marks]** for recommendations related to your problem statement based on your analysis (what can users learn from your product, how does it help them solve problems related to your problem statement, other ideas for how to extend your project, or other avenues that could be explored related to the problem)

2. **Demo [25 marks]:** In addition to your code, submit a short video pitch/presentation (no more than 5 minutes in length), giving a brief demo of your product, how it is used, and what information people can learn from it.
    a. **[5 points]** for showing a working user-interface (not just code you would expect a user to run)
    b. **[5 points]** for showing how a user could input their own data (uploading their own datasets, filling out fields in your GUI, etc)
    c. **[15 points]** for showing the feedback your product gives, explaining what it means, relevant manipulation/filtering of visualizations, and how a user could use it to help them solve a problem/answer a question.

Submission Format

By the due date, each team must submit a single zip file containing the entire project, with a top level README, and subdirectories for each Phase of the project.

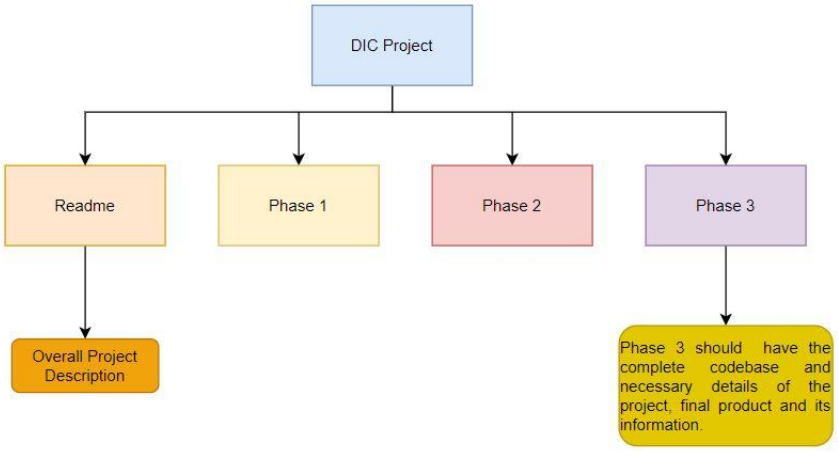

Figure 2: Final Project Directory structure