Project phase 3

Due: 12/5/2024 23:59 p.m.

Content Covered

Building a data product

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). The project will be broken into 3 phases, with Phase 3 covering the steps 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.

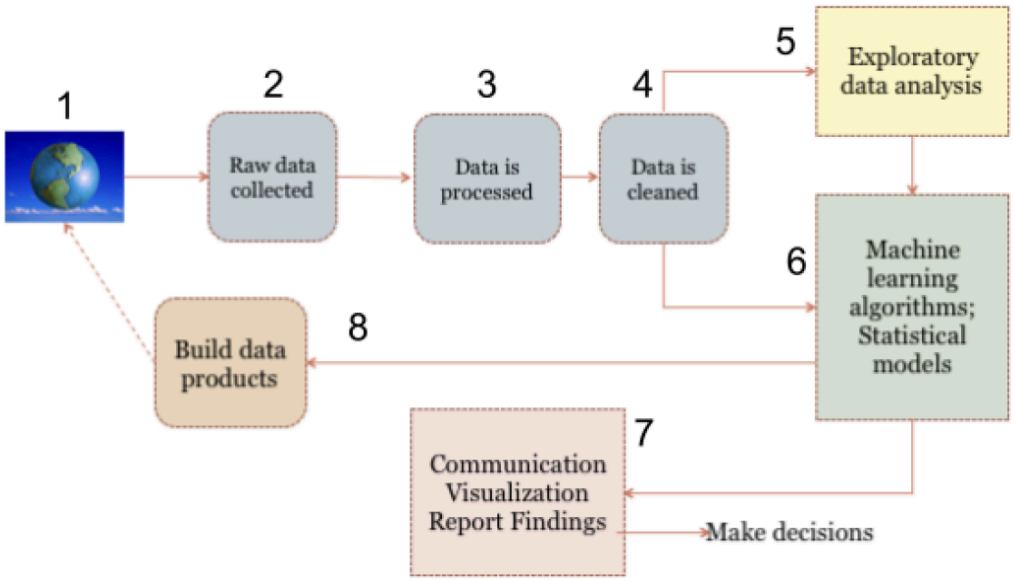

Figure 1: The Data Science Pipeline

Figure 1: The Data Science Pipeline

Learning Outcomes for Phase 1:

1. Identify problems prevalent in public application domains.

2. Research and identify data sets (preferably structured data) to address the problems and collect the relevant data sets.
3. Clean and provision the data for downstream explorations and analytics.
4. Understand the basic characteristics of the data by performing John Tukey's exploratory data analysis (EDA).

Learning Outcomes for Phase 2:

5. Identify suitable ML, MR, and/or statistical modeling algorithms. Model and apply algorithms to get insights into the behavior of the data. It could be classification, regression, clustering, etc.
6. Understand and explain the differences in each of the algorithms used.
7. Visualize the analytics using appropriate charts and graphs. You can use Seaborn or any other plotting tool.

Learning Outcomes for Phase 3:

8. Understand how to persist data by using databases.
9. Build a data product utilizing modern web UIs. The interactions include data input/output/storage/visualization/exploration with different parameters.

Description

An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov, pew research) have their data available to the public. Social network applications such as Twitter and Facebook collect enormous amounts of data contributed by their numerous and prolific users. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make a subset of their data available for use by registered users for free. Some of them, as downloadable data files (.csv, .xlsx), others as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as web page content. In this case, typically a web scraper is used to crawl the web (pages) and scrape the data from these web pages to extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products. You'll research these organizations that deal with data and data solutions, and decide on a domain and set of data, and the goals of your analysis of this data. In particular, what are the questions you want to answer by analyzing this data. For this project you'll work with structured data sets. If your dataset is unstructured, then you need to program to make it structured.

General Project Requirements

1. Work Environment: Required language for the project is Python. You can use any IDE of your choice: Jupyter notebook, VS code etc. For webapps any programming language can be used.
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. See the course syllabus for more detail.
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 receive a penalty.
    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 (but make sure your dataset is not used by other teams). Once Phase 1 is complete you will no longer be allowed to change, which makes it critical that you carefully complete Phase 1.
    b. During Phase 2 you will be applying ML, MR, and/or statistical modeling algorithms to your datasets that you have cleaned and analyzed in Phase 1, with the goal of gaining deeper insight into the data and answering questions related to your problem statement.
    c. During Phase 3 you will be building a data product from the work you have in Phase 1 and Phase 2, targeted at end users in your problem domain.
5. Teams: For the duration of the project you may work in groups of 3 to 4 only. Please make sure that you have registered your team via Google form, which must be completed before asking for project guidance. Project discussion should only occur between you and your teammate, or you and course staff. Each team member must contribute each part of the project. There will be one submission per team..

Submission Requirements

1. Deadlines: Your submission is due by **11:59PM on Thursday, 12/5/2024**. For each day your submission is late, there will be a 20% penalty. You must submit Phase 2 to begin work on Phase 3. Please start the project as soon as possible.
2. Submission: Project deliverables should be submitted via UBLearns. There should be one final submission per group. You can submit multiple times before the deadline but we will grade the final submission. For the final submission you are required to submit a zip file containing all the required deliverables. The zip file must be named:

[member 1 ub number]_[member 2 ub number]_[member 3 ub number]_[member 4 ub number]_phase_3.zip.

It should contain the following items:

1. One single complete PDF report under the root folder.
2. A folder named “app” that contains all app code.
3. A folder named “exp” that contains the final version of the previous python notebook code.
4. A short video < 5 mins under the root folder
5. A readme.md file under the root folder

Below are the detailed tasks of each item.

Tasks [50 marks total]

1. **Report [10 marks team]**: The report should be well-written in academic paper format which wraps up the whole project.
2. **App [30 marks team]**: The data product should demonstrate the use of database and user interactions with the data without having to open the jupyter notebook. We expect to see
    a. **[5 marks team]** Deployment of a suitable database for data persistence. All structured and unstructured dataset throughout the project must be persisted. **(It can be cloud-hosted)**
    b. **[5 marks team]** An interface that allows users to lookup, add, modify and remove the data entries in the database.
    c. **[20 marks team]** Pick at least **four** of the most significant questions, implement the model or algorithms, compute the result or prediction based on the fetched data from the database and the parameter from the user input. Display the result to the user. **You are allowed to improve/change your questions in this phase too.**

    The app can be built using popular web frameworks.

3. **Short video [5 marks team]**: The screen capture video demonstrates the building process and the major interactions of your app.
4. **readme.md file [5 marks team]**: It contains
    a. The questions listed for each team member.
    b. The information of where the experiment code associated with each question is located (the file or the line number of the file if only one ipynb)
    c. The analysis of the question is located. (It could be the same location of the code)
    d. The folder structure information e.g. app/ contains all app code exp/ contains all ipynb code exp/ contains experiment results etc.....
    e. The instructions to build the app from source code.

Additional Information and References

Some popular frameworks:

https://streamlit.io/

https://plotly.com/

https://redash.io/

https://solara.dev/