Project Phase #2

Due: 4/10/23 @ 11:59PM

Content Covered
Machine Learning and Statistical Analysis

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 2 covering the steps 6 and 7 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.

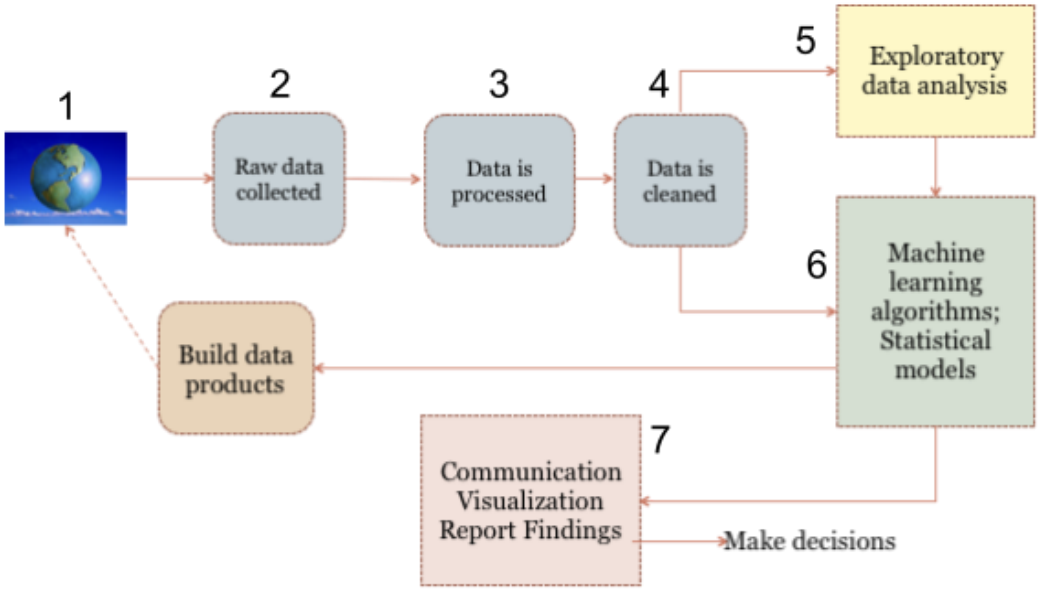

Figure 1: The Data Science Pipeline

Learning Outcomes for Phase 2:

1. 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.
2. Understand and explain the differences in each of the algorithms used.
3. Visualize the analytics using appropriate charts and graphs. You can use Seaborn or any other plotting tool.

Description:

Now that you have cleaned, processed, and exploratory analysis on your data you will be applying more significant algorithms from Machine Learning (ML), statistical modeling, and/or MapReduce (MR) to start gaining intelligence from your data. We have discussed a number of different algorithms/approaches in class, and there are more references in the course textbooks as well as in chapters 4 and 5 of [2]. This phase also includes visualization, as the output of the algorithms are often displayed as charts and graphs. The algorithms you use and the visualizations you produce should be targeted towards answering questions related to your initial Phase 1 problem statement.

General Project Requirements

1. **Work Environment:** Required 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. 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.
    a. 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.
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 contribute each part of the project. There will be **one submission per team.**
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.

Submission Requirements

1. **Deadlines:** Your submission is due by 11:59 PM on Monday, 4/10/2023. For each day your submission is late, there will be a 25% penalty. You must submit Phase 1 to begin work on Phase 2. 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: **`member1_member2_phase_2.zip`**. It should contain a PDF for your project report named `report.pdf` and a `src/` directory with your commented code files.

Deliverables **[50 marks total]**

1. **Algorithms/Visualizations [25 marks]:** Apply **5 different significant and relevant algorithms** (ML, MR, and/or statistical models) to your data and create visualizations for the results. **For 487 students**: at least 1 of the 5 algorithms must be one that was not discussed in class. **For 587 students**: at least 2 must be from outside of class. Algorithms discussed in class are: Linear Regression, k-Means, k-NN, Naive Bayes, and Logistic Regression. The outside algorithms can come from the class textbooks, or other sources. **Cite the appropriate sources for each outside algorithm you choose to apply.**
2. **Explanation and Analysis [25 marks]:** For each of the 5 above algorithms, provide justification for why you chose the particular algorithm for your particular problem, work you had to do to tune/train the model, and discuss the effectiveness of the algorithm when applied to your data to answer questions related to your problem statement. This should include discussion of any relevant metrics for demonstrating model effectiveness, as well as any intelligence you were able to gain from application of the algorithm to your data.

Additional Information and References

Some references and tutorials for a variety of algorithms and visualization techniques can be found here:

1. https://seaborn.pydata.org/tutorial.html
2. https://scikit-learn.org/stable/
3. https://plotly.com/python/

References

[1] C. O'Neill and R. Schutt. Doing Data Science., O'Reilly. 2013.
[2] J. VanderPlas, Python Data Science Handbook., O'Reilly. 2016.