Project

Due: 04/26/2023 at 11:59pm EST

Collaboration Policy: This project will be completed in groups containing around fixe students. IF (and I doubt this will happen) a group member(s) is not contributing, I want to know about it as **soon as possible**. I cannot do anything about it if I find out **after** the work has been done. As I said earlier, I doubt that this will happen in this course, but just in case it does, please bring it to my attention ASAP.

Overview: This project is intended to give you a chance to explore data mining on real data for a task of your choosing. After you have formed your group (or used the task as a mechanism for finding similarly-interested students to form a group with), please come and talk to me so we can find you some data. You will then use this data and try to accomplish your task using the data mining techniques we have learned in class. You are of course, allowed to learn on your own other techniques which we do not talk about it class if you are interested, but I do not require you to do so. We will use the last two remaining lectures to host project presentations. Later in this document is the rubric I will be using, so please keep this is mind when doing your project and when making your presentation. Good luck!

Finding Data: Once your group has decided on a task, please send me an email. In this email, I need to know the following information:

1. The members of your group.
2. The task you are interested in.
3. If you have already found data for your task.

If you have **not** found data for your task (I suggest checking `kaggle.com`), I will help you find data. Please do **NOT** pay for data or for computing services (such as AWS, etc.). You should **not** spend your own money for this project. Come talk to me if your data is behind a paywall.

In order to keep projects on schedule, please send me this email **NO LATER** than Sunday 04/16/2023. This is critical, as I **CANNOT GIVE EXTENSIONS FOR THE PROJECT**.

What to Submit: I want the code that you wrote as well as the presentation (i.e. powerpoint slides, pdf, etc.) that you create. In your presentation/code, please include either a link to the data that you used (that I can download the data from), or clear instructions on how I can acquire the data on my own.
Submission instructions: Please submit your code as well as your presentation slides to me via email (one email per group).

1 Rubric/Outline

This is the rubric I will be using to grade your project. The questions I will ask here I expect you to consider in your project, and I also suggest that you view this rubric as an outline for your presentation.

What is your task?

What task are you trying to do? What problems/applications does this have? Why should we care?

What is your data?

What information does your data have? What format is it in? Where did you get it from? What are the statistics of your data?

How are you going to represent the data to the machine?

Why did you choose this representation? What information are you losing/preserving by using your representation?

How are you preprocessing your data?

What steps are you taking to convert your data into your chosen representation? Why did you choose these steps?

What model are you using to solve your task?

What does your model take as input / output? Why did you choose this model?

How are you evaluating your model?

What performance metrics are you using? Any criteria that you're looking for specifically?

How did you model perform?

Did you model perform well? What limitations does your model have? When does it do a good job and when does it do a bad job? What did you do to try and make your model perform better?

What are your conclusions?

Do you think any of your representation choices affected your model? What did you learn from this project? What would be your next steps if you were to continue this project?