Final Project Submission deadline

May 8th @ 11:59 PM (So it must be submitted *BEFORE* the clock hits 11:59)

Collaboration

You can make a group of up to 3 people, so having less than 3 people is fine, but you will still be expected to have a project of the same quality as if 3 people worked on it, so keep that in mind. Each member is expected to make technical contributions (so one person can't just make the video or presentation). Your work effort will be reported anonymously in the submission by your other teammates.

Project

1. Form group
2. Choose a data set from the Machine Learning Repository
3. Choose a project type that matches with your data set (You *must* choose only one):
    a. Classification
    b. Clustering
    c. Regression
4. Using what you've learned so far this semester, you will perform whatever project type you chose using *at least* 3 methods to compare (ex. You choose classification, then you should use 3 classification methods like Naïve Bayes, Random Forests, Decision Trees, etc.). Your analysis should be robust, i.e., you should take care to explore your data, evaluate your features, collinearity, scale appropriately, split, remove missing values, and your results should be explained well with graphs/charts. This is the final project, so it should showcase your ability to do a complete data science project from start to finish. You should also be able to communicate your findings with a deep understanding of the data set and models you chose.
5. Create a PowerPoint or other presentation with the following format:
    a. Introduction / Problem Statement
    b. Methods / Implementation
    c. Results
    d. Discussion / Explanation of results (not just, "oh we got 68% accuracy"), as well as pros and cons of your model, your data, etc.
6. Create a video of your group presenting. You can do this a few different ways, but I would suggest everyone hopping on a Teams meeting and presenting it, and you can record this. Other screen recorders might work, too. The presentation should be at least 8-10 minutes long, but anything

longer than that is fine. Heavy emphasis will be placed on your explanation of your project, as well as your interpretation. Simply telling me what you coded will cost heavily in grading. You must show that you understand the concepts behind the methods you're employing.

Extra Points on Final

As an optional assignment, any individual can do, *in addition to their group project*, another project with another data set from the ML repository mentioned above to receive +4 points on the Final Exam. This project must be done individually and turned into the same HW5 assignment drop box on Blackboard. It has the same requirements in terms of analysis that the group project should have. Should you choose to do one or both of the extra assignments, you will submit a zip file. Each extra project you do should also have its own .ipynb file and PDF report (at least 3 pages, should include graphs and charts) with the same structure as the PowerPoint (Introduction, methods, …). Each completed extra projected must be done thoroughly to receive full credit and extra points on the final.

TIPS:

- Try not to choose a data set with too many features, as it might make your models unstable and complex (which will make your computer's processor hate you). If you do choose to have many features, use some dimensionality reduction / feature selection method, but keep in mind that interpretability of your models is key.
- Having data sets with too few observations / instances (< 150) will make your model prone to overfitting, because the training set won't have much data to go on.
- Work with people who are actually committed to working hard and doing their part. Sucky partners = a sucky project.
- Reach out to Chad Weatherly on Teams or by email (cdweathe@central.uh.edu) if you have any questions, but I respond quicker on Teams.

Submission

- You will upload your video to the HW5 assignment in Teams since Blackboard doesn't allow uploads > 250 MB. Only one student needs to submit, with the file name LASTNAME1_LASTNAME2_LASTNAME3_HW5.mp4
- *EACH* student should submit their group's single .ipynb file (you can have extra .py files if you want to make your code cleaner and easier to read. If you don't know how to do that, don't worry about it). In the comments of your submission on Blackboard, include the contributions of each

team member. This should include what each team member specifically did, along with a percentage of workload done by each team member. Your submission here is hidden from the other students, so please be honest.

- Should you choose to do any extra projects, you will not only submit your comments, you will submit a zip file with the following files contained:
    - Group's .ipynb file:
        - LASTNAME1_LASTNAME2_LASTNAME3_HW5_.ipynb
    - Individual project files (you will have 2 of each file should you choose to do both extra projects)
        - LASTNAME_PROJECT TYPE_HW5.ipynb
        - LASTNAME_PROJECT TYPE_HW5.pdf
        - Where PROJECT TYPE is classification, regression, or clustering
- If you don't choose to submit any extra work, that's fine, and you will only submit your group's .ipynb file.
- Finally, the grading will be done as follows:

Step	**Points**
Data Exploration and Preprocessing	20
Data Analysis	20
Results / Discussion / Explanation	50
Presentation	10