Final Project

Perform an Exploratory Data Analysis (EDA)/predictive modeling on the Heart Disease UCI Kaggle Dataset or a Dataset of your choice.

In this project, you will create a github repo and notebooks to perform EDA/ML on a dataset of your choice. If using the proposed Kaggle Dataset, please predict heart disease outcomes with at least one model you have implemented. If using your own dataset, please implement at least one predictive ML algorithm that's appropriate for your dataset.

Final notebook(s)/repo should have:

- annotated code with markdown
- several graphs
- executable ML model with score
- executive summary

Personal datasets must be approved by end of class, Wednesday Jan. 27th. Please consider having a backup dataset just in case.

Dataset considerations:

- Is there enough data?

- Can ML be performed on the data?

- suggest > 1,000 samples. More is likely better

- if using images, CPU time could be an issue

Week 22

Project Part 1: EDA due Wednesday, Feb. 3

- Create a github repo for your mini project. Post the link to your GitHub repo in Canvas for Project Part I: EDA
- Your notebook should address each of the following:
  - Data issues: missing values, duplicate values, outliers
  - Data cleaning solutions: imputation/estimation, dropping entries -- justify your choices!
  - Describe the relationship of features to your target (should include at least a few plots).
  - Feature engineering (transformation, normalization, creating new combinations of features, etc.) if you think this is necessary. Describe your rationale.

Week 23

Project Part 2: Modeling due Wednesday, Feb. 10

- Post the same link to your GitHub repo in Canvas for Project Part II: Modeling. You can either create a second notebook for modeling, or create a second section in your EDA notebook.
- Your modeling notebook/section should include each of the following:
    - Feature engineering, if not captured in the EDA notebook.
    - Splitting data into train/test sets
    - Build at least one model
    - Predict test set using model(s)
    - A quantitative metric of model(s) performance

PRESENTATIONS: please be prepared to give a 5-min maximum presentation of your project in class on 2/10. The format of the presentation is up to you – you can prepare a few slides if that's helpful, or simply take us through your notebooks. Keep it short and sweet and have fun!