CS 250 – PROGRAMMING FOR DATA APPLICATIONS
PROJECT 1 SUBMISSION GUIDELINES

Submit a Jupyter Lab Notebook and a presentation video per team of two. Use a lot of Markup cells to explain everything.

Data and the Purpose of the Project

Specify the name of the dataset. If the data is publicly available, give a link to it. If not, you can submit it as a CSV file if it is less than 10MB in size. If it isn't publicly available and it is too large to submit, mention that. This is one of the scenarios where you may be called to my office to hand in the data or demo the project in person. Next, describe what the data contains and specify what problem you are planning to solve using that data. For instance, "The dataset contains dimensions and weights of different species of fish sold at a fish market, and we are planning to predict the species of fish based on the weight and dimensions."

Analysis of the Data

Here you analyze the data and make some charts. The charts will be outputs from a Jupyter notebook. There is no hard and fast rule about what analysis you need to do, but it should help me understand your data. So, the number of rows, columns, classes and number of items in each class (if it's a classification problem), max, min, mean etc if it makes sense. Maybe a clustering visualization, or scatter/bubble chart. Understand your data thoroughly at this stage and be creative. If you are working on image data, you can include some sample images.

Solving the Problem

Here you try to solve the problem you proposed earlier by some kind of Supervised/Unsupervised Machine Learning technique covered in class. Split the data into independent and dependent variables, and training/test sets for supervised techniques. Fit your model on the training set, and test its performance on the test set. Evaluate this using numeric error or success values and appropriate visualizations. If this performance isn't great, don't panic. Write down why you think it isn't good, and what in your opinion may be causing it.

Work as a group and submit one *.ipynb* file per group on Moodle by the beginning of class on Monday, March 17. This should contain, all code in the form of a Jupyter Notebook, and the dataset if needed (see above).

Presentation

Each team has to present the project in class during the week of Monday, October 20, and the following week if necessary. The presentation must contain a walkthrough of the Jupyter notebook, with ALL team members participating in explanation of the data and the results. Under no circumstances should this exceed 10 minutes.