CS 250 – PROGRAMMING FOR DATA APPLICATIONS

PROJECT 2 SUBMISSION GUIDELINES

Submit ONE copy of the following PER PERSON

A one-or-two-paragraph text document which says, in your own words, what yours and the other team members' roles were in your project. In other words, outline the work that each one of you did. Write this in your own words, do not copy-paste what your teammates write.

The following things should be done ONCE PER GROUP

1) Submit ONE copy of a Jupyter Notebook containing the things listed below. The length of the document should be similar to what you did for Project 1. Use a lot of Markup cells to explain everything in your Jupyter Notebook.
2) A presentation of your project in the class. All members of your group should participate in this presentation.

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. 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.

Keep in mind that

1) ***Your Project 2 must use a different machine learning technique from Project 1.***
2) ***Your Project 2 must use a different dataset from Project 1.***