DAV 5400 Spring 2019 Project 4 (Due Week 13: 100 Points)

Predictive Analysis using scikit-learn

As data analysts, we're often tasked with taking data in one form and transforming it for easier downstream analysis. In this Project, you'll use what you've learned in the course to prepare data for predictive analysis and then construct a predictive model using tools available within the **`scikit-learn`** library. ****You may work in small groups of no more than three (3) people for this project.*** **

The data set you will be using is the UCI Mushroom Data Set (which we first used in **Week 5** as part of an in class "Hands On" exercise): https://archive.ics.uci.edu/ml/datasets/mushroom. The fact that this is such a well-known dataset in the data science community has made it a good dataset to use for comparative benchmarking. For example, if someone was working to build a better decision tree algorithm (or other predictive classifier) to analyze categorical data, this dataset could be useful. In this Project, we'll use **`scikit-learn`** to answer the question:

"***Which other attribute (i.e., aside from the poisonous/edible indicator) or attributes are the best predictors of whether a particular mushroom is poisonous or edible?"***

The work you will need to do for this Project can be separated into two distinct phases: all of the work required for **Phase I** can be completed using the Python and Pandas skills you developed through Week 10 of this class (i.e., without any prior **`scikit-learn`** knowledge), while Phase II will require the use of **`scikit-learn`** to assess the predictive qualities of the non poisonous/edible indicators contained within your DataFrame.

Phase I: **Data Acquisition, Data Preparation & Exploratory Data Analysis**

- First study the dataset and the associated description of the data (i.e. "data dictionary").
- Create a pandas DataFrame **with a subset of the columns in the dataset**. You should include the column that indicates **edible or poisonous**, the column that includes **odor**, and **at least two other columns** of your choosing.
- Add meaningful names for each column in the DataFrame you created to store your subset.
- Convert the "e"/"p" indicators in the first column to digits: for example, the "e" might become 0 and "p" might become 1. For each of the other columns in your DataFrame create a set of dummy variables. This is necessary because your downstream processing in Project 4 using **scikit-learn** requires that values be stored as numerics. See the pandas **get_dummies()** method for one possible approach to doing this.
- Perform exploratory data analysis: show the distribution of data for each of the columns you selected, and show plots for edible/poisonous vs. odor as well as the other columns that you selected. It is up to you to decide which types of plots to use for these tasks. Include text describing your EDA findings.
- Include some text describing your preliminary conclusions about whether any of the other columns you've included in your subset *(i.e., aside from the poisonous/edible indicator)* could be helpful in predicting if a specific mushroom is edible or poisonous.

Phase II: **Build Predictive Models**

- Start with the mushroom data (including the dummy variables) in the pandas DataFrame that you constructed in Phase I.
- Use **scikit-learn** to determine which of the predictor columns that you selected (odor and the other columns of your choice) most accurately predicts whether or not a mushroom is poisonous. How you go about doing this with **scikit-learn** is up to you as a practitioner of data analytics.
- Clearly state your conclusions along with any recommendations for further analysis.

HINT : If you understand the process used in the DataSchool videos on Machine Learning with scikit-learn to predict iris species from four predictor variables, you should be able to apply what you've learned to complete this Project.

Save all of your work for this project within **a single Jupyter Notebook** and upload it to your online DAV5400 GitHub directory. Be sure to save your Notebook using the following nomenclature : **first initial_last name_Project2**" (e.g., J_Smith_Project2). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***

As a reminder, Project 4 is due no later than 11.59pm on Sunday April 21.