DSA 610, Homework #7, Spring 2025

Part I:
Instructions: Answer each discussion question in your own words. You may use posted resources or other online resources to answer the questions (**cite your sources**). Thoroughly explain your responses with a minimum of one paragraph (3-5 sentences) in length. Be thoughtful. We will discuss the answers in class.

1. Consider a start-up company working out of someone's basement. (What they do exactly is up to you.) Consider their data requirements at the start and as the company grows. How would their data needs change as they grow? Be specific about when or if they might be able to use a spreadsheet, database, data warehouse or data lake, or some combination of these.

2. What are ACID requirements? Why might non-relational databases need to relax some of these standards?

3. Why is domain knowledge important in data analysis? Describe a situation in which insufficient domain knowledge can/has led to poor conclusions.

4. Choose one of the data analysis tools mentioned in the lecture notes and compare it to Python. (If it's one you've never used before, use the product documentation as your guide.)

5. Among the classification models we considered are binary classifiers (two options), and multi-class classifiers (2 or more options). Investigate multi-label classifiers (each observation can receive more than one label). Describe at least one algorithm that can be used for multi-class classification and describe a potential use case.

Part II:
Instructions: Use the attached dataset (**beer_preference_data_hw1.xlsx** and **wine_hw7.xlsx**) to complete the following tasks in Python. Report your answers to the questions on this homework sheet. Include your Jupyter notebook file along with your homework submission, saved as a PDF. Make sure that any graphs you create are quality graphs with a legend (as needed), axis titles, a descriptive title, appropriate ranges (bar graphs start at 0, etc.). They should be able to stand on their own. You may need to relabel some elements (such as replacing 0s and 1s with string labels).

Use Python to answer the following:

1. Clean the data by removing the Individual column. The Beer Preference and Beer_Pref_Class columns represent the same data, so remove the Beer Preference column (keeping track of which label is 1 and which is 0).

2. Construct and evaluate the following models on the beer preference data:
   a. A Logistic Regression model
   b. A Support Vector Machine model
   c. A decision tree model
   d. One other model of your choice

   Be sure to check the impact of feature scaling (on logistic and SVM), and plot an ROC/AUC curve, and a confusion matrix.

3. Switching to the wine data, construct a pair plot and correlation table. Are there are any potentially problematic variables?

4. Construct and evaluate the following models on the wine data to predict wine type:
   a. A decision tree
   b. A random forest model
   c. A KNN model
   d. An XGBoost Model
   e. A neural network model

   Be sure to check the impact of feature scaling on KNN and the neural network models. Create a table of variables of importance for the decision tree, Random forest and XGBoost. Create a confusion matrix for all the models.

5. Optional. You can also use clustering models as semi-supervised learners. Use K-Means as illustrated in class on both the beer preference data and the wine data. Assess its performance relative to the traditional classification models.