DSA 610, Homework #8, 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. What are some advantages and disadvantages of parallel processing? Are there circumstances when parallel processing can't be easily used?

2. Describe how clustering methods can be used in a semi-supervised process for classification. What are the benefits and drawbacks of such a process?

3. What are the major concerns of data ethics and how is each addressed in data management?

4. What are the specific ethical concerns around data reuse, particularly in the form of reselling data?

5. Describe some of the effects of rescaling variables in machine learning algorithms. How might different distance metrics impact the results of algorithms?

Part II:
Instructions: Use the attached dataset (**marketing_hw8.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. Explore the dataset briefly by evaluating which variables may have a significant influence on whether the customer has tried the lasagna or not (the Has Tried variable).

2. Prepare the data for your model. Convert any categorical variables to dummy variables as needed. Remove the Person variable. Rescale the variables (if the models you select in Question 3 require it).

3. Select two different classification models. For each of your models, apply the following validation strategies:
   a. Hold-Out CV
   b. K-Fold CV
   c. LOOCV
   d. Stratified CV
   e. Bootstrapping

4. Create appropriate visualizations to compare the validation methods for each model. Assess the results of the validation (not necessarily the models, themselves).