Homework Three

MSDS 598 Spring 2022

Directions

- Submit a .ipynb notebook to Canvas.
- The Notebook should begin with a Markdown cell with your **Name** and the title of the Assignment, **Homework 3**. Failure to do so will result in points lost.
- Use Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions. Failure to do so will result in points lost.
- The filename for your notebook should be formatted like

  $$FirstName\_LastName\_AssignmentName.ipynb.$$

  Failure to do so will result in points lost.
- This is due on March 7th at 6:30 PM Pacific time. Solutions will be posted Thursday after this.

For the following **five** questions you will need to use the `heart disease` dataset which can be obtained from

https://www.kaggle.com/ronitf/heart-disease-uci

Here the target variable is labeled `target` and is 0 if there is no heart disease and 1 if there is heart disease present.

1. What is the percentage of patients with heart disease? Would you consider this a balanced of imbalanced dataset (no wrong answer here!).

2. Suppose we want to use the variables `age` and `sex` to predict the presence of heart disease.

(a) Use logistic regression to create a model predicting `target` from `age` and `sex`. Report the Accuracy.

(b) If we had a model that predicted heart disease for every patient, what would the accuracy be?

3. Perform a train-validation split (use 50 patients for your validation set). What is the accuracy of your model (same variables as the previous question, created on the training set) on the validation set?

4. Pick another variable. Use this validation set and the accuracy metric to decide if this variable improves the model or not.

5. Pick your favorite set of independent variables and run a logistic regression using train-validation set from above.

(a) Produce the Confusion Matrix.

(b) Report the Accuracy, Precision, Recall, and AUC.

(c) Considering the task of predicting whether a person has heart disease or not, which metric in part (b) is the most important?