HPE DSI 311 – Introduction to Machine Learning – Summer 2021
Homework Assignment #2
Due Tuesday (July 6), 11:59 pm (Central)

Your assignment is to create a Jupyter notebook that demonstrates how to do the following (use methods discussed in the class materials shared so far):

1. Load the dataset in the file named winequality_white.csv and produce at least one table and one graph that summarize the dataset statistics. Separate the data into training and testing datasets and set up a classification problem: predicting the quality value (variable with seven classes labeled 3, 4, 5, …, 9) based on the values of all the other variables (acidity, alcohol, pH, etc.). (2 points)
2. Train and tune (via cross-validation) at least two different models based on Decision Trees (e.g., DecisionTreeClassifier, RandomForestClassifier); Consider at least two different hyperparameter options (e.g., tree depth). (5 points)
3. Train and tune (via cross-validation) at least two different SVM models based on different kernel options (e.g., linear and sigmoid) and regularization parameters (different values of C). (5 points)
4. Use the make_pipeline() method to study and describe the impact of feature selection (using different n_component values for PCA) and data scaling (e.g., MinMaxScaler) on the performance of the tuned SVM from Step 3. (5 points)
5. Test the performance of the best method you found using the test set you created. Discuss your overall results. (3 points)