HPE DSI 311 – Introduction to Machine Learning – Spring 2023
Homework Assignment #2
Due Sunday (February 19), 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 (a single variable with seven classes labeled 3, 4, 5, …, 9) based on the values of all the other variables in the file (acidity, alcohol, pH, etc.). (2 points)
2. Train and tune (via cross-validation) at least two different models; one based on Decision Trees (e.g., DecisionTreeClassifier, RandomForestClassifier) and one based on SVMs (e.g., different kernel SVC). (6 points)
3. Consider at least two different hyperparameter options for each model (e.g., tree depth, regularization value C). (6 points)
4. Use the make_pipeline() method to study and describe the impact of feature selection on the performance of the tuned SVM from Step 3. You can try dimension reduction (e.g., using different n_component values for PCA) and/or data scaling (e.g., MinMaxScaler). (6 points)
5. Train the DummyClassifier() on your training set and then compare the performance of the best method you found to the DummyClassifier() using your test set. Discuss your overall results. (2 points)

DummyClassifier():
https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

What to submit: Please name your h/w submission as follows:
311_lastName_firstName_assignmentNumber.ipynb

How to submit: Please submit homework in Moodle.