HPE DSI 311 – Introduction to Machine Learning – Summer 2024
Homework Assignment #2
Due Monday, June 24, 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. 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.) and split the dataset into separate training and test sets in a reproducible way; (**2 points**)
2. Train two models to solve this classification problem: one based on Decision Trees (e.g., DecisionTreeClassifier, RandomForestClassifier) and one based on SVMs (e.g., an SVC with your choice of kernel). Use the training set you created in part 1 to cross-validate the performance of each model. Report statistics for the scoring method of your choice (e.g., accuracy, weighted precision, macro recall, f1 score); (**6 points**)
3. Use GridSearchCV() and the training dataset from part 1 to tune the Decision Tree family model from part 2; compare its performance when changing at least two different hyperparameters (e.g., tree depth) across a range of values. (**6 points**)
4. Use the make_pipeline() method and the training dataset from part 1 to study and describe the impact of data transformations on the performance of the SVM family model from part 2. 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. Use your test set to compare the performance of this DummyClassifier() and the best model versions from parts 3 and 4. 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.