HPE DSI 311 – Introduction to Machine Learning – Summer 2024
Homework Assignment #1
Due Monday, June 10th, 11:59 pm (Central)

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

1. Load the dataset in the file named BDOShoham.csv and produce at least one table and one graph that summarize the dataset statistics; (**4 points**)
2. Set up a classification problem: predicting the FlowPattern value based on the values of the variables named Vsl, Vsg, and Ang, and split the dataset into separate training and test sets in a reproducible way; (**4 points**)
3. Train at least two models (e.g., k-NN, logistic regression) to solve this classification problem. Use the training set you created in part 2 to cross-validate the performance of each model. Report on three different scoring methods (e.g., accuracy, weighted precision, macro recall, f1 score); (**6 points**)
4. Pick a model and a scoring method from part 3. Use cross-validation to evaluate the improvement/degradation of performance when you modify at least two hyperparameters (e.g., n_neighbors, weights, metric, penalty) as compared to the model's default settings; (**4 points**)
5. Test the performance of the best model+hyperparameters combination you found in part 4, using the test set you created in part 2. Discuss your overall results. (**4 points**)

There is no "perfect solution." The objective of this assignment is to provide you with hands-on practice and an opportunity to learn. The goal is to see how the different choices you make in training affect the results that you obtain, not how to obtain the best performance in the class. Good luck!

What to submit: Please name your notebook file as follows:
311_lastName_firstName_assignmentNumber.ipynb

How to submit: Please submit your files in Moodle.