

HPE DSI 311 – Introduction to Machine Learning – Summer 2021
Homework Assignment #1
Due Tuesday (June 22), 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 BDOSham.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. Train at least two models (e.g., k-NN, logistic regression) to solve this classification problem; (4 points)
3. Evaluate each model's performance using cross-validation on the training set you created; report on at least two different scoring methods (e.g., confusion matrix, weighted precision, macro recall, f1 score); (4 points)
4. Modify at least two hyperparameters (e.g., n_neighbors, weights, metric, penalty) and describe the improvement/degradation of a model's performance compared to its default settings; (4 points)
5. Test the performance of the best model+hyperparameters combination using the test set you created. Discuss your overall results (4 points)