BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 4

Please submit a single **PDF** file with all your answers, discussion, plots, etc. **on BrightSpace and Feedback Fruits**. You can use any software package/language you like, though I recommend Python/pandas/sklearn. Weka, R, and MATLAB should also work for these tasks.

You are given a 3-class dataset of 300 homes labelled by their overall energy efficiency: 0=low, 1=med, 2=high. Each home is described using five features that can be measured from an instrumented vehicle parked on the street in front of the home. Our goal is to develop a diagnostic tool that will classify each house.

1. Load the dataset in `A4.txt`. The column names correspond to the five features plus the class ID:
   `colNames = ['Thermal', 'Area', 'Glazing', 'Clading', 'Roofing', 'Efficiency']`

2. Split your data into train/test using a 75/25 split and stratified sampling. **Report** the number of samples from each class in your train and test subsets.
3. Using the training set, for each feature, **plot** the feature distribution for each class. You can either use five histograms or five 1D kernel density plots. Label each sub-plot by the feature **name**. The distribution of feature values should be visible for all three (potentially overlapping) classes on each of the five plots. **Which feature looks most useful and why? Which home efficiency class do you think will have the lowest accuracy and why?** *(60 words max)*
4. Complete 5-fold-cross-validation over the train subset using an SVM classifier with a polynomial kernel with degree=3 and C=0.8. **Report** the accuracy over each fold, the average accuracy across all five folds, and the standard deviation across the five accuracy measurements.
5. Train another SVM model (same kernel & C) on all of your training samples. Test on the test subset. **Report** the accuracy on the test subset. Does it fall within 1 standard deviation of the average accuracy observed in Step 5?
6. For this question only, assume that the misclassification costs are as follows:

		Actual		
		Low	Med	High
Predicted	Low	0	1	2
	Med	1	0	1
	High	2	1	0

   a. What is your total misclassification cost for the test set predictions from Q5 above?
   b. How could you incorporate this loss information into your classifier design? *(60 words)*
7. Using 5-CV across only the training subset, perform a hyperparameter sweep of the number of hidden nodes in a 3-layer feedforward neural network. **Report** your accuracy for `numH=[1, 10, 100]` hidden nodes. Use the 'adam' solver, a hyperbolic tangent activation function for the hidden layers.
8. Returning to Question 3, compare a naïve Bayes classifier trained using only the 'most useful' feature to a naïve Bayes classifier trained using all five features. Describe how you split/used your data, how you tested the hypothesis (null hypothesis, alternative hypothesis, test metric, etc.), what p-value you obtained, and your conclusion.