AI 221: Machine Exercise 1

Instructions:

- Read and answer each problem using computer code. This MEX should be done *individually*.
- Each item should be answered as either a Jupyter Notebook or a MATLAB Live Script, exported as a single PDF file for the entire MEX. Make sure to HIGHLIGHT your final answers.
- When done, submit the PDF file through UVLE.

Problem 1: Energy Efficiency in Buildings

Go to https://archive.ics.uci.edu/ml/datasets/Energy+efficiency.

Download the Energy efficiency dataset. The dataset contains 768 samples of simulated buildings with 8 attributes (X1 to X8) and two targets (Y1 and Y2). The following are their meanings:

X1 Relative Compactness	y1 Heating Load
X2 Surface Area	y2 Cooling Load
X3 Wall Area	
X4 Roof Area	
X5 Overall Height	
X6 Orientation	
X7 Glazing Area	
X8 Glazing Area Distribution	

The predictive model for heating and cooling loads of these buildings are useful for analyzing their energy consumption, in particular, in cold countries. In this problem, let's predict the *heating load only*:

a. **[30 pts]** Split the samples into 60% Training, 20% Validation, and 20% Testing data at random. Build a pipeline with Standard scaler then linear ridge regression. Set your own 10 different choices of regularization, find the best choice that gives the highest accuracy on the validation data, then make one final evaluation on the test data. What is the best model's coefficients, intercept, and its training, validation, and test accuracy?

b. **[30 pts]** Based on your answer in item (a), what are the top 5 features among X1 to X8? If you repeat the procedure above using only the 5 top features, what are the results?

Based on your results for this Problem, what insights did you gain?

Problem 2: Classifying Breast Tumors

Go to https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Download the Wisconsin Breast Cancer dataset. The dataset contains 699 instances of breast tumors with the following attributes:

1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10

4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	(2 for benign, 4 for malignant)

From the raw data set, remove rows with missing values, remove the column “Sample code number” and replace the “Class” values into 0’s and 1’s (0 for benign, 1 for malignant). You may choose to do this in Python (Pandas) or manually in Excel. The goal is make a classifer for the tumor status.

a. **[20 pts]** Split the samples into 70% Training and 30% Testing at random. Make sure to use “stratify=y” in the test_train_split function. Build a pipeline using the Standard scaler and logistic regression. Use the default penalty settings of Logistic Regression. After fitting the data, what is the model’s training and testing accuracy? Which features are most important?

b. **[20 pts]** From your answer in item (a), generate a confusion matrix, then calculate the other metrics: F1-score, Precision, Recall, and False alarm rate. Finally, plot the ROC curve and report the AUC. For this item, make a result for both the training and testing data, separately.

Based on your results for this Problem, what insights did you gain?

END OF EXERCISE