AI 221: Machine Exercise 2

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. Palmer Penguin Species Data Set

Download the Palmer Penguins Data set from:

https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data

The dataset contains 345 penguins from the Palmer Archipelago, Antarctica. Similar to the Iris Flowers Data set, it contains 4 numerical features of the penguins, namely: culmen length, culmen depth, flipper length, and body mass index.

Your task is to classify the penguins into their species (Adelie, Chinstrap, and Gentoo) based only on the **culmen length** and **flipper length** features.

a. **[5 pts]** First visualize the 4 numerical features of the data using Seaborn's pair plot, then set the hue to the penguin species.

b. **[20 pts]** Split the samples into 75% Training and 25% Testing data at random with stratification (stratify=y). Build a pipeline with Standard scaler then SVC. Train the model using the default settings for multi-class SVC in sklearn. Report the accuracy, macro-averaged F1-score, and confusion matrix of the trained model separately for the training data and testing data.

c. **[5 pts]** For your answer in letter (a), visualize the decision boundary in the space of culmen length vs. flipper length. Add a scatter plot of the training and test data set (use different markers for the two sets).

d. **[20 pts]** Find a better model by varying the box constraint, kernel function, kernel parameter, and multi-class strategy. Evaluate only at least 5 candidate SVC models, each having a certain combination of settings. Report the metrics of the best SVC model that you found, same as in (a).

Problem 2: Predicting Bike Sharing Demand in Seoul, South Korea

Go to https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

Download the Seoul Bike Sharing Demand dataset. The dataset contains 8760 instances of weather characteristics, holiday, season, and the number of bikes rented for that hour.

Take **only the weather data + hour of the day** as inputs: Hour, Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, and Rainfall. We'll also take only the **Winter** data to keep the number of data points manageable for kernel methods.

Your goal is to predict the bike demand (**Rented Bike Count**).

a. **[5 pts]** First, visualize the weather data and the Rented Bike Count using box plots.

b. **[20 pts]** Split the data into 70% training and 30% testing at random. Make a pipeline using Standard Scaler and SVR. Train the model using the training set, then report the RMSE (root mean squared error) on the Test Set. You can fine-tune your own SVR by changing the kernel function, kernel parameter, and epsilon.

c. **[20 pts]** Do the same as (b) but now using Standard Scaler + KRR. You can fine-tune your KRR by changing the kernel function, kernel parameter, and regularization (alpha).

d. **[5 pts]** Do the same as (b) but now using simple Linear Regression. Compare the results of your SVR, KRR, and Linear Regression.

END OF EXERCISE