AI 221: Machine Exercise 2

Instructions: Read and answer each problem using computer code. Each item should be answered as a Jupyter Notebook. Make sure to HIGHLIGHT your final answers.

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. **[25 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 multi-class SVC in sklearn (You can try any arbitrary choice of kernel function, values of kernel parameters, multi-class strategy, and other hyper-parameters). Report the accuracy, macro-averaged F1-score, and confusion matrix of the trained model separately for the training data and testing data.

c. **[20 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).

Problem 2: Predicting Chlorophyll-A (Green color) Levels in Laguna Lake

Go to https://github.com/kspilario/predict_chlorophyll/blob/main/chlor.csv

Download this Chlorophyll-A (Chl-a) dataset collected from Laguna Lake in 2023. The dataset contains Chl-a level as outputs (1st column) and nutrient levels as inputs (3rd to 6th columns). All other columns are irrelevant for this problem.

Your goal is to predict the Chl-a levels in Laguna Lake. This is important because Metro Manila gets drinking water from treating Laguna Lake water. If the Chl-a level is too high, water shortages may occur. Answer the following:

a. **[10 pts]** First, remove the rows with missing data via Pandas. Visualize the data as a pair plot.

b. **[20 pts]** Split the data into 70% training and 30% testing at random. Make a pipeline using Standard Scaler and Gaussian Process Regressor. 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 model by changing the kernel function, kernel parameter, and epsilon.

c. **[20 pts]** Do the same as (b) but now using Standard Scaler + Kernel Ridge Regression (KRR). You can fine-tune your KRR by changing the kernel function, kernel parameter, and regularization (alpha). Compare the results of GPR versus KRR in terms of accuracy.

END OF EXERCISE