MS Hall - Data Analytics Day 2

Problem Statement:

A botanist aims to classify different species of iris flowers based on various physical characteristics, such as sepal length, sepal wi-dth, petal length, and petal width, using the Iris Dataset. The goal is to develop a predictive model that can accurately identify the species of an iris flower given these measurements. The data is shared in the WhatsApp Group. Develop a Python code to perform exploratory data analysis, correlation analysis, and build a classification model from the data.

Tasks

Task 1: Do Data Imputation if required.

Task 2: Plot histograms for all features. Tabulate the correlation coefficients for all the columns (including the target). Plot the corresponding correlation matrix heatmap.

Task 3: Divide dataset into two subsets: dataset _features (having all columns except target) and dataset _target (having only target column). This constitutes data and corresponding labels. Divide each of dataset_features and dataset _label into training and testing subsets (90% data for training, 10% data for testing, no shuffling). Print the shapes of both training and testing subsets for dataset_features and dataset_label.

Task 4: Predict the outcomes using regression algorithms. Study the effect of different learning rates (a) 0.001, (b) 0.01, (c) 0.1. Calculate the RMSE of predicted data with the different learning rates. Report the optimal learning rate out of the abovechoices, and print the corresponding optimal values for coefficients and intercept.

Bonus Task: Also use any clustering algorithm to predict the outcomes.