Cohort 10
C3 PGP AI/ML
Assignment 1 - K-Nearest Neighbours algorithm

The objective of this assignment is to train a classification model for the given dataset using KNN algorithm. You are free to use Sklearn implementation directly for this assignment. Following the basic model, you need to experiment with different k values and observe classification accuracy changes with changes in the k-value. Detailed instructions are following:

Dataset: Attached with assignment in Canvas, please download the CSV from there

Tasks:

1. Load the dataset ‘pumpkin seeds’ mentioned above and apply following analysis/feature engineering techniques on it:
    a. Check numerical columns for outliers and remove, if any
    b. Normalize the numerical columns in the dataset
    c. Plot frequency distribution of the class variable (“Class” in the CSV file) using countplot() function from Seaborn library [1+1+1 marks]
2. Split the dataset into training and testing sets using train_test_split from sklearn.model_selection [0.5 mark]
3. Use KNeighborsClassifier from sklearn.neighbors to train a classification model with k=3 and predict on the training data. Print the classification report for training data [1+0.5 mark]
4. Use model trained above to predict on test set and print the classification report for test predictions [1 mark]
5. Implement a loop to try different values of 'K' (number of neighbors). Train with following k values [5,7,9,11] [2 marks]
6. Plot the classification accuracy obtained on test dataset on y-axis for k=3,5,7,9,11 with k on x-axis [1 mark]