Tasks to do:
1. Install the numpy, matplotlib, and scikit-learn libraries.
2. Familiarize yourself with the functionalities available in the sklearn.datasets submodule https://scikit-learn.org/stable/datasets.html#datasets. The tasks in this lab will be performed on the Iris dataset.
3. Familiarize yourself with the documentation of the KNeighborsClassifier class and the description of the k-nearest neighbors method.
https://scikit-learn.org/stable/modules/neighbors.html#classification.
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
4. For the Iris dataset, do the following:
Iris: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris
(a) Generate/load the dataset. For Iris, there will be four variables (columns) and 150 samples (rows).
(b) Split the data into training and testing sets using the train_test_split function with a ratio of 0.7 for training data and 0.3 for testing data:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection. train_test_split.html.
(c) Train the KNeighborsClassifier on the training data with a specified number of neighbors N.
(d) Calculate the accuracy of the classification on the test data.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html.
(e) Repeat steps 4c and 4d to find the optimal number of nearest neighbors (the one that results in the highest accuracy). You can loop through the previous steps and plot a graph showing the relationship between the accuracy and the number of neighbors N.
(f) For the selected value of N, do the following:
i. Train the classifier on the training data with the selected value of N.
ii. Create a confusion matrix showing the accuracy of the classification on the test data (https://scikit-learn.
org/stable/auto_examples/model_selection/plot_confusion_matrix.html#confusion-matrix).
iii. Create a scatter plot showing the actual class division of the data, and then create the same plot using the predicted results of the trained classifier instead of the actual class vector. The points on the plot must be colored according to their class!
For the Iris dataset, the visualization should be done using the first two variables (sepal length, sepal width).
iv. Additionally, for the Iris dataset, create a three-dimensional scatter plot based on three variables (sepal length, sepal width, petal length). The points must be colored according to their class.