**BT-3172: Special Topics in Bioinformatics**

**Lab10: Machine learning in Bioinformatics-I: Introduction.**

**Name: Anushka Udara**

**Index number: s14234**

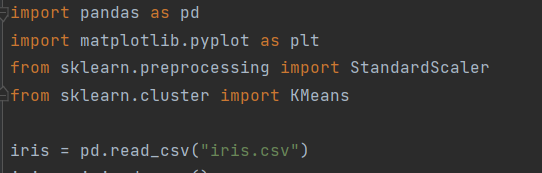
In this practical, you will learn how to use Python scikit-learn package to implement K-means and K-nearest neighbors (KNN) algorithms to learn from the popular iris dataset.

After using PyCharm to write your scripts, **copy the codes to the appropriate space below the questions**. Also, submit the Python files separately so we can test them. Use the following format to name each script: YourIndexNo\_PrimaryQuestion.py (submit two programs for the two questions)

1. K-means algorithm.

In this problem, you will learn how to implement the K-means clustering algorithm on the famous iris data set (Fisher 1936). The iris data is available in the “iris.csv” file.

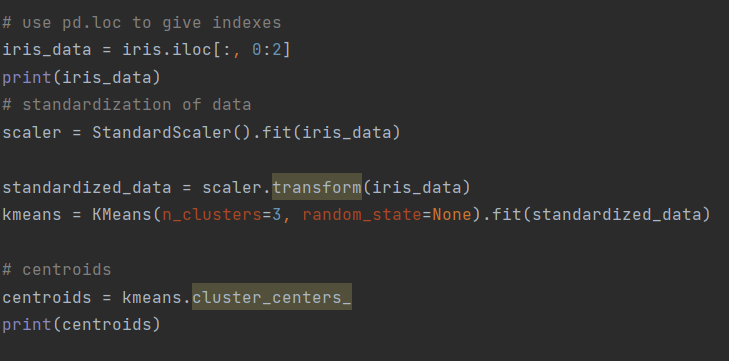
* 1. Import the necessary packages and read the iris data file into a Pandas DataFrame.

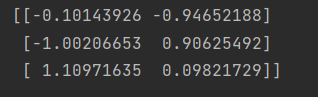


* 1. Implement the K-means algorithm using the scikit-learn package on **sepal length and sepal width** data. Pick the K value for the algorithm and explain the reason for the K value selection below. Print and write the sepal length and width values for the centroids below.

K: 3

Reason: This value was selected based on the plot results of the kmeans algorithm. The number of clusters that can best separate the data into clusters, is chosen as the K value.

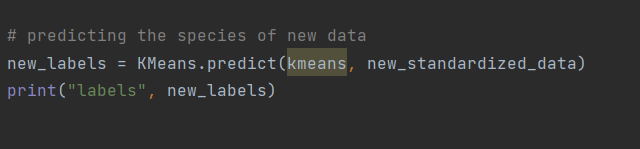




* 1. A student has measured the sepal length, sepal width, petal length, and petal width of two species-unknown iris plants (measured in centimeters), which is given below in the above specified order.

Plant1: 4.6,3.0,1.5,0.2

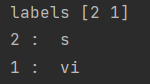
Plant 2: 6.2, 3.0,4.1,1.2

****He suspects that these plants fall into either one of the three iris species: *Iris setosa*, *Iris versicolour*, and *Iris virginica*. Use the K-means model you trained in the question (II) to predict the species of the two plants**using only the sepal length and width variables**. Write down the predicted labels below.

* 1. Draw a scatter plot for the sepal measurement data. Represent each cluster in different colors (except red).Show the centroids of each cluster in red color and the predicted two new observations in another color. Show the cluster label for each data point except the centroids.

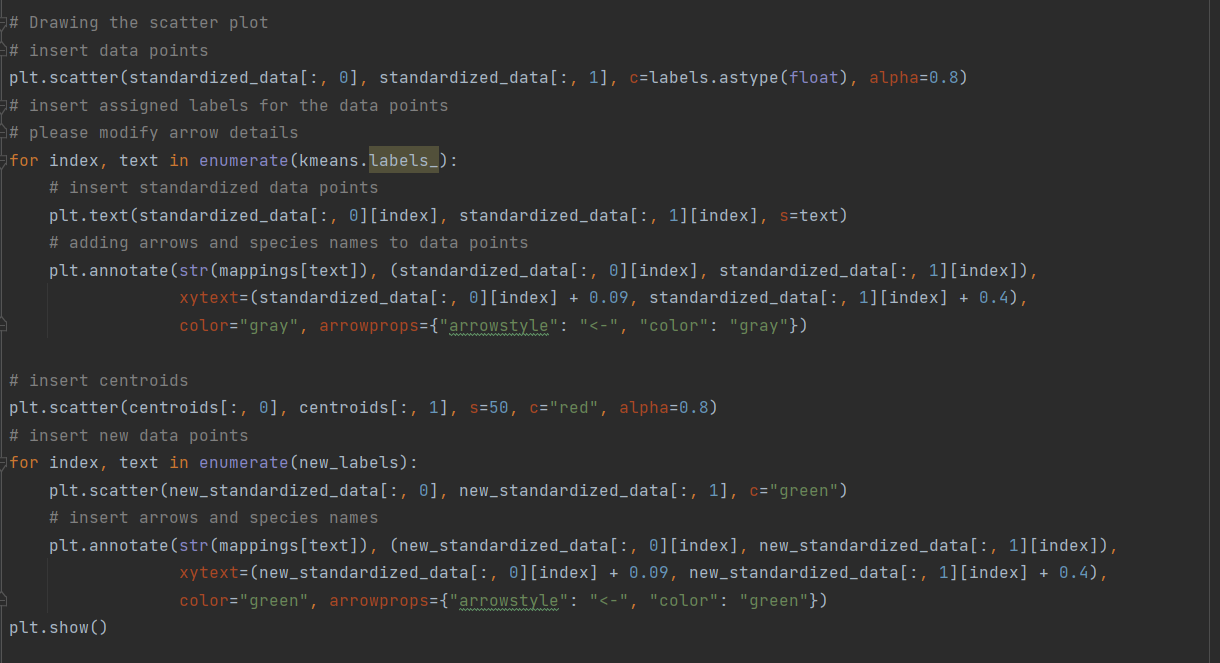
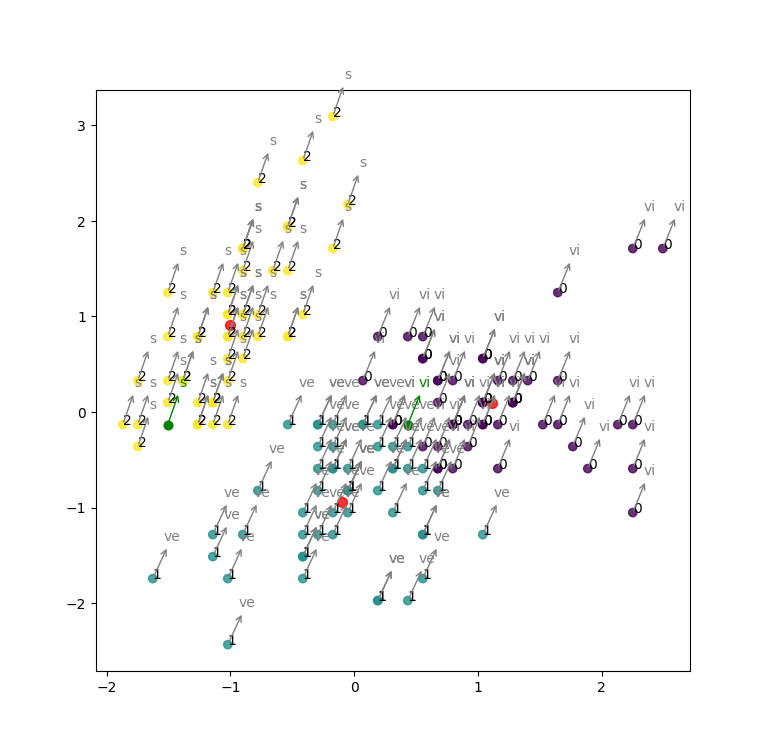
Hint: use plt.annotate() function in matplotlib to label data points.

Show the original species labels in the scatter plot using an arrow to point at each data point. Use the v\_short column to obtain the shortened labels. The species names and cluster labels should not overlap. Using these labels identify the species of the above predicted plant labels and write them for each plant in the space below.

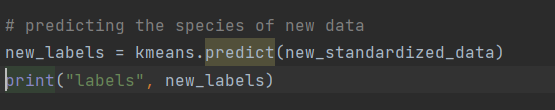


2: *Iris setosa*

1: *Iris virginica*



* 1. Now, implement the K-means algorithm using the scikit-learn package on petal length and petal width data. Use this trained model to predict the species of the two plants (question III) **using only the petal length and width variables**. Write down the predicted labels below.

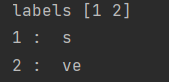




* 1. Draw a scatter plot for the petal measurement data. Represent each cluster in different colors (except red). Show the centroids of each cluster in red color and the predicted two new observations in another color. Show the cluster label for each data point except the centroids.

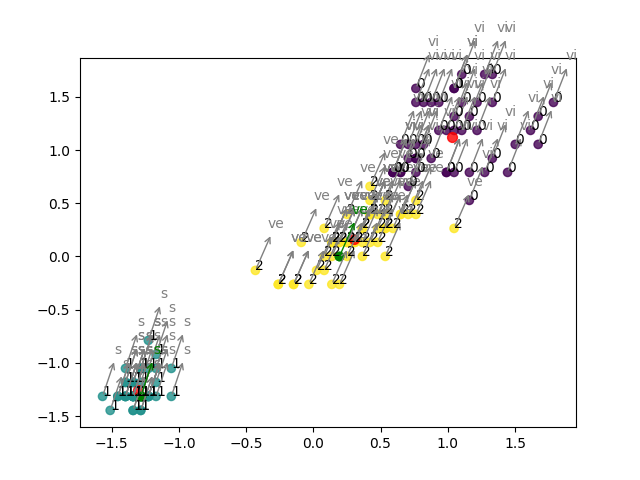
Hint: use plt.annotate() function in Matplotlib to label data points.

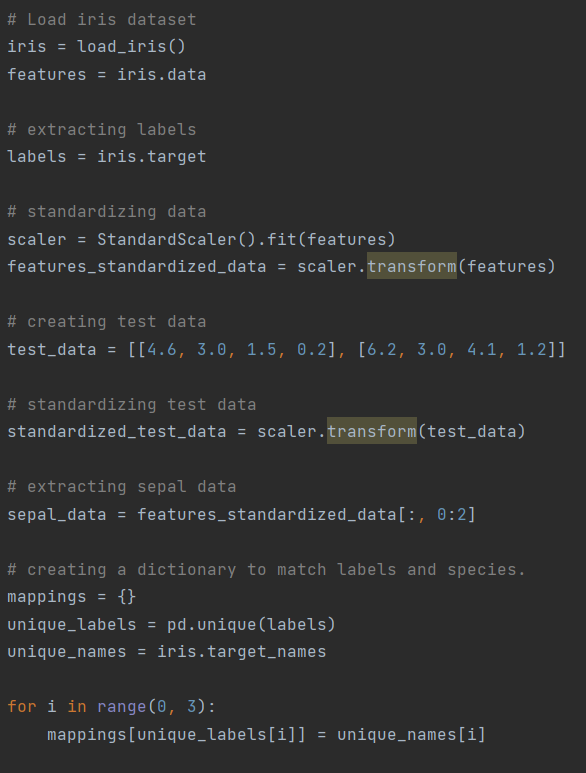
Show the original species labels in the scatter plot using an arrow to point at each data point. Use the v\_short column to obtain the shortened labels. The species names and cluster labels should not overlap. Using these labels identify the species of the above predicted plant labels and write them for each plant in the space below.

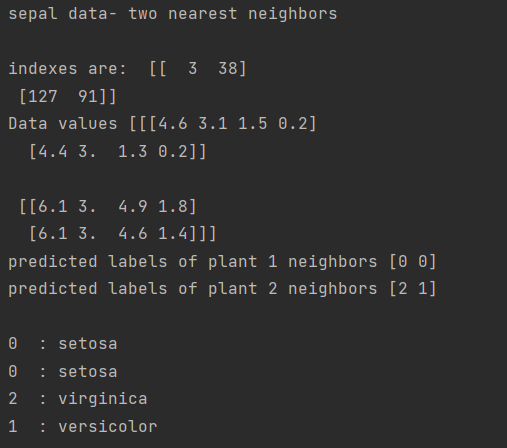
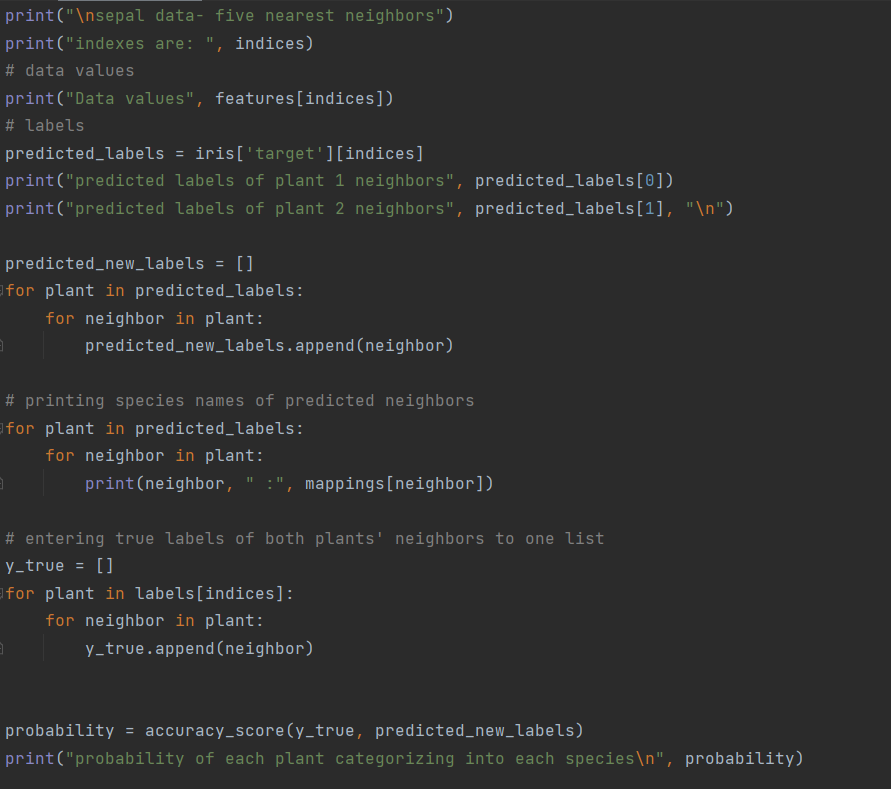


1: *Iris setosa*

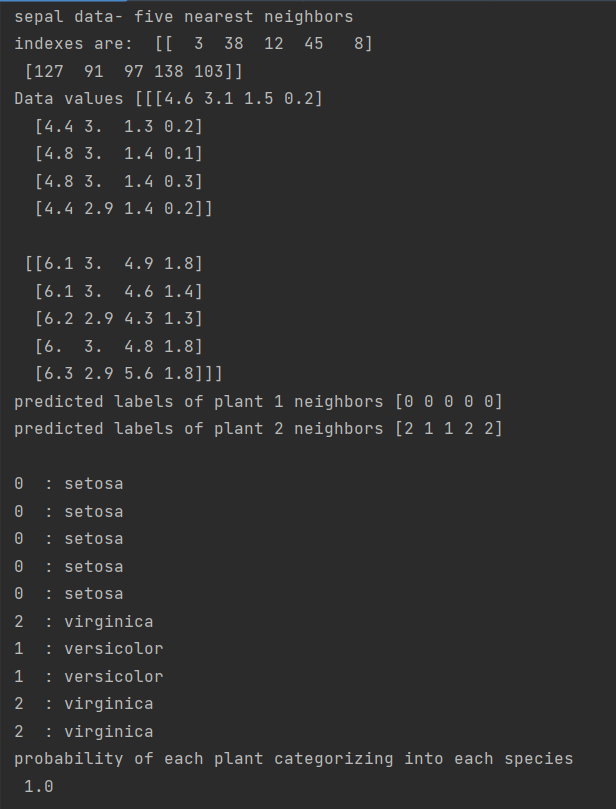
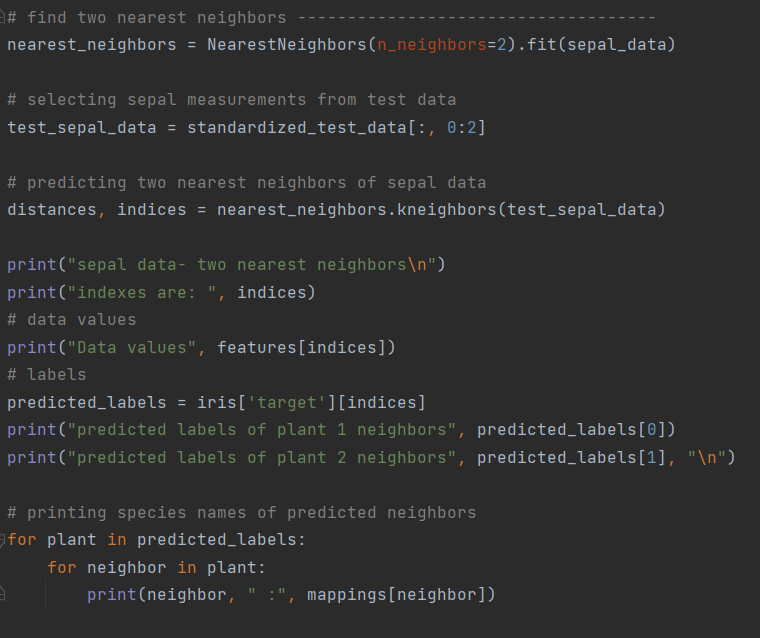
2: *Iris versicolour*

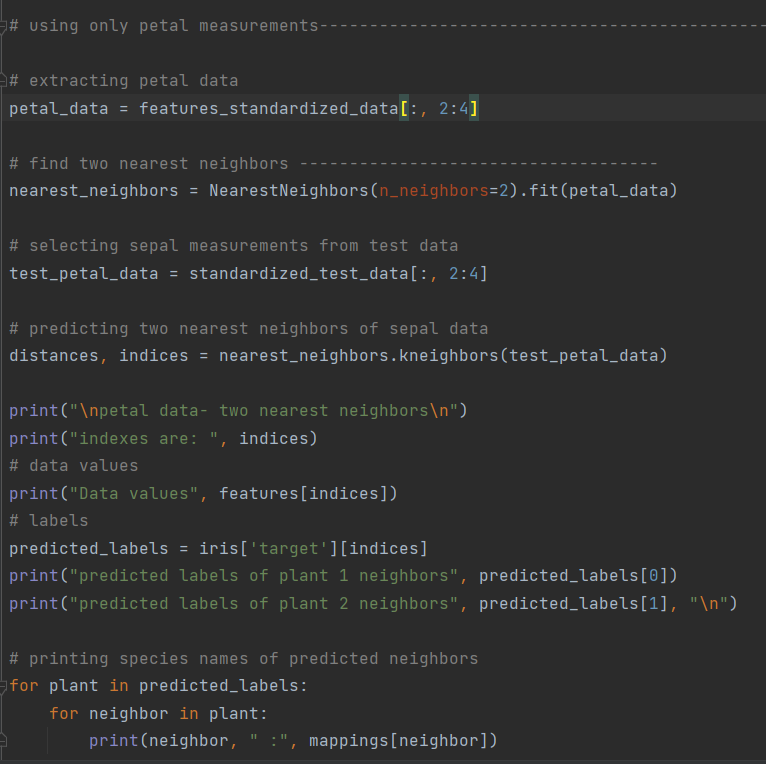
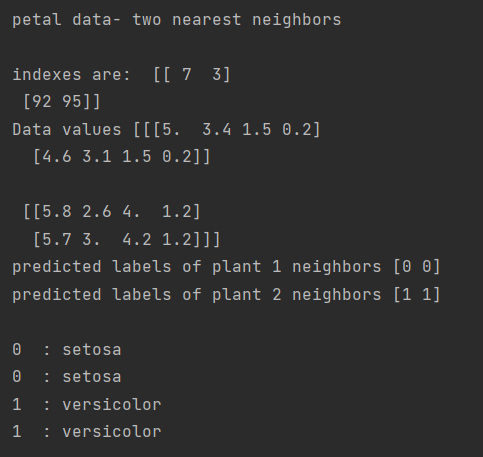
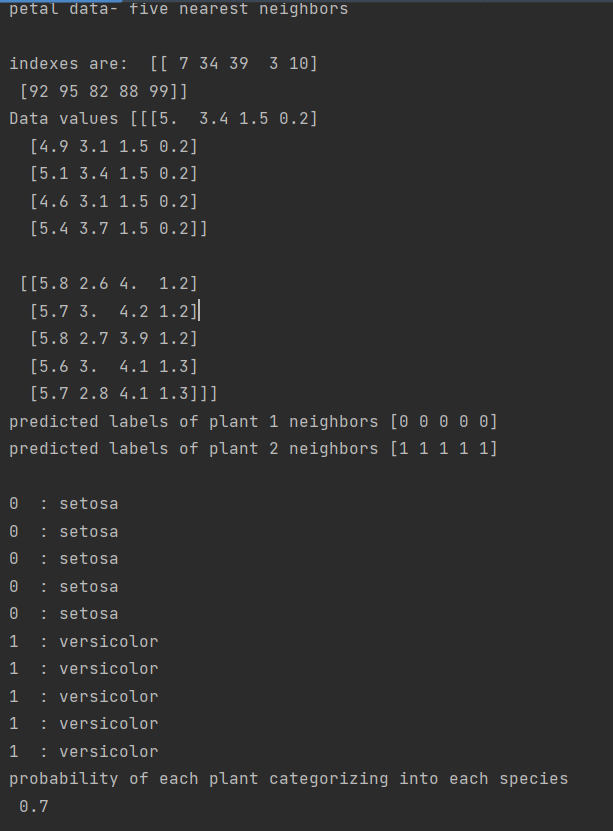


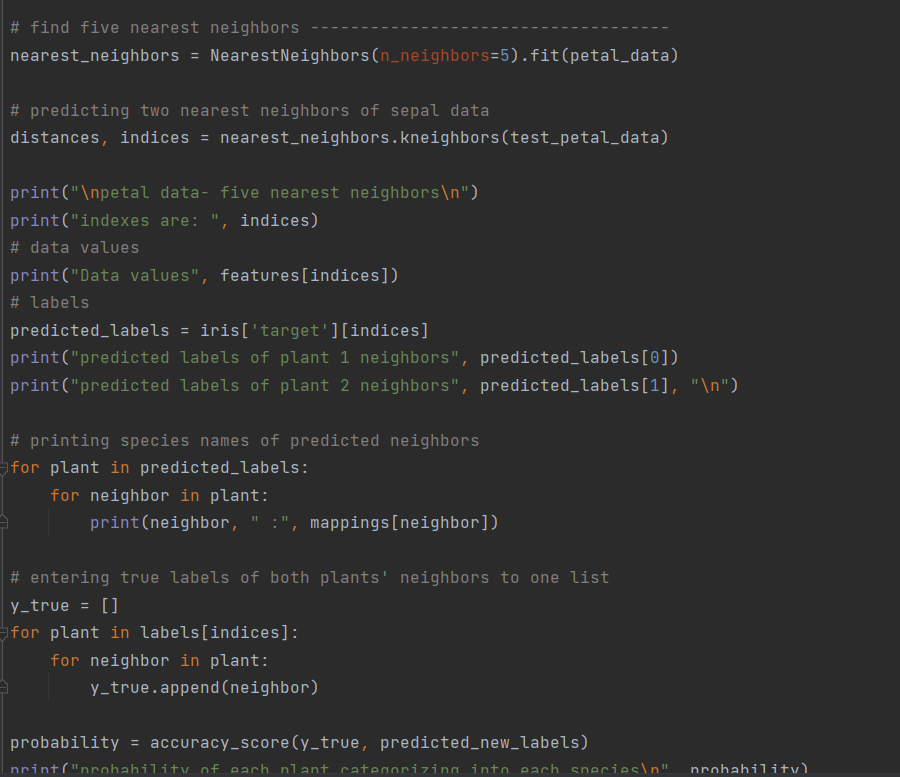
1. K-nearest neighbors (KNN) algorithm.
   1. Now, train the iris data set you used for the question 1 using the KNN algorithm. However, for this instance, do not use the CSV file to read the data, instead use the bundled iris data set in the scikit-learn package by importing the datasets sub module. Load the iris data and extract the labels. Then, standardize the data using the StandardScaler class in the preprocessing sub module. Also, read the Plant 1 data record in question 1(III) as a test data record and standardize it using the same transform model used for standardizing the iris data set. Furthermore, create a 2-d array for the two data records (Plant 1 and Plant 2) and standardize the array as before.
   2. First, train the KNN model using **only the sepal measurements**.
      1. Extract the standardized sepal data and train the KNN model to find the two nearest neighbors of the test data record (question I). Find the two-nearest neighbors and print the indices, data values, labels, and species of the two-nearest neighbors.

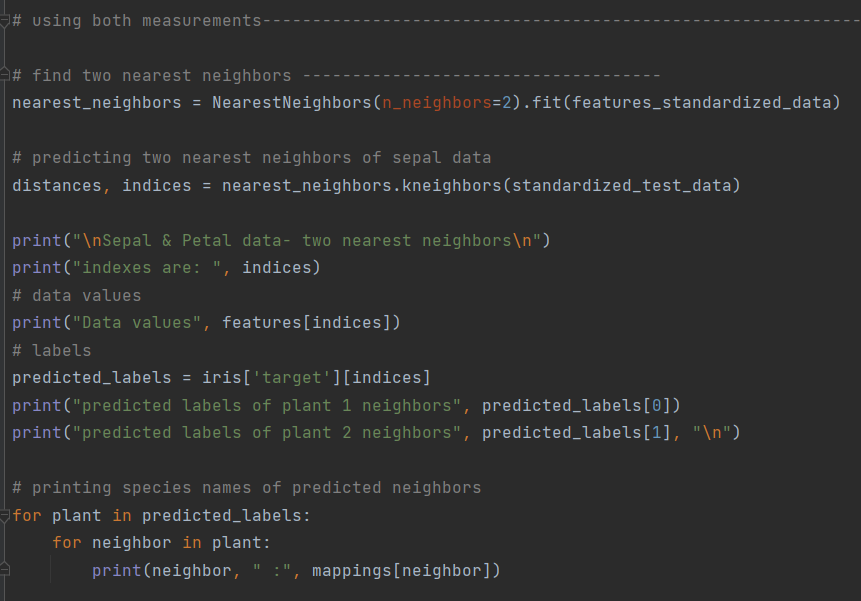


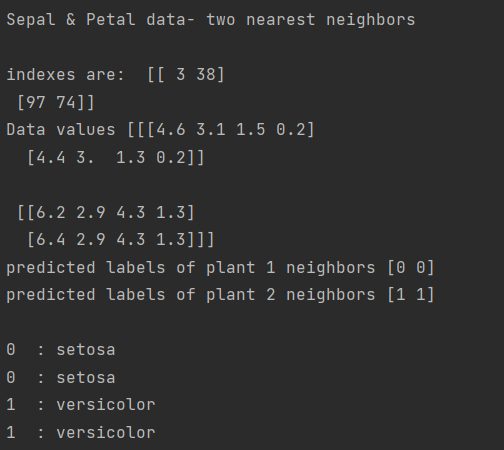
Then, train the model again for 5-nearest neighbors for standardized sepal data and predict the probability of each plant categorizing into each species and write them below. Further, predict the class labels for the two plants and write the labels and the species predicted from the KNN algorithm.



* 1. Train the KNN model using **only the petal measurements**.
     1. Extract the standardized petal data and train the KNN model to find the two nearest neighbors of the test data record (question I). Find the two-nearest neighbors and print the indices, data values, labels, and species of the two-nearest neighbors.
     2. Then, train the model again for 5-nearest neighbors for standardized petal data and predict the probability of each plant categorizing into each species and write them below. Further, predict the class labels for the two plants and write the labels and the species predicted from the KNN algorithm using only the petal measurements.

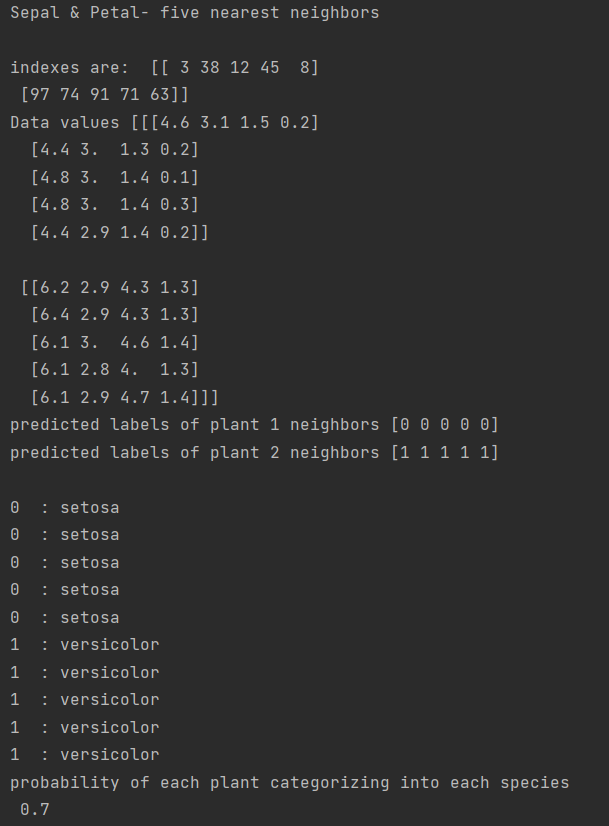


* 1. Train the KNN model using **both the sepal and petal measurements**.
     1. Train the KNN model for all the standardized measurement data to find the two nearest neighbors of the test data record (question I). Find the two-nearest neighbors and print the indices, data values, labels, and species of the two-nearest neighbors.



* + 1. Then, train the model again for 5-nearest neighbors for all measurement data and predict the probability of each plant categorizing into each species and write them below. Further, predict the class labels for the two plants and write the labels and the species predicted from the KNN algorithm.





**References**

* Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." *Annals of eugenics* 7.2 (1936): 179-188.