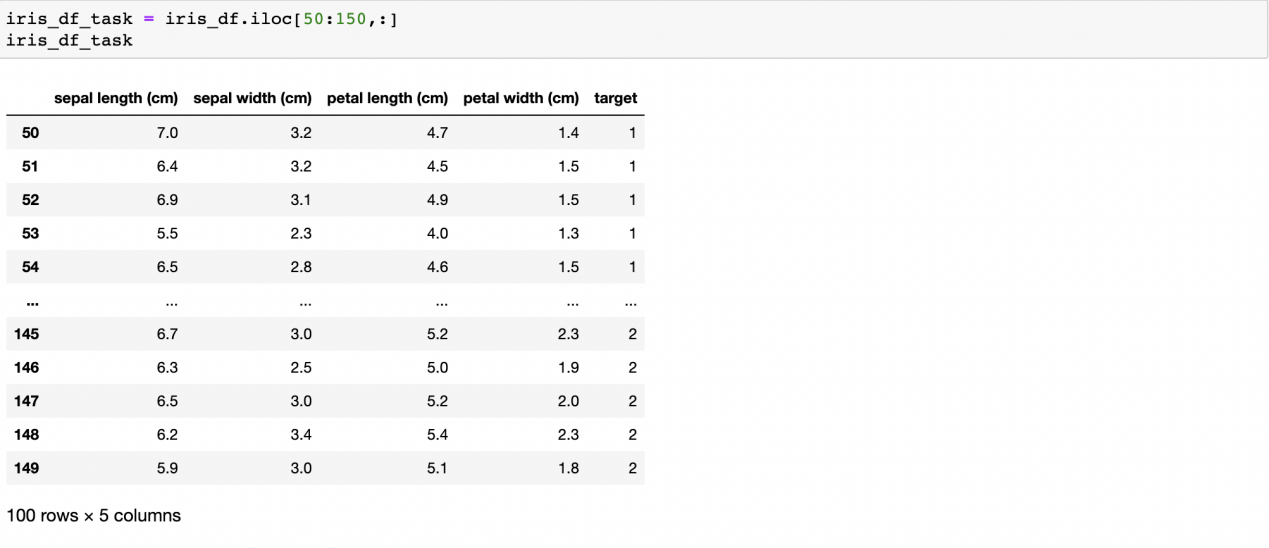
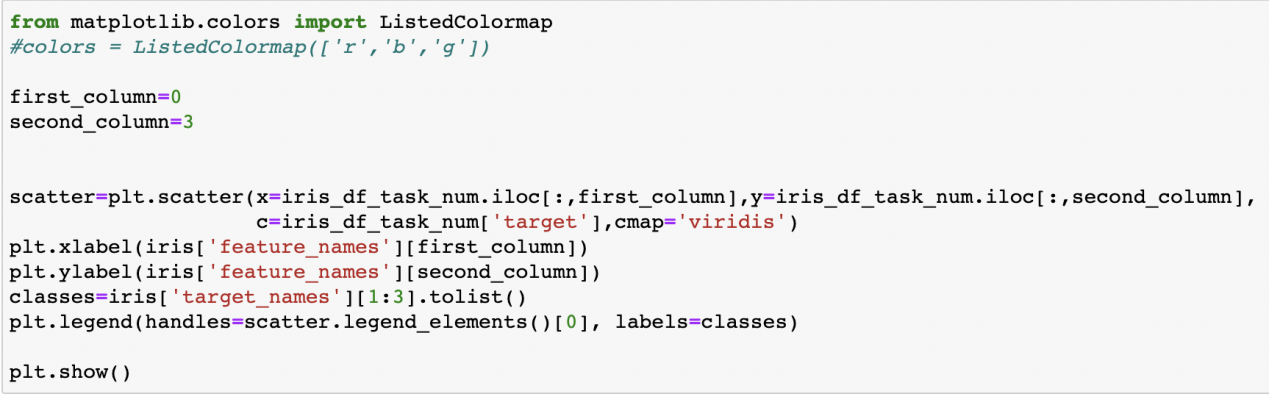
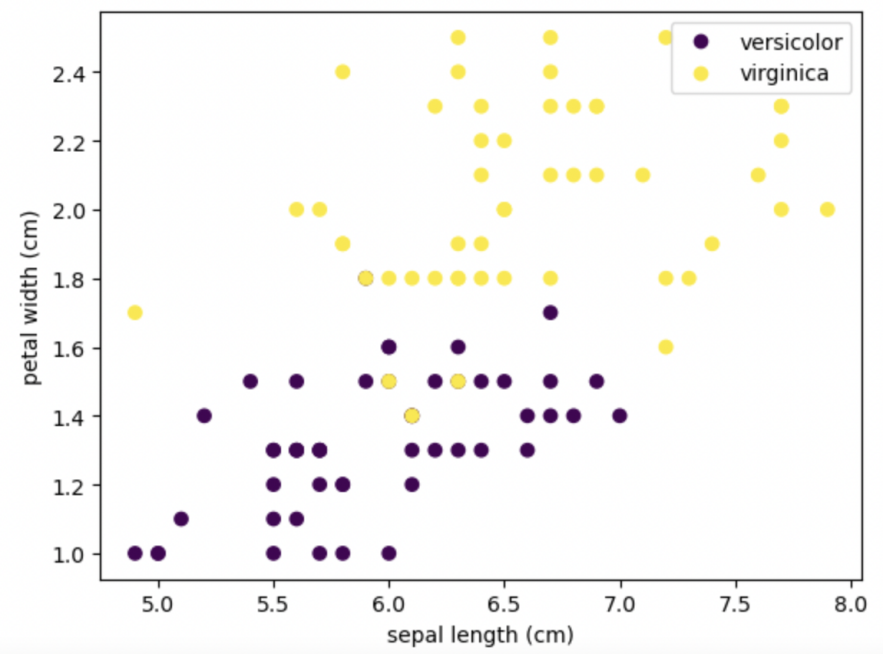
**Report - Lab1**

Iris flower Data exploration, visualization (plotting) and basic ML in Python.

**1. Make a scatterplot with Sepal length on x-axis and Petal width on the y axis. Species are color labelled and a Legend shows the color labelling. Comment on the graph, with respect to Machine Learning classification tasks.**

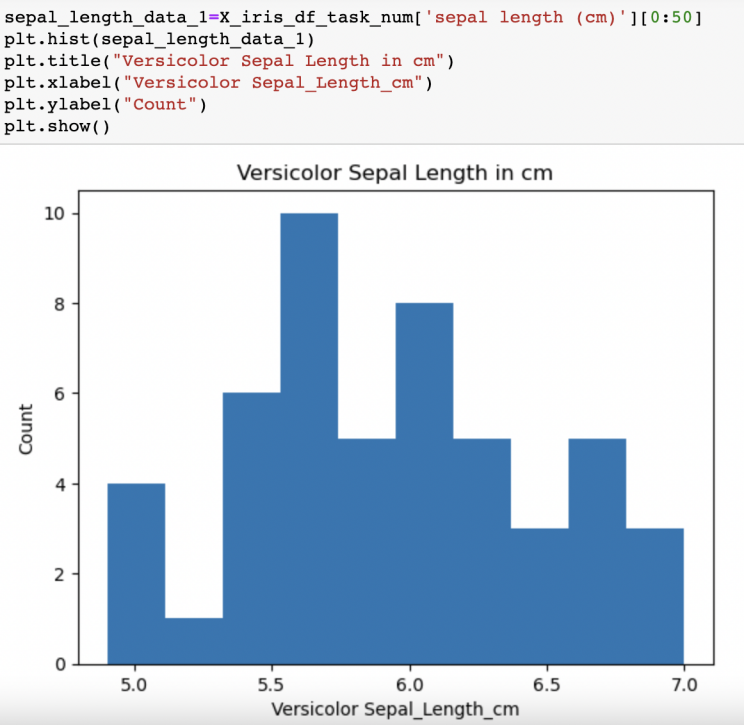


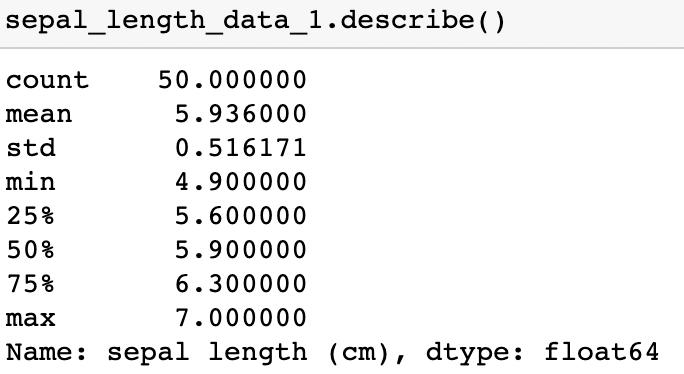


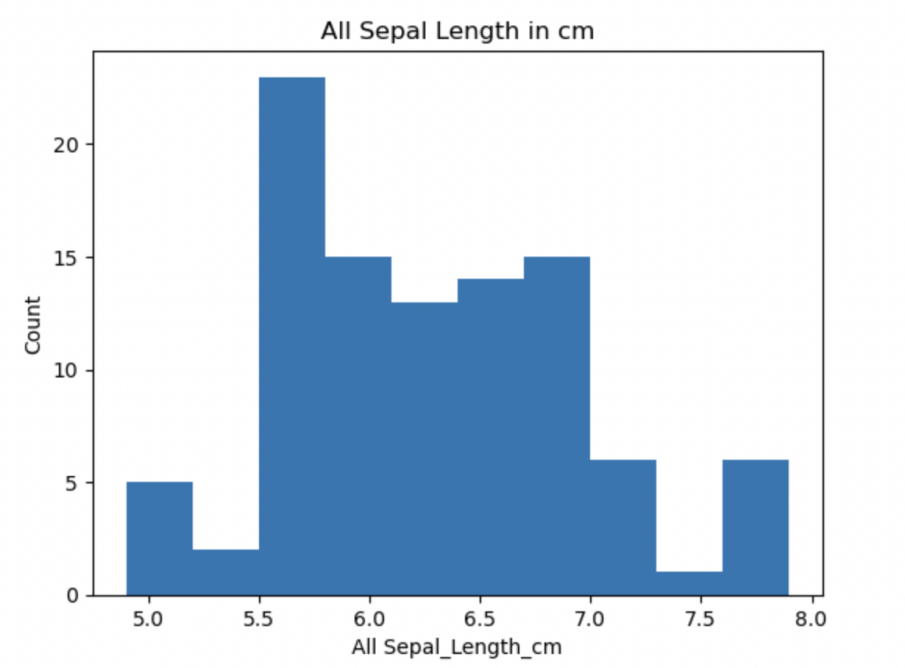


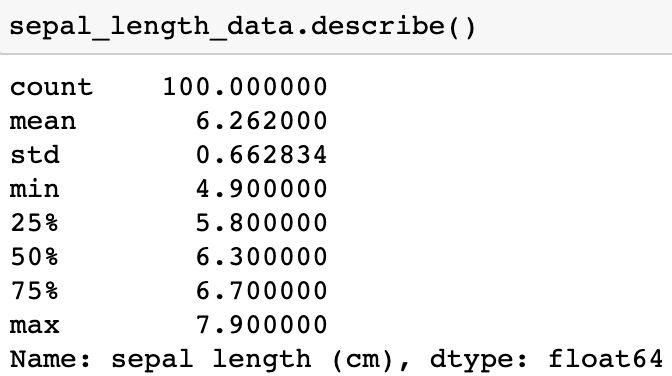
The table does not accurately display 50 data points for each category. In fact, the displayed number is less than 50. This is because some data belong to different data points but have the same corresponding values, so the number of points that can be counted is less than that of the number of original data.

**2. Make a histogram of Versicolor Sepal Length. Make another histogram of both species Sepal Lengths. Calculate corresponding mean values and standard deviations.**





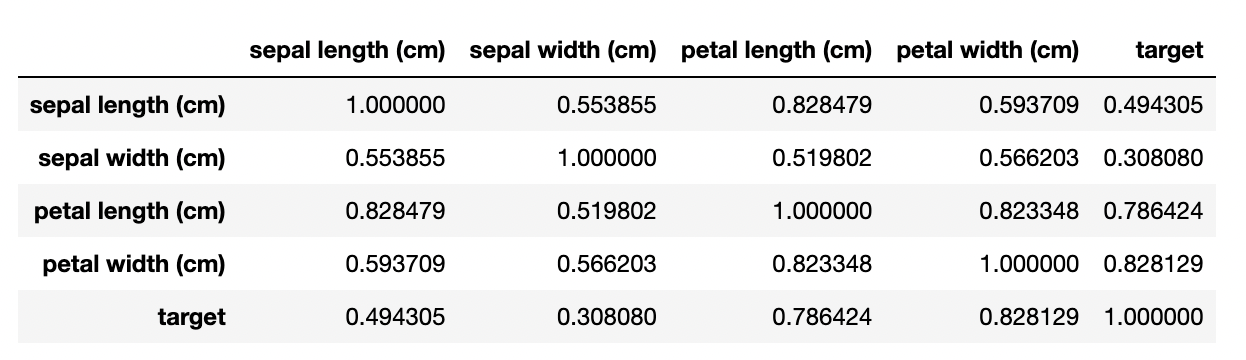




For Versicolor, the average value is about 5.936 , and the standard deviation is about 0.516. For these two species, the average value is about 6.262, and the standard deviation is about 0.663.

The standard deviation can reflect the degree of dispersion among individuals in the group, and is a measure of the degree of dispersion of the mean value of a group of data. A larger standard deviation means that most values are farther from their mean; a smaller standard deviation means that the values are closer to the mean. The smaller the standard deviation, the more clustered the data; the larger the standard deviation, the more discrete the data.

**3. Calculate the correlations of the dataframe columns.**

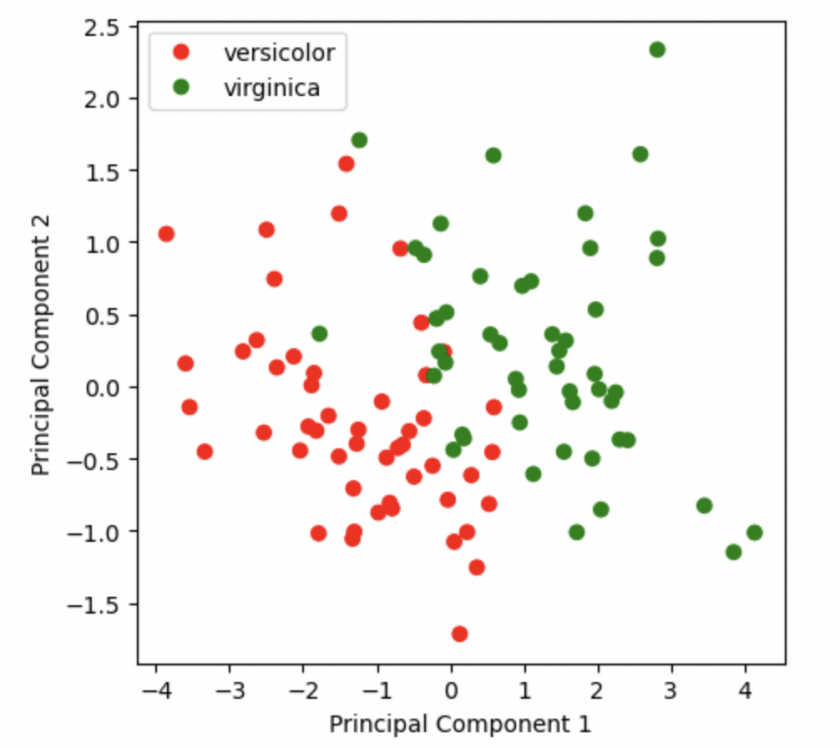
 **a. Which variable correlates best with the Species column ?**

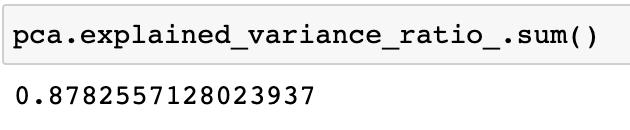
It can be seen from the above table that the correlation between petal width and target columns is stronger, because its Pearson correlation coefficient is close to 1 (0.828129), so if I can only input one column of data, I will choose petal width.

**b. Can any column be dropped from a correlation perspective ?**

From the above table, it can be seen that the correlation between petal length and sepal length is stronger, because the Pearson correlation coefficient is close to 1 (0.828479), so if I need to discard or drop a certain column, I will choose petal length or sepal length.

**4. Make a 2D PCA plot labelled with species, and explain how well it approximates all 4 dimensions.**

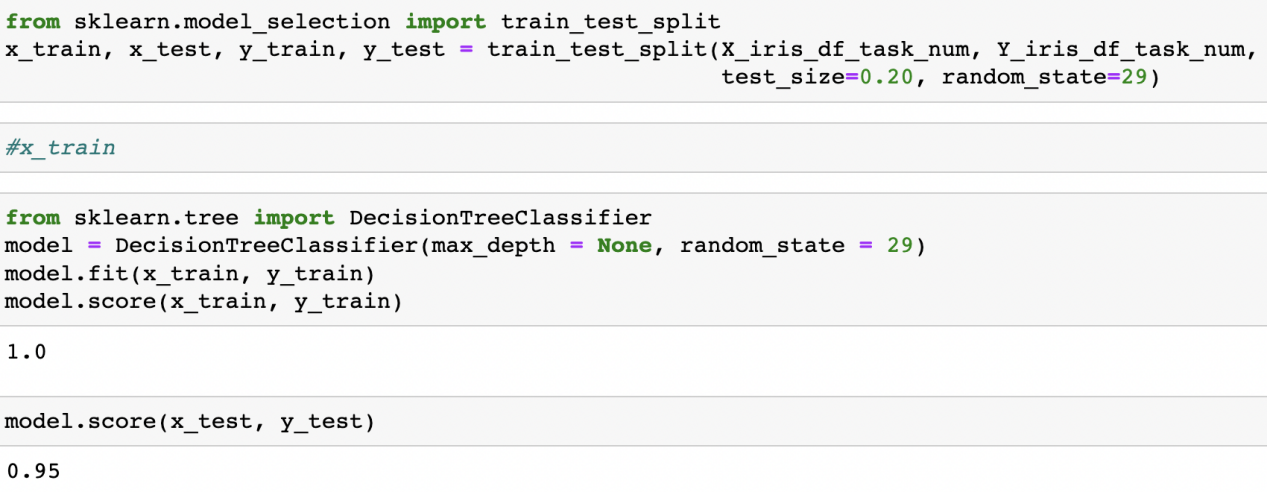




Attribute “explained\_variance\_ratio” could view the ratio of the information amount of each new feature vector to the total information amount of the original data after dimension reduction.

**5. Split the data in 80% training and 20% testing, and perform a Decision Tree ML**

**calculation. Report the accuracy (you will get help during the lab of this task)**



Set random\_state equal to 29, and calculate the accuracy of the test data to be 95%. The accuracy obtained by setting different random\_state values is not unique. The decision tree is shown in the figure below.

