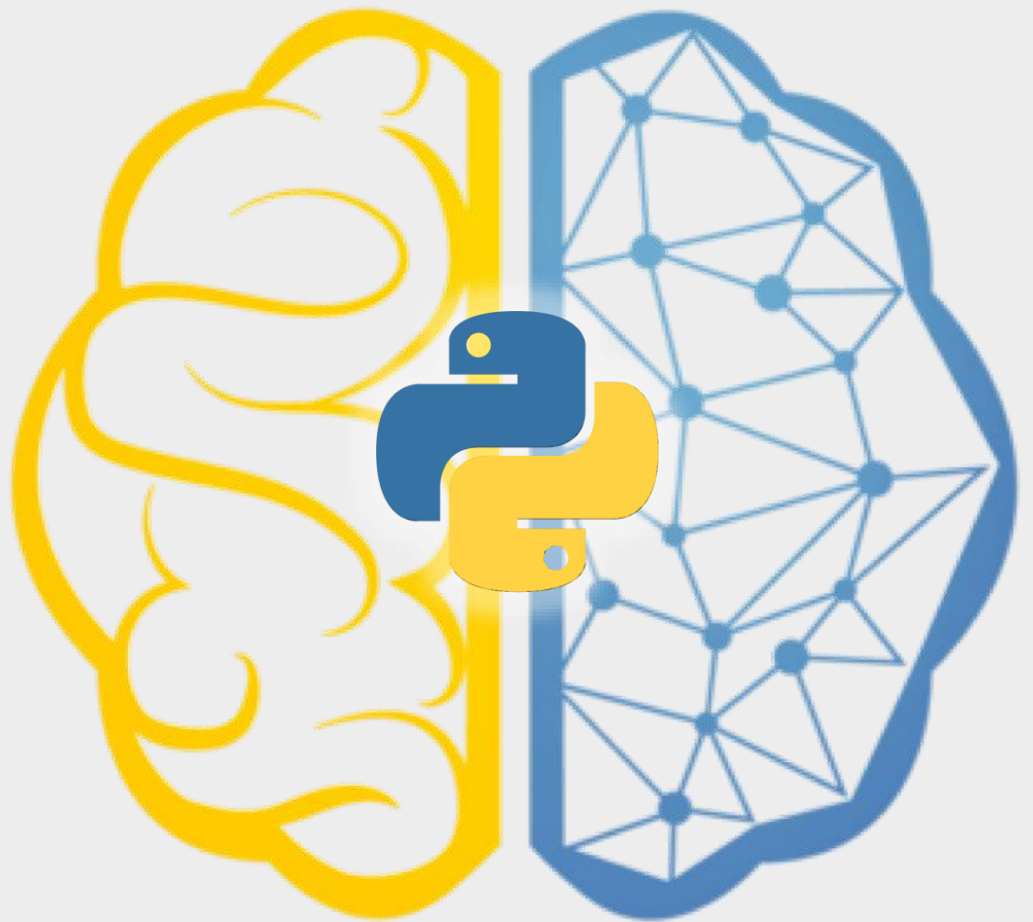


# FIND THE PRINCIPAL COMPONENTS OF IRIS DATASET



*-By Shivam Kumar Giri*

## PROBLEM STATEMENT

To examine the first 2 principal components of X. These components contain lots of information about our data set. Create a scatter plot with each of the 150 rows of X projected onto the first two principal components. In other words, the horizontal axis should be first principal component, the vertical axis should be second principal component.

Additionally, try to discriminant for each object using tree different colors for three different categories.

## ABSTRACT

The Iris dataset represents **3 kind of Iris flowers (Setosa, Versicolour and Virginica)** with 4 attributes: **sepal length, sepal width, petal length and petal width.**

**Principal Component Analysis (PCA)** applied to this data identifies the combination of attributes (principal components, or directions in the feature space) that account for the most variance in the data. Here we plot the different samples on the 2 first principal components.



**Iris Versicolor**



**Iris Setosa**



**Iris Virginica**




### Objectives of PCA:

- It is basically a non-dependent procedure in which it reduces attribute space from a large number of variables to a smaller number of factors.
- PCA is basically a dimension reduction process but there is no guarantee that the dimension is interpretable.
- Main task in this PCA is to select a subset of variables from a larger set, based on which original variables have the highest correlation with the principal amount.

### Uses of PCA:

- It is used to find inter-relation between variables in the data.
- It is used to interpret and visualize data.
- As number of variables are decreasing it makes further analysis simpler.
- It's often used to visualize genetic distance and relatedness between populations.

## LEGEND

SI No.	Symbol	Flower	Image
1.	0	Iris Versicolor	
2.	1	Iris Setosa	
3.	2	Iris Viriginica	

## METHODOLOGY

### Step1: Import all libraries

Basic library such as matplotlib, pandas, seaborn, NumPy and sklearn

For **data analysis**: NumPy, pandas

For **data visualization**: matplotlib, seaborn

For **classification/clustering**: sklearn

### Step 2: Import the dataset

Load the iris dataset: `iris=datasets.load_iris()`

### Step 3: Get the basic data exploration of iris dataset

Explore various features of dataset:

`iris.target_names`

```
array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

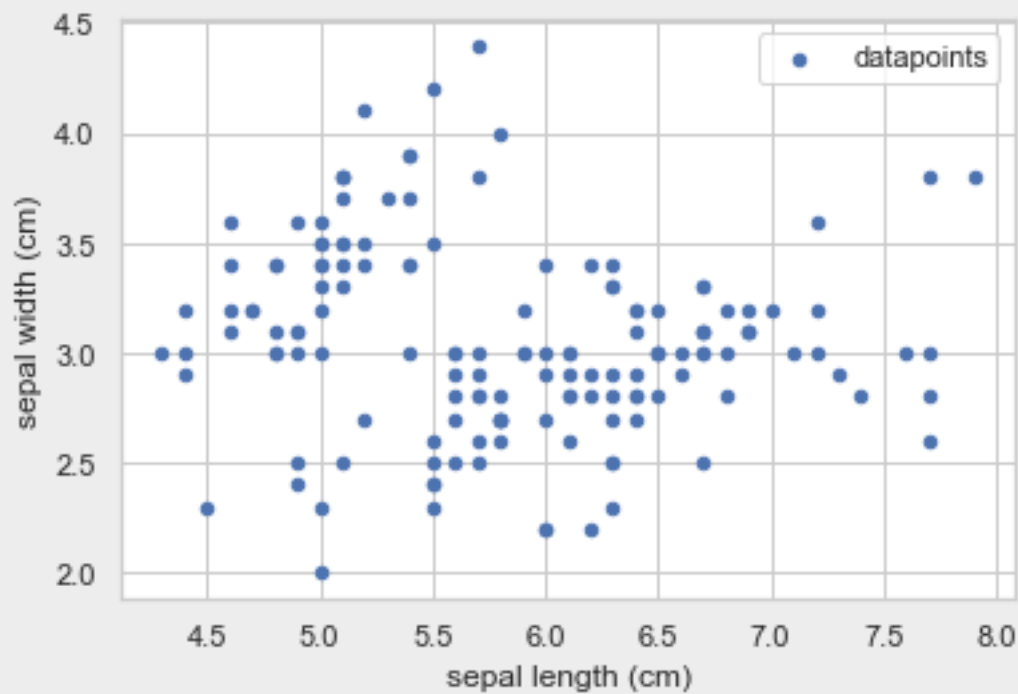
#### Step 4: Create Dataframe Perform Basic Data Exploration

Create a pandas dataframe and perform basic Exploration of data

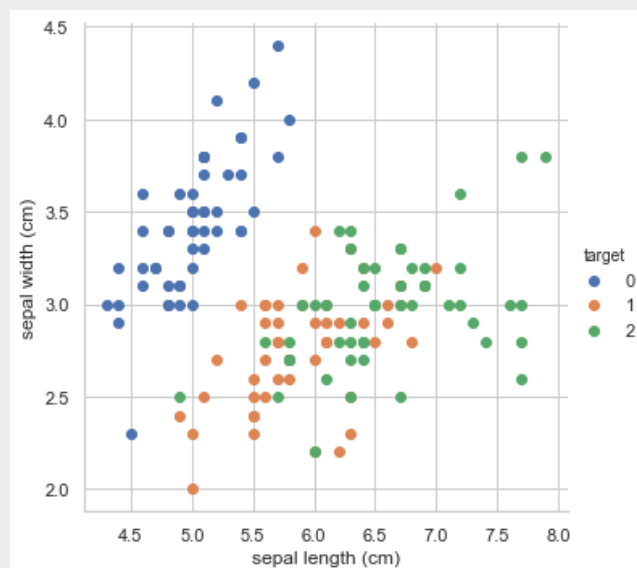
```
data.shape
```

```
(150, 5)
```

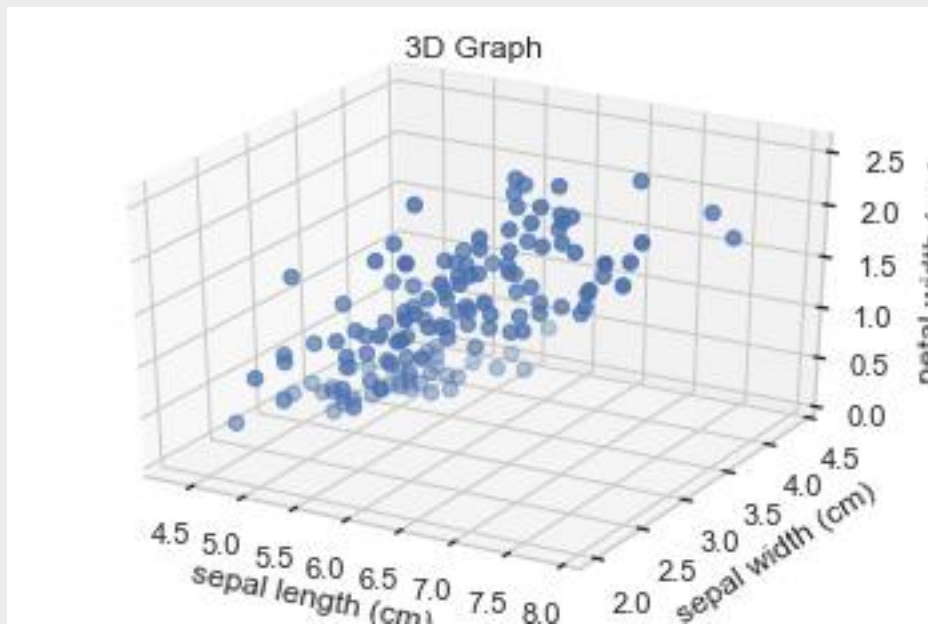
#### Step 5: Create 2D-Scatter Plot for representing 2 Dimensions



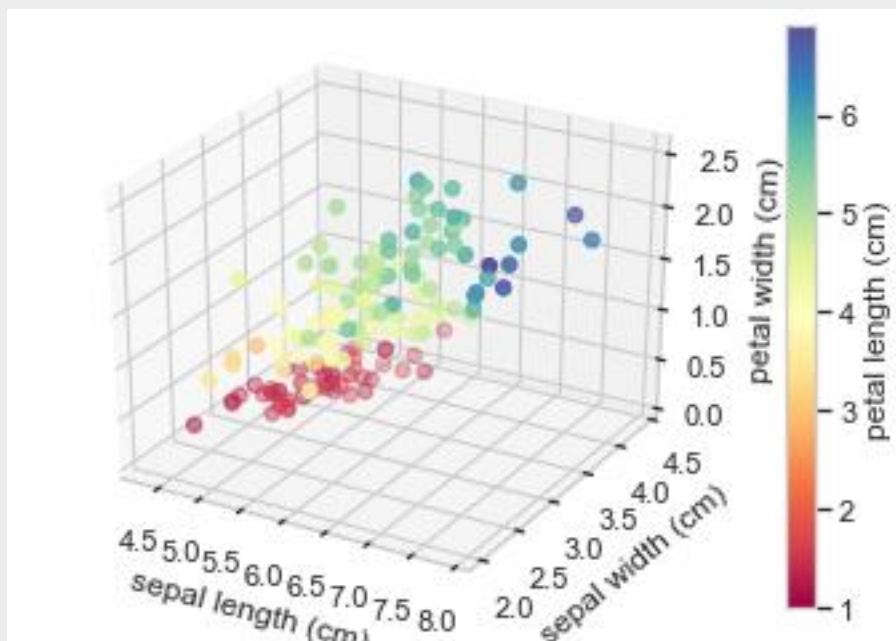
Stylize the Graph with distinct target



**Step 6: Create 3D-Scatter Plot for representing 3 Dimensions**



**Step 6: Create 4D-Scatter Plot for representing 4 Dimensions**

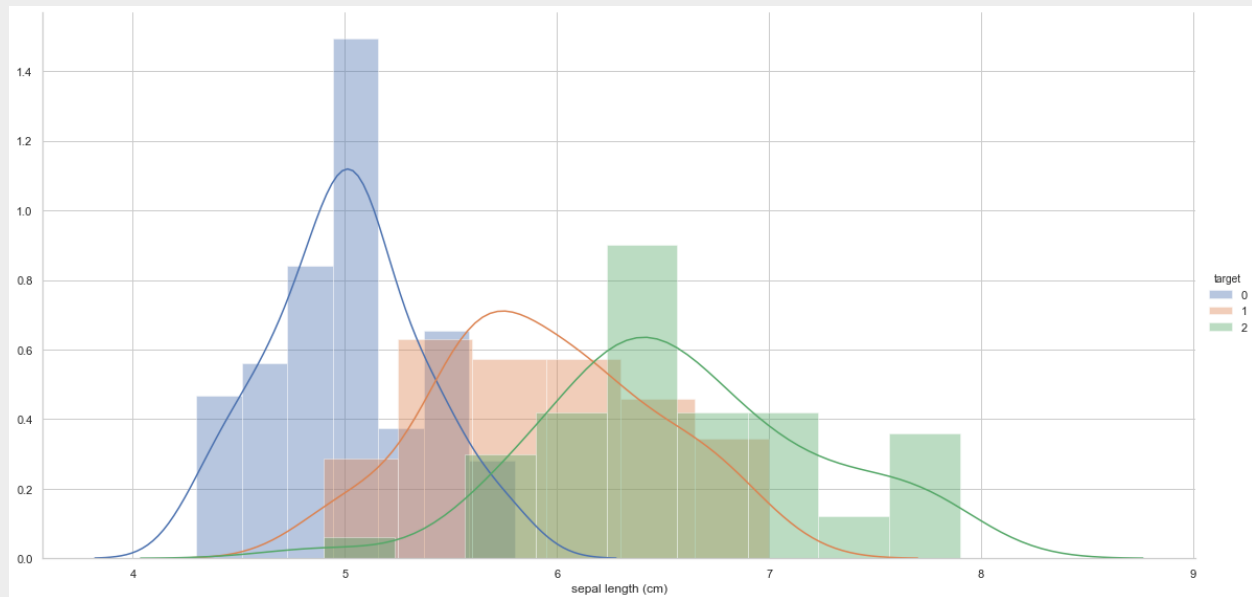


**Basic observation** is it's hard to analyze the flowers in 3rd and 4th Dimension of Data

**Step 7: Basic observation for all the dimensions of flower analysis**

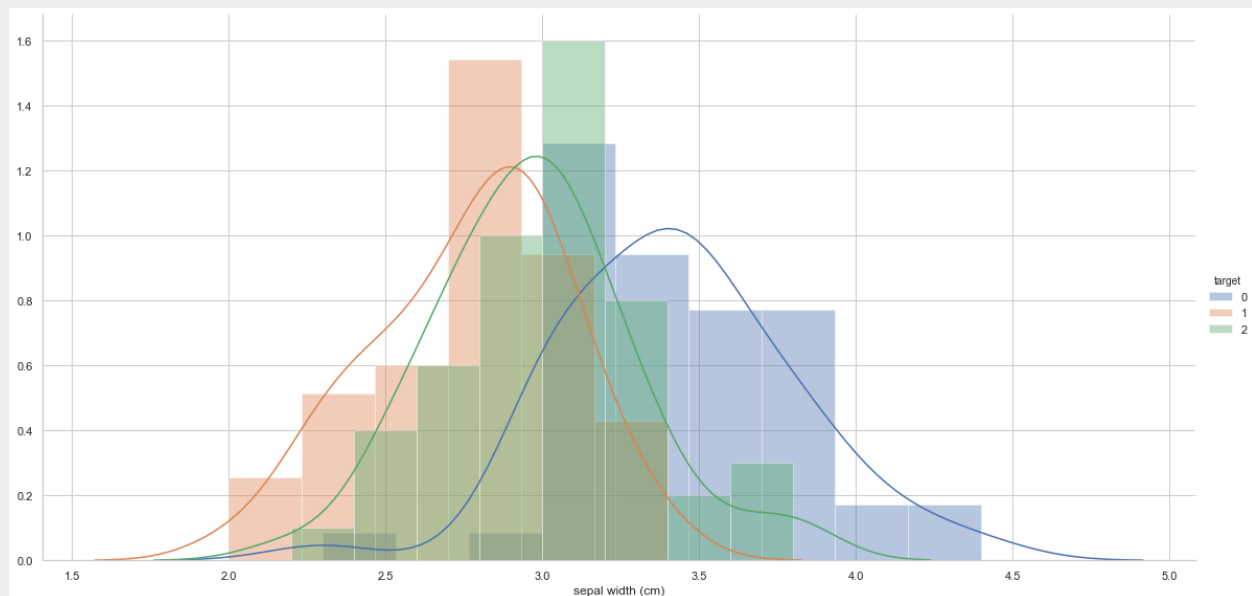
**Sepal length:**

It's easy to segregate Iris Versicolor from sepal length



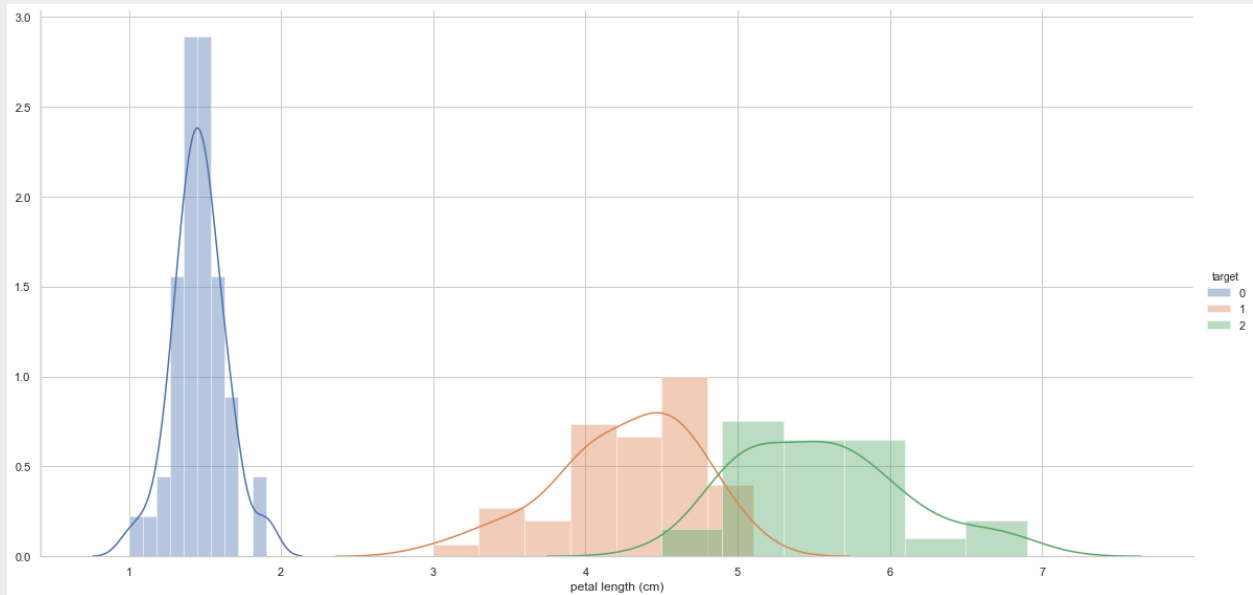
### Sepal width:

It's hard to segregate any species of Iris from sepal width



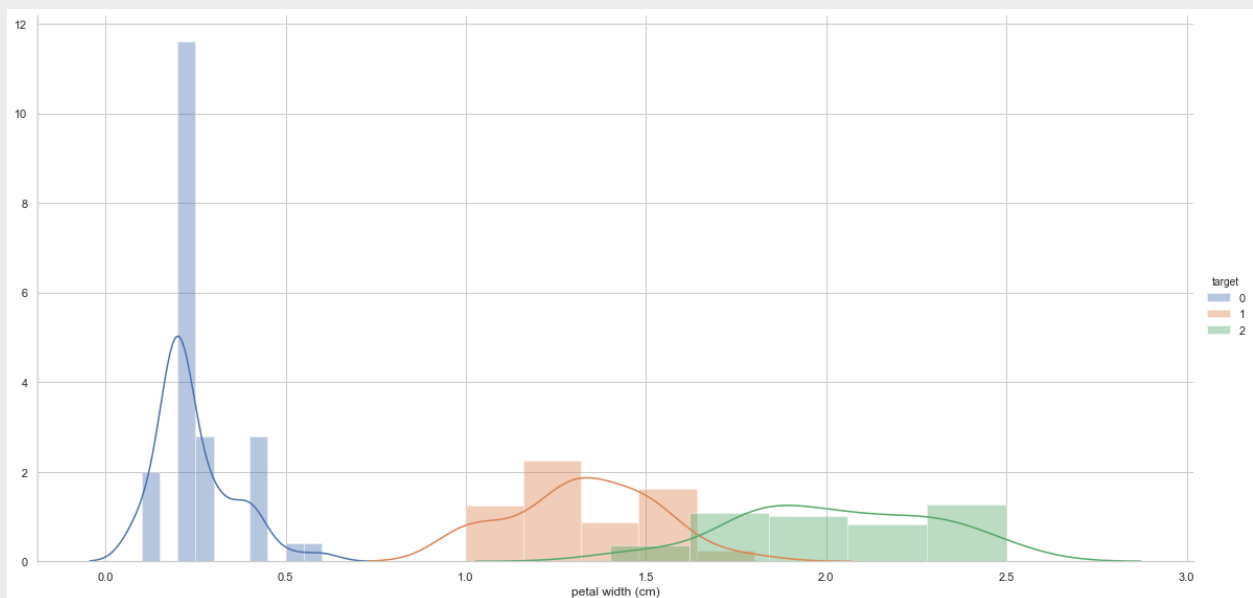
### Petal length:

It's very easy to segregate Iris Versicolor from petal length out of three



### Petal width:

It's very easy to segregate Iris Versicolor from petal width out of three



### Step 8: Create pair plots to visualize n-th dimensions

A pairs plot allows us to see both distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis and, fortunately, are easily implemented in Python. We can visualize the above graph along with pairwise plotting of data.





In the above case for 4D or 4 features we have 6 plots, consider 100D or 100 features, number of plots will be  $^{100}C_2$  plots.

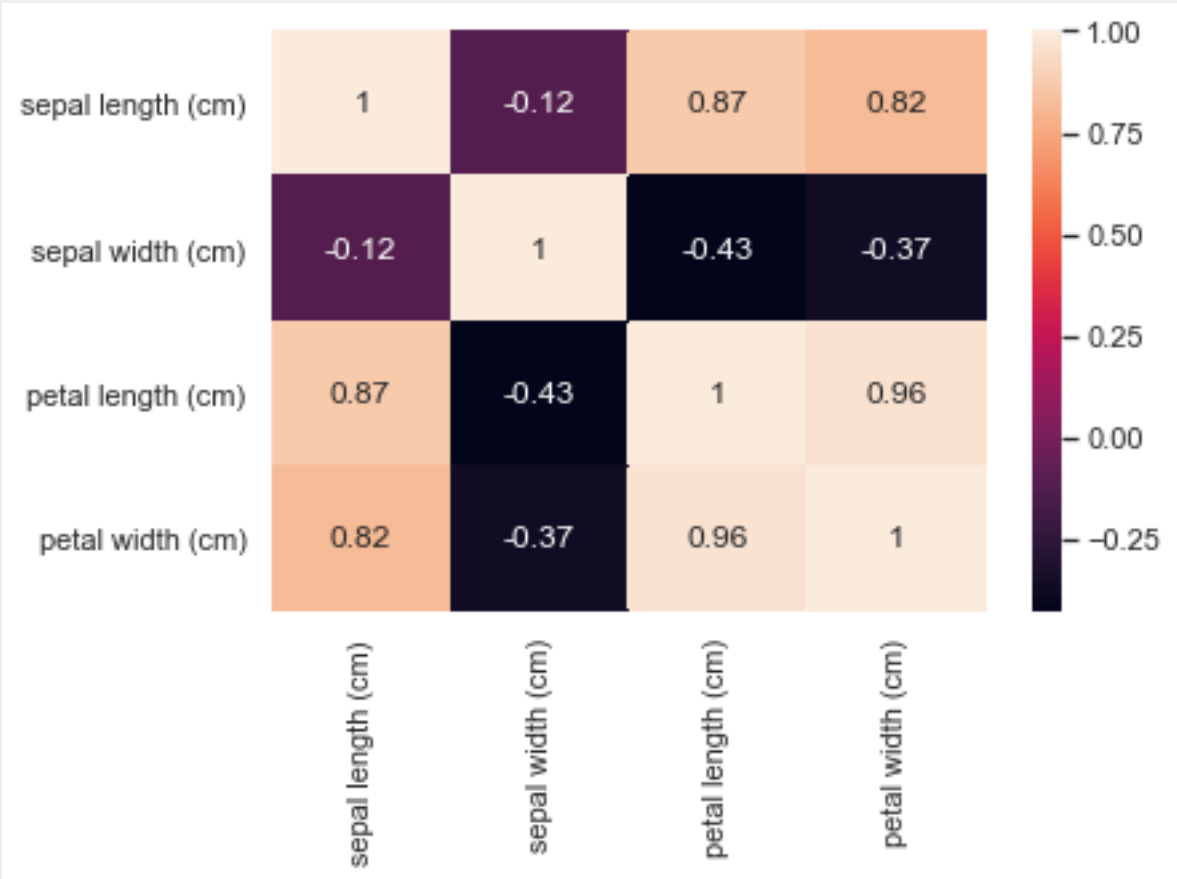
### **Step 9: Get the correlation of dimensions**

Correlation is interdependence of variable quantities.

Along with that there is graphical representation of correlation with the heat map.



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
sepal length (cm)	1.000000	-0.117570	0.871754	0.817941	0.782561
sepal width (cm)	-0.117570	1.000000	-0.428440	-0.366126	- 0.426658
petal length (cm)	0.871754	-0.428440	1.000000	0.962865	0.949035
petal width (cm)	0.817941	-0.366126	0.962865	1.000000	0.956547
target	0.782561	-0.426658	0.949035	0.956547	1.000000



### ***Step 11: Divide basic Data for PCA***

X=> dimensions variables

Y=> Target variables

### ***Step 12: Perform preprocessing part by fitting Standard Scale***

Doing the pre-processing part on training and testing set such as fitting the Standard scale.

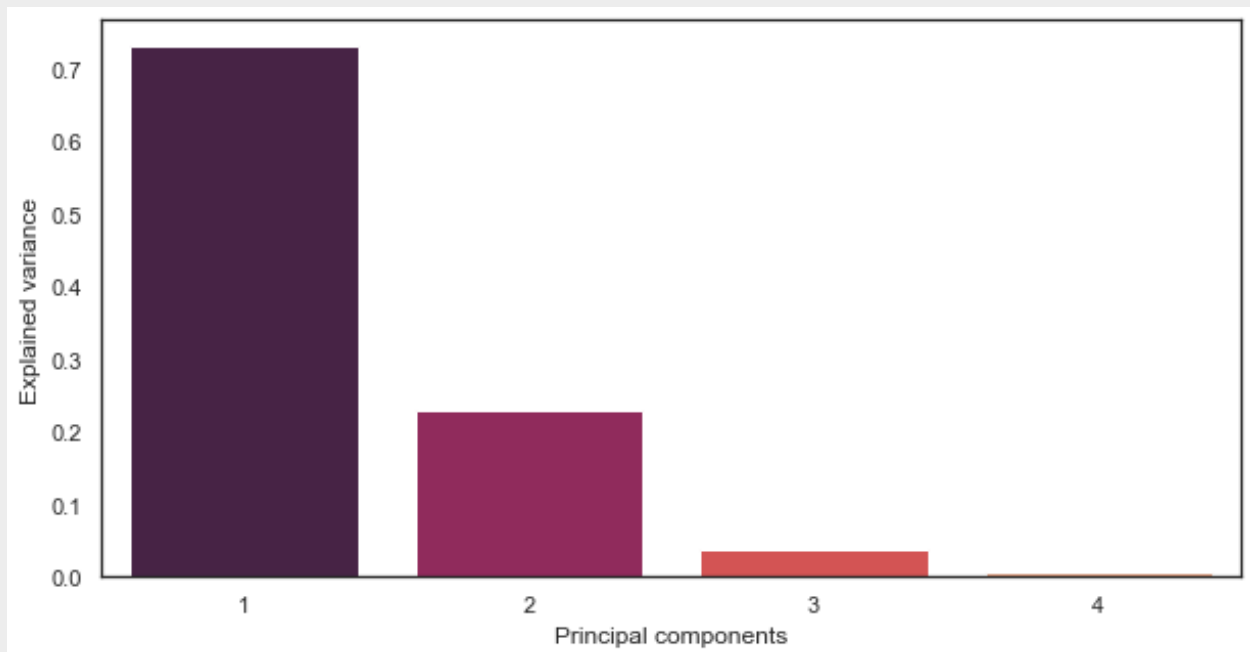
### ***Step 13: Perform Basic PCA and check the Variance***

Applying the PCA function into training and testing set for analysis.

The variance for each Principle components are:

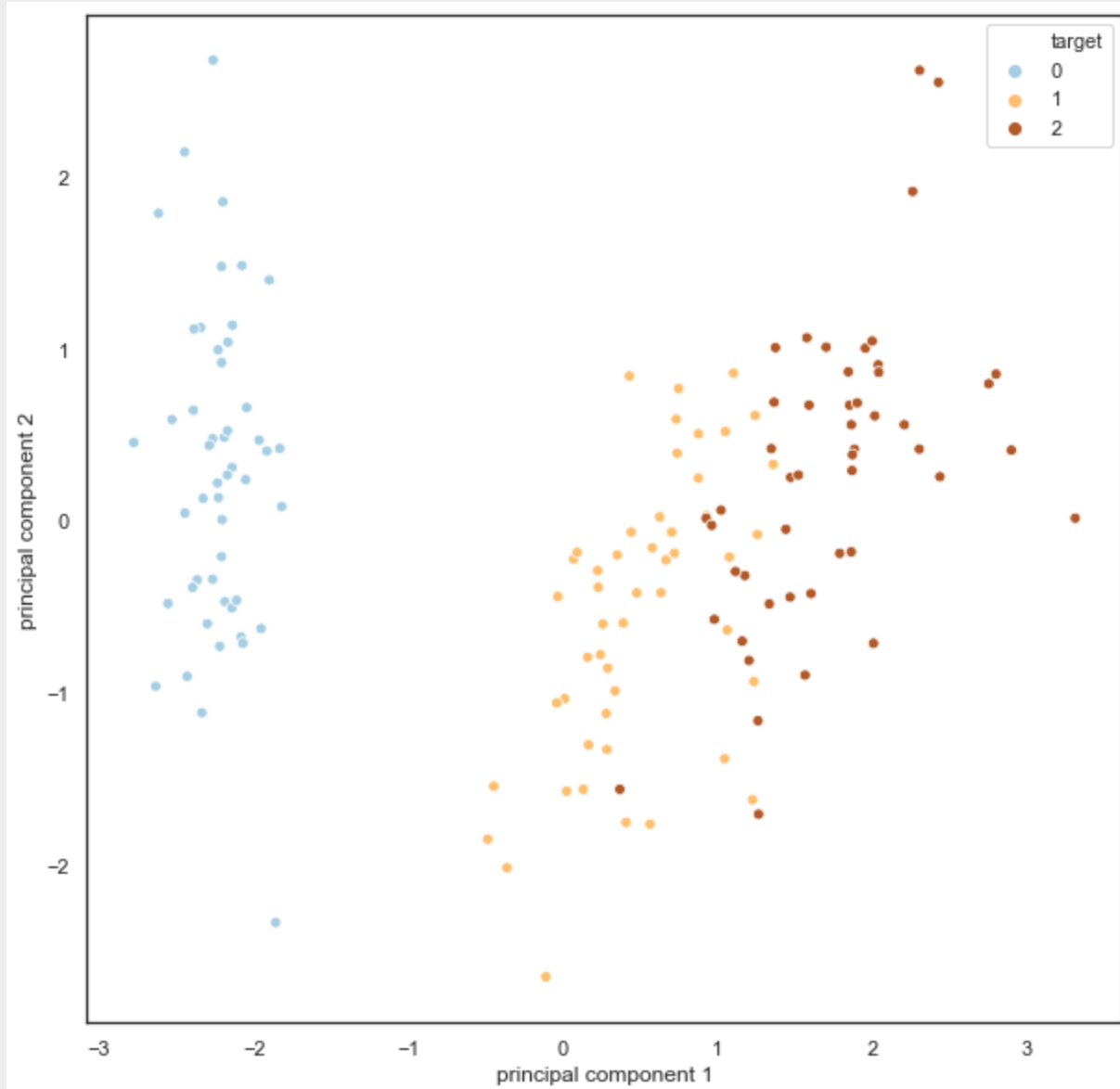
[0.72962445, 0.22850762, 0.03668922, 0.00517871]

### ***Step 14: Plot variance of each Principle components***

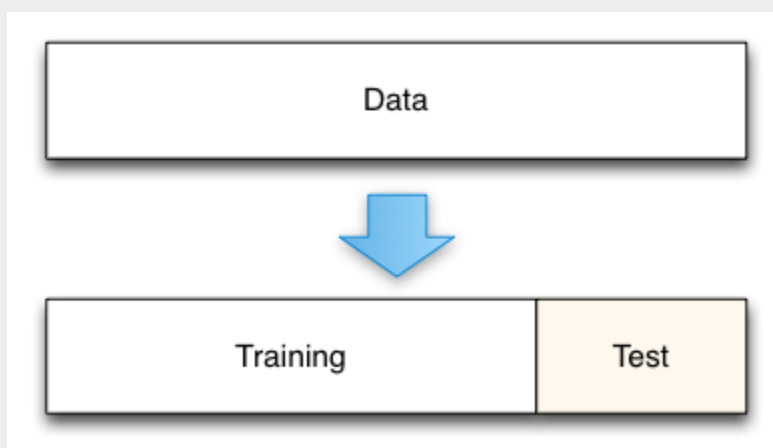


### ***Step 15: Interpreting the output of PCA***

Here is the basic Interpretation of PCA obtained



**Step 16: Splitting the dataset into the Training set and Test set**



**Step 17: Perform the basic Principle components Analysis**

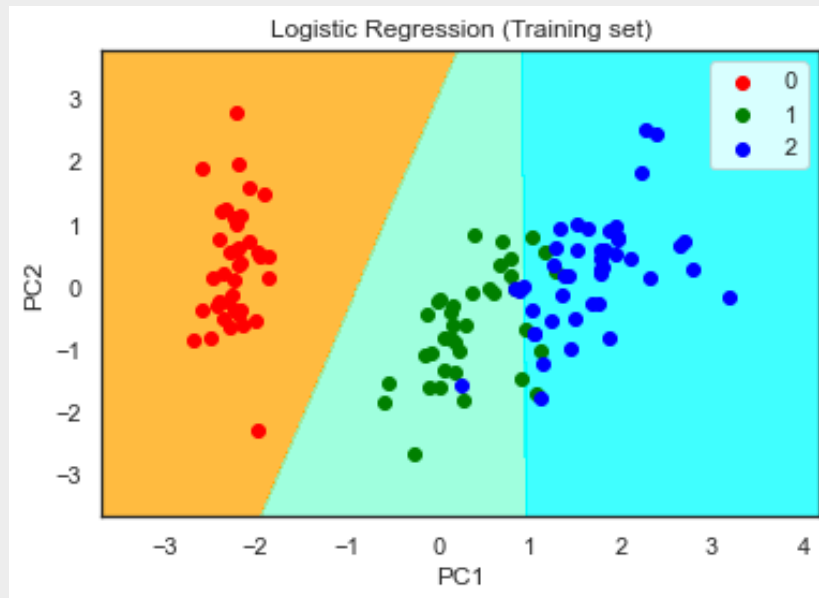
The PCA is performed for dimension =2.

[0.75079979 0.24920021]

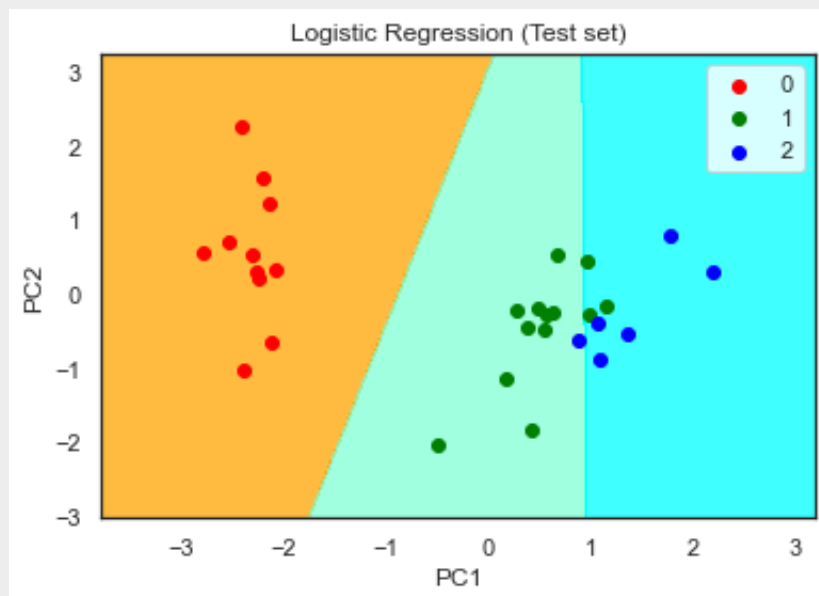
**Step 18: Fitting Logistic Regression To the training set**

Fitting Logistic Regression To the training set imported from sklearn.linear\_model

**Step 19: Visualizing the training set results through scatter plot**



**Step 20: Visualizing the test set results through scatter plot**

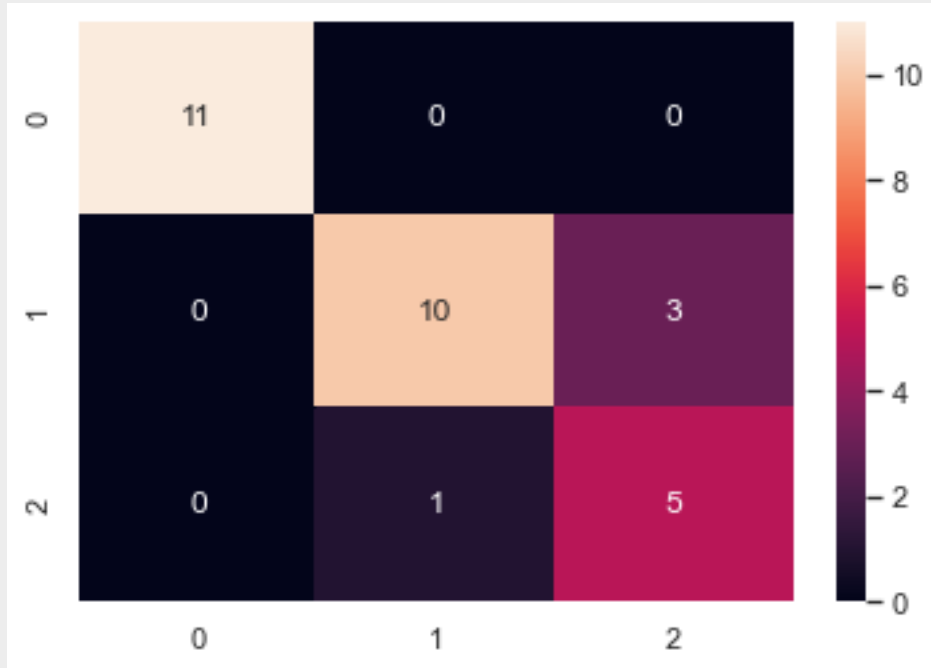


### Step 21: Creating and Visualizing the confusion

Confusion Matrix of the classifier with two major PCA on 3 species of flower are:

```
[[11  0  0]
 [ 0 10  3]
 [ 0  1  5]]
```

The heatmap generated is as follows:



### Step 22: Predict the accuracy of the Principle Components 1 and 2 for test cases

Accuracy Achieved is:86.66666666666667%

## CONCLUSION

Hence, we can find out that its **easy to classify Versicolor** from the other species of Iris. There may be little inaccuracy attained by PCA in **Setosa and Virginica** species of Iris. The Principle components hence is attained here for two dimension out of 4 dimensions here

# The End