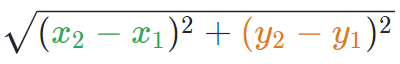
**DATA MINING ASSIGNMENT -2**

**KNN REPORT**

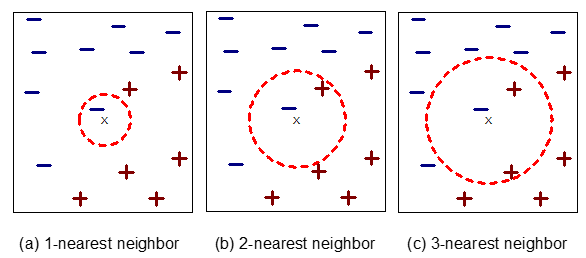
1. **Describe Nearest Neighbors method**

Nearest Neighbor is used to classify the class label using neighbors. One way is to find all the training data that are relatively similar to the attributes of the test data and classifying. When given a test example we compute the proximity to the rest of the data points in the training set. In this method, we first train a model using training set and use that model to classify the class of testing set.

KNN algorithm makes use of Euclidean distance

E. D. = 

where we find distance between training data and testing. The class labels with least distances are considered for classification. Suppose if k=3 then least 3 distances are to be considered and the class label which is more no. of times should be considered as final classifier. If there are 2 different classifiers then it is better to choose k value as odd num.



In above figures the center X mark is considered as testing sample. X have to classified either into – or +. In figure (a) k = 1 which actually a negative example because k=1 directly classifying X as – so we should preferably choose other than 1.

In figure (c) k = 3 so considering X as center of the circle and 3 nearest neighbors are two are – and one + so X will be classified as –

1. **Explain what was your criteria for selecting the two attributes**

If we see the features, we have sepal length, sepal width, petal length, petal width we

have to select features that belong to sepal or petal but not one from each. The feature, which gives high accuracy must be selected. When we plot other features such as petal length and petal width, we get less accuracy than sepal length and sepal width. If we

plot petal length and petal width, “Over Fitting“occurs. Sepal length and sepal width has least overfitting and high accuracy. Once it is done, sepal length and sepal width gives more accuracy than that of petal length and petal width.

1. **Visualizations of the classifier in a 2D projection, for all three different number of neighbors.**

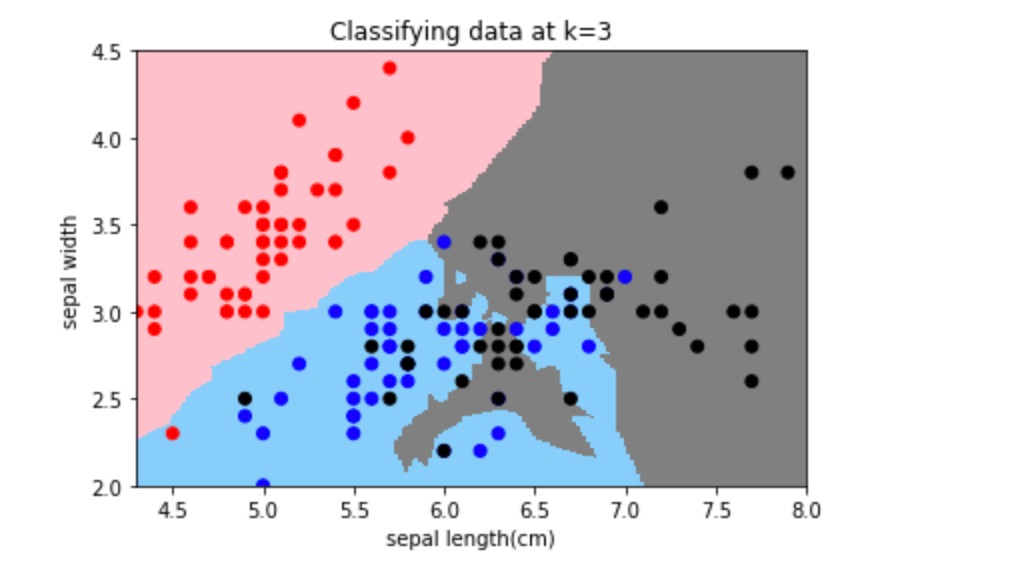
Here there are 3 different background colors each represents 3 different class labels, represented as 0, 1 and 2.

Pink background color is for “0”

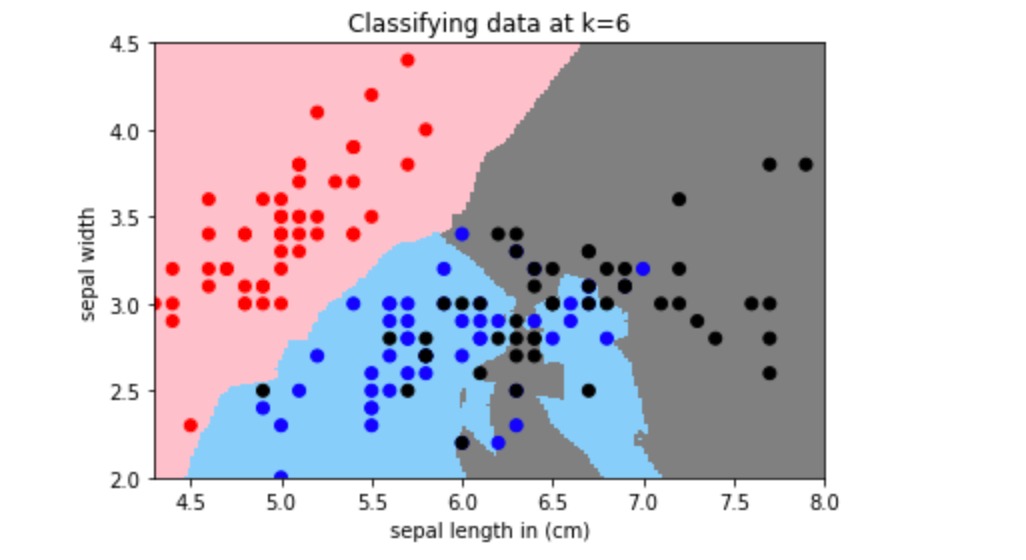
Blue background color is for “1”

Grey background color is for “2”.

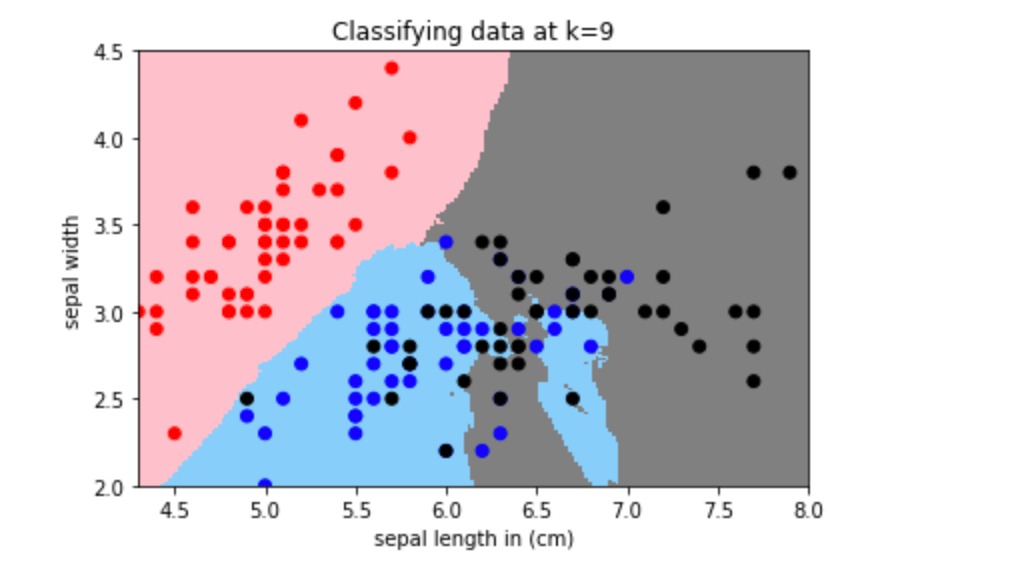
Similarly, 3 different colored dots also represents same class labels as mentioned above but these are of testing data class labels.



The above Figure has sepal length as x-axis and sepal width as y-axis and k value is 3.The plot is a scatter plot as mentioned in code.



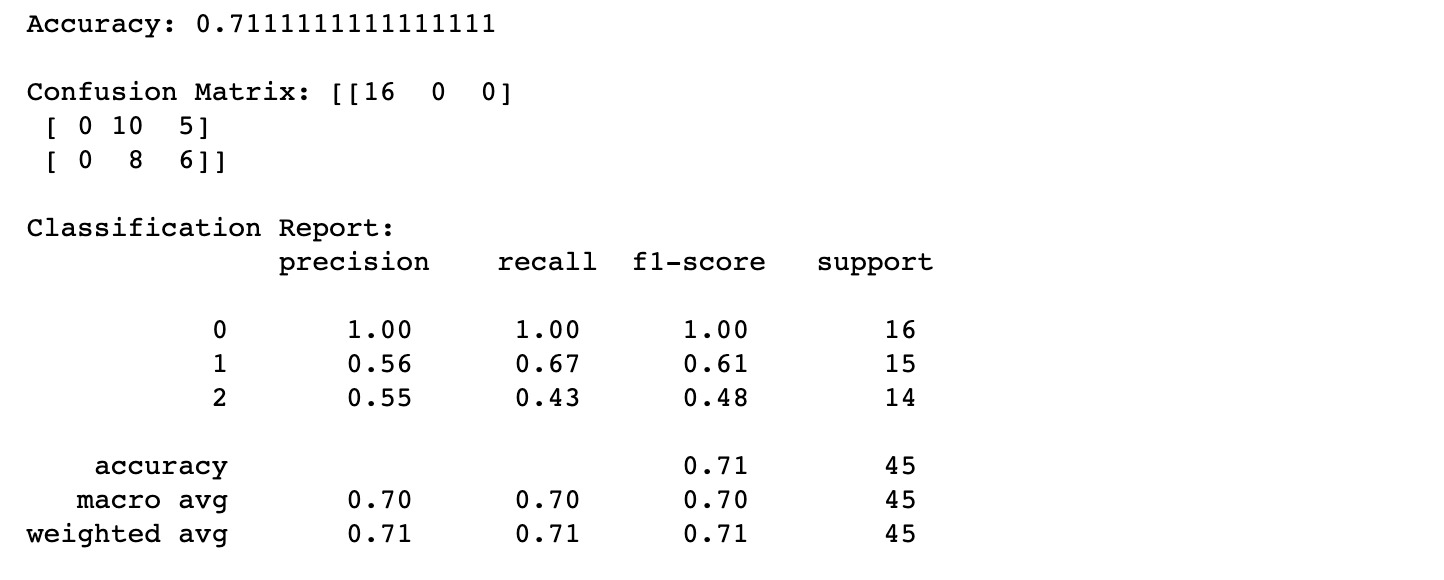
The above scatter plot and has sepal length as x-axis and sepal width as y-axis and k value = 6.



The above scatter plot and has sepal length as x-axis and sepal width as y-axis and k value is 9.

1. **Interpret and compare the results**

* For k=3:



Here, column represent predicted values and rows represent actual values, which

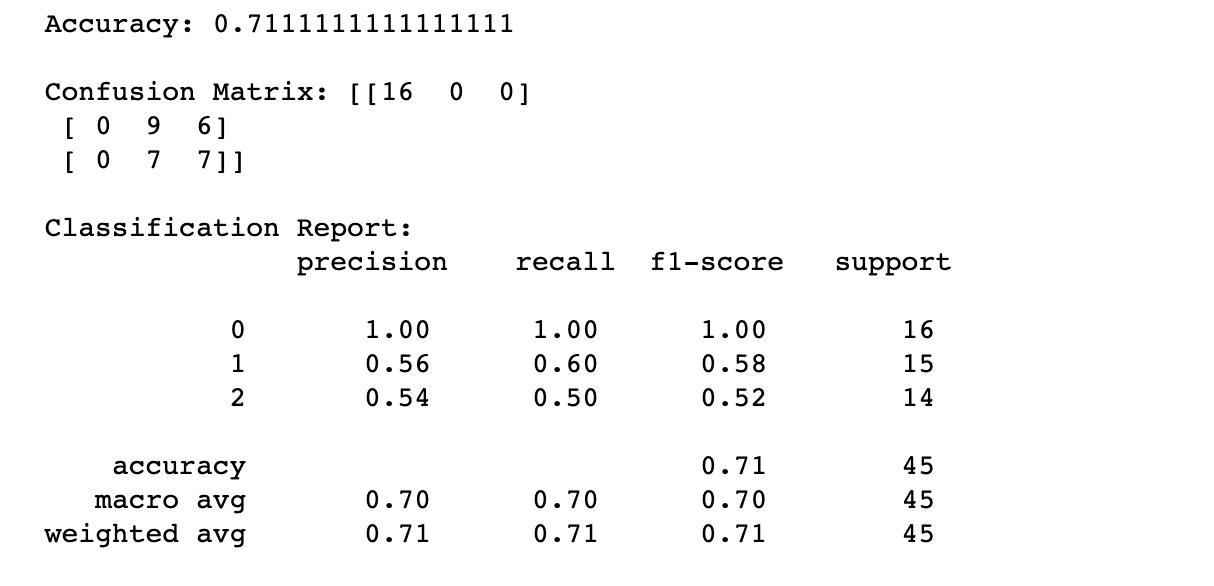
means in the 1st  row class 0 are of 16 values and there are zero wrongly predicted

classes. In the 2nd row class 1 are of total 15 values among which 10 are correctly

classified and 5 are wrongly classified as class 2. In the 3rd row there are 14 values in class 2 among which 6 are correctly classified as class 2 and 8 are wrongly classified as class 1.

Accuracy of knn when k=3 is 0.71111111111111111

* For k=6:



Here, column represent predicted values and rows represent actual values, which

means in the 1st  row class 0 are of 16 values and there are zero wrongly predicted

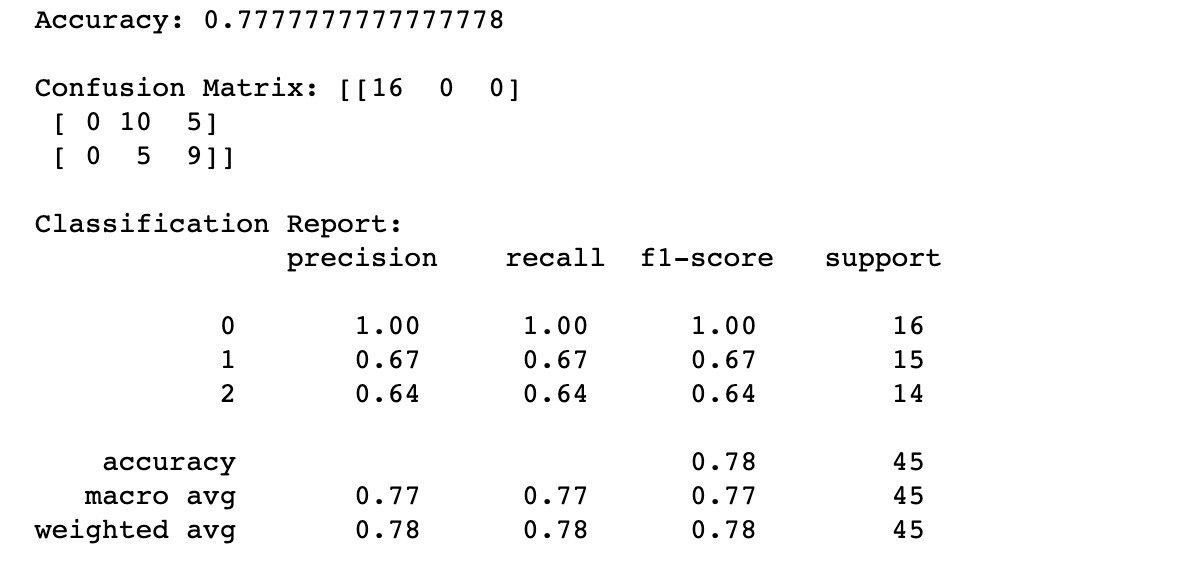
classes. In the 2nd row class 1 are of total 15 values among which 9 are correctly

classified as class 1 and 6 are wrongly classified as class 2. In the 3rd row there are 14 values in class 2 among which 7 are correctly classified as class 2 and 7 are wrongly

classified as class 1.

Accuracy of knn when k=6 is 0.71111111111111111

* For k=9:



Here, column represent predicted values and rows represent actual values, which

means in the 1st  row class 0 are of 16 values and there are zero wrongly predicted

classes. In the 2nd row class 1 are of total 15 values among which 10 are correctly

classified and 5 are wrongly classified as class 2. In the 3rd row there are 14 values in class 2 among which 9 are correctly classified as class 2 and 5 are wrongly classified as class 1.

Accuracy of knn when k=9 is 0.7777777777777778

When we observe the accuracies of three models k=9 has more accuracy of 0.7777 which

means in this case more number of predicted values are same as actual values than of models k=1 and k=2.