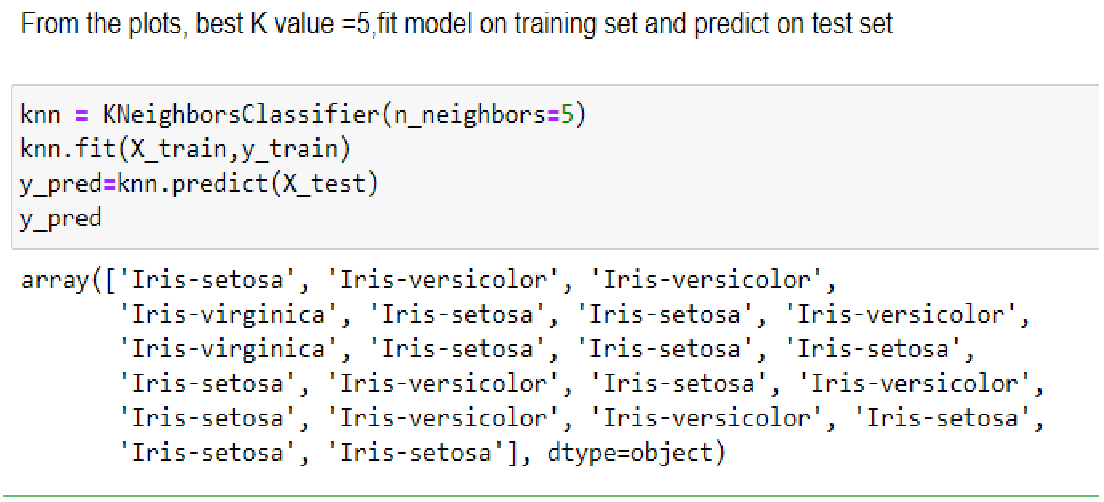
[CM6] Using the best found parameter, the model is fit on the entire training set and target is predicted on the test set.

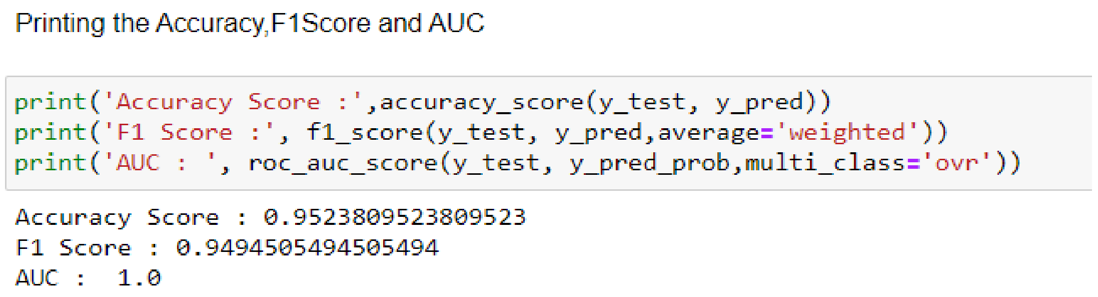
- **For iris,** on test set,



From the plots, best K value =5,fit model on training set and predict on test set

```
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train,y_train)
y_pred=knn.predict(X_test)
y_pred
```

```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-versicolor',
       'Iris-virginica', 'Iris-setosa', 'Iris-setosa', 'Iris-versicolor',
       'Iris-virginica', 'Iris-setosa', 'Iris-setosa', 'Iris-setosa',
       'Iris-setosa', 'Iris-versicolor', 'Iris-setosa', 'Iris-versicolor',
       'Iris-setosa', 'Iris-versicolor', 'Iris-versicolor', 'Iris-setosa',
       'Iris-setosa', 'Iris-setosa'], dtype=object)
```



Printing the Accuracy,F1Score and AUC

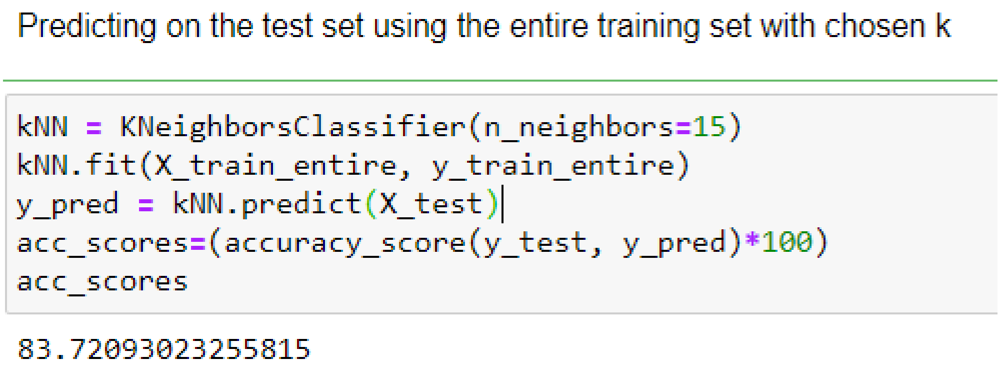
```
print('Accuracy Score :',accuracy_score(y_test, y_pred))
print('F1 Score :', f1_score(y_test, y_pred,average='weighted'))
print('AUC : ', roc_auc_score(y_test, y_pred_prob,multi_class='ovr'))
```

```
Accuracy Score : 0.9523809523809523
F1 Score : 0.9494505494505494
AUC :  1.0
```

The accuracy, f-score and AUC results obtained are tabulated below:

k	Accuracy	f-score	AUC
5	95.23%	0.94	1.0

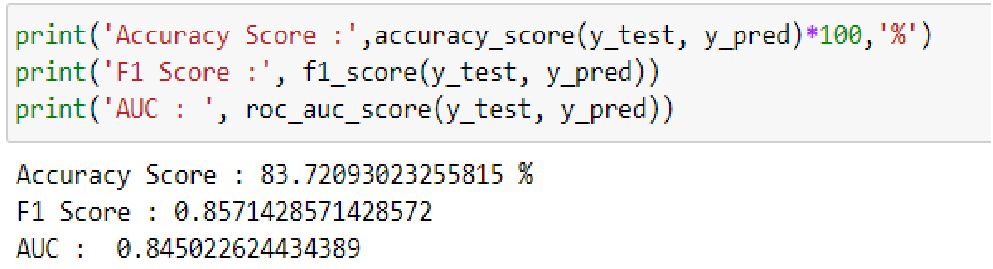
- **For Heart Disease**, on the test set,



Predicting on the test set using the entire training set with chosen k

```
kNN = KNeighborsClassifier(n_neighbors=15)
kNN.fit(X_train_entire, y_train_entire)
y_pred = kNN.predict(X_test)
acc_scores=(accuracy_score(y_test, y_pred)*100)
acc_scores
```

```
83.72093023255815
```



```
print('Accuracy Score :',accuracy_score(y_test, y_pred)*100,'%')
print('F1 Score :', f1_score(y_test, y_pred))
print('AUC : ', roc_auc_score(y_test, y_pred))
```

```
Accuracy Score : 83.72093023255815 %
F1 Score : 0.8571428571428572
AUC :  0.845022624434389
```

k	Accuracy	f-score	AUC
15	83.72 %	0.86	0.85

Lower k value causes higher variance and overfitting and can be highly susceptible to noise .Higher k causes higher bias and under fitting. So we should always try to balance the bias and variance while choosing the right k value.

For **iris dataset,**

As k value increased, lower k values [1, 5, 10] had higher accuracy and at higher k values i.e. k > 25, accuracy started decreasing. Therefore, accuracy was not affected the same way with increase of k.

Iris dataset is relatively small and hence with higher k values, can develop higher bias. Hence k=5 is chosen as the optimal k value.

For **heart disease dataset,**

Ask value is increased, at lower k values [1, 5, 10] ,high accuracy is exhibited and for k>15,the accuracy falls. Again, accuracy was not affected the same way with increase of k.

Heart disease dataset involves a large number of features including categorical, numeric, binary and ordinal data. Lower values of k can cause overfitting and higher values, underfitting. So as a balance, K =15 is selected as the optimal value which also happens to have a good accuracy value.