Assignment 7

**Question:**

1. Download the iris data
2. Apply K-nearest-neighbor algorithm in iris classification
3. Decide the best K value in this problem and give a brief explanation

**Answer:**

The K-nearest-neighbor algorithm (KNN) is one of the simple algorithms for regression analysis. KNN algorithm uses the given data and identifies the near datapoints which have same similarities.

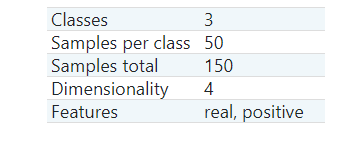
**Steps for using KNN algorithm:**

1. Determining the parameter K by analyzing the nearest data points.
2. Calculating the difference between all training samples and query instance.
3. Filter the nearest datapoints based on the minimum K radius.
4. Group the nearest neighbors in group x
5. Calculate the predictive attribute by using the nearest neighbor group.

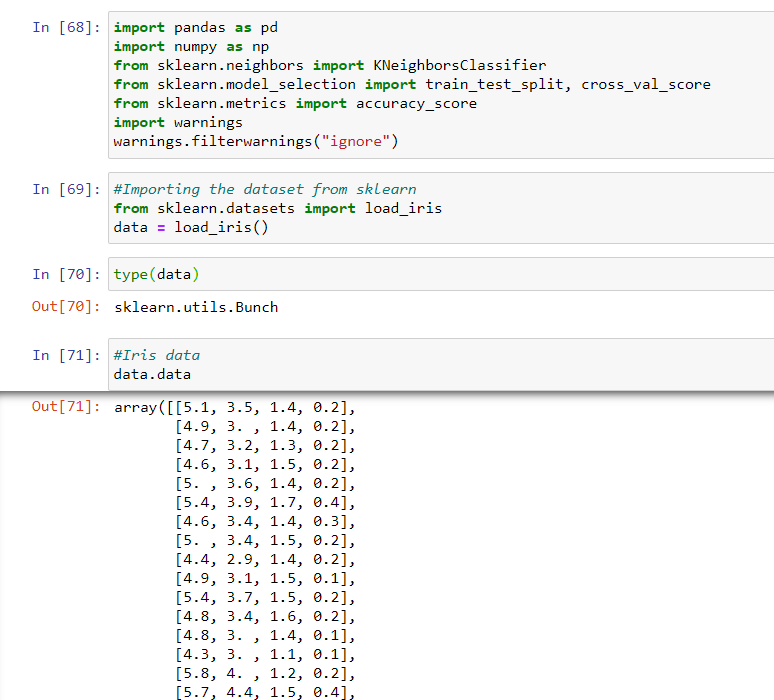
**KNN algorithm on iris dataset using scikit-learn library**

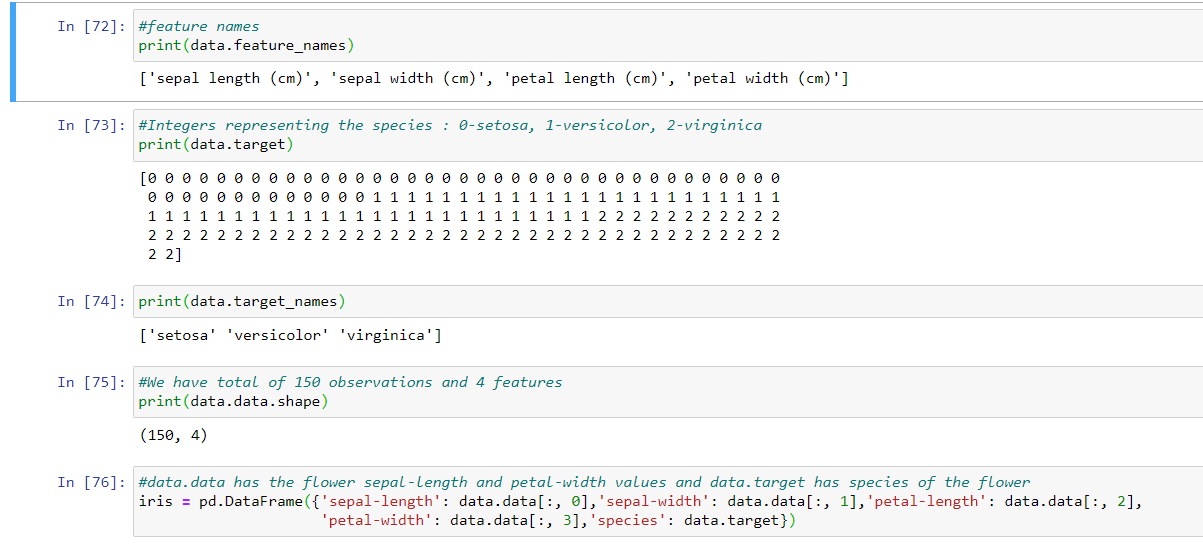
The iris dataset has total of 150 samples with 50 samples for each different species of iris flower. Each sample has sepal length, width and petal length and width and species name.

The dataset properties are shown in below figure.



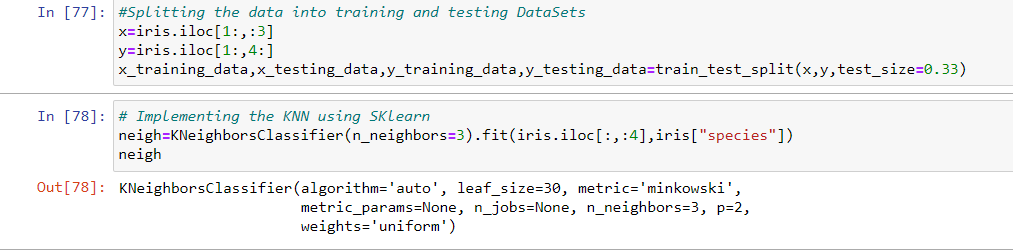
Code:





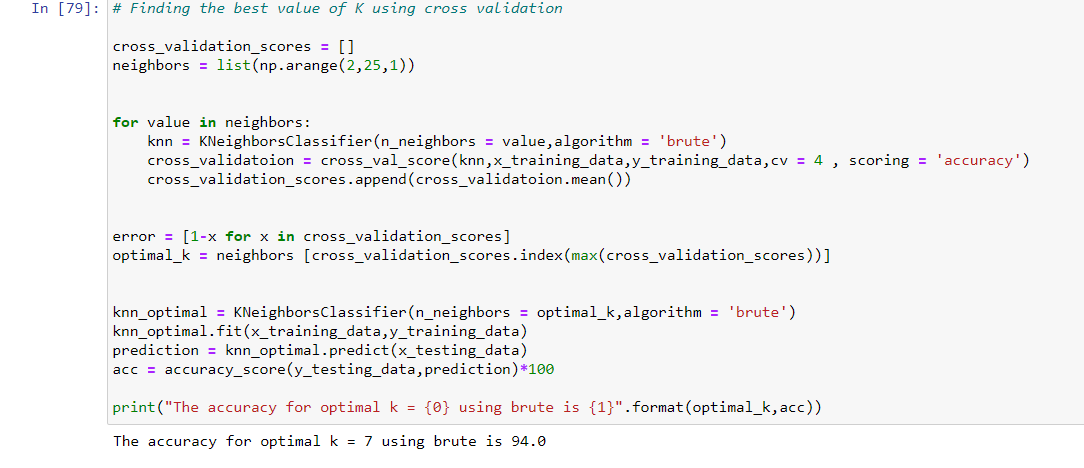
**Splitting the Data**

Training and testing on the same data is not an optimal approach, so we are splitting the data in to two parts one is training set and other is testing set. We are using train\_test\_split function to split the data and the parameter test size determines the split percentage. Here we are splitting the dataset and performing training and testing on different sets of data the resulting testing accuracy will be better estimate the performance of the model.



The KNN algorithm is implemented using the SKlearn. Import the class ‘KNeighborsClassifer’ from ‘neighbors’ module. ‘fit’ method is used to train the model on training data (X\_train, y\_train) and ‘predict’ method to do the testing on testing data (X\_test). Choosing the optimal value of K is critical, so we fit and test the model for different values for K using for loop and record the KNN’s testing accuracy.

**Cross validation**



**Problems faced**

* Deciding the value of K
* Calculating the optimal K value and accuracy for optimal K value

The accuracy optimal K value is 7 using the brute is 94.0

In KNN model, the complexity is defined by value of K. The high K value represents simpler model and low K value represents complex model.

The model accuracy for given model fo K value 2 to 25 is 94 percent.