**SD03Q07**

**FRUITS**

* ANISH ET

1832007

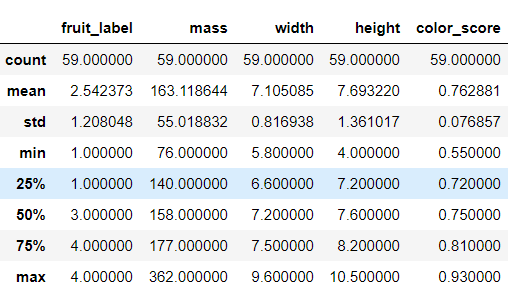
**DATA SET:**

The given Fruit classification dataset consists of 6 columns and 59 rows. “Fruit\_label, Fruit\_name, mass, width, height, color\_score” are the respective columns where “fruit\_name” column is string and consists the name of the fruits and the rest have the details (in numbers) of the respective sample of the fruits.

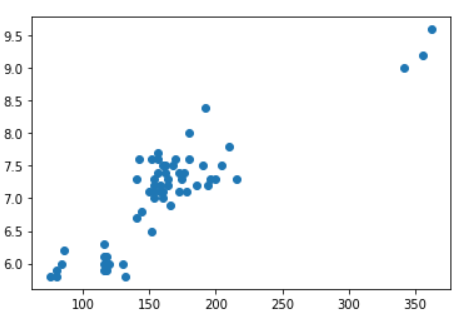
**DATA PREPROCESSING:**

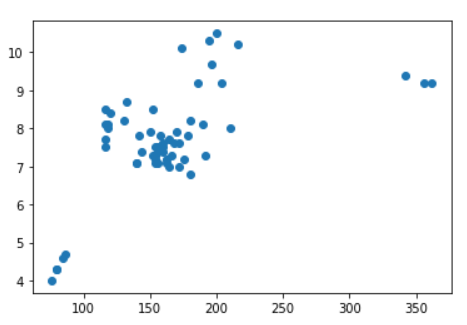
Data processing is, generally, "the collection and manipulation of items of data to produce meaningful information." In this sense it can be considered a subset of information processing, "the change of information in any manner detectable by an observer.The objective is to find the best combination of parameters to build the classifier. So, Pre-processing of the data has been done by plotting scatter plots for each and every combinations of the given parameters. Among those, The combination between “width” and “color\_score” has been found to be the best. The feature variable has been taken as “width” and “Color\_score” dataframes which has been already framed into arrays using numpy and pandas. Since there is no missing value, there is no need of filling the places. And also “Fruit\_label” has been taken as predictor variable. Using Sklearn.metrics, Scaler has been implemented into the dataset to scale the dataset further more.

**DATASET SNIPPET:**

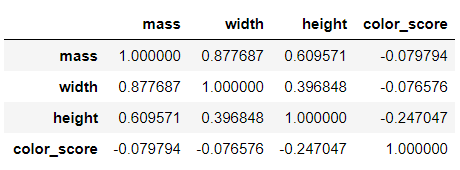
****

**SCATTER PLOTS:**

****



**CORRELATION TABLE:**



**KNN**

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in [Machine Learning for regression](https://quantra.quantinsti.com/course/trading-with-machine-learning-regression) and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function).

Using Sklearn\_model\_selection library, the implementation of splitting of train and test data set is done. The test and train sets are scaled using scaler library.

To find the neighbours, Euclidean Distance is found out for the variables. And the splitted datasets such as train and test sets are fitted into KNN classifier model. After finding out the best K value for the model, it was also fitted into model for predicting the outcome using parameters(feature variable and predictor variable). The model is created and accuracy of 0.916 out of 1 is obtained using sklearn.metrics library - accuracy\_score function

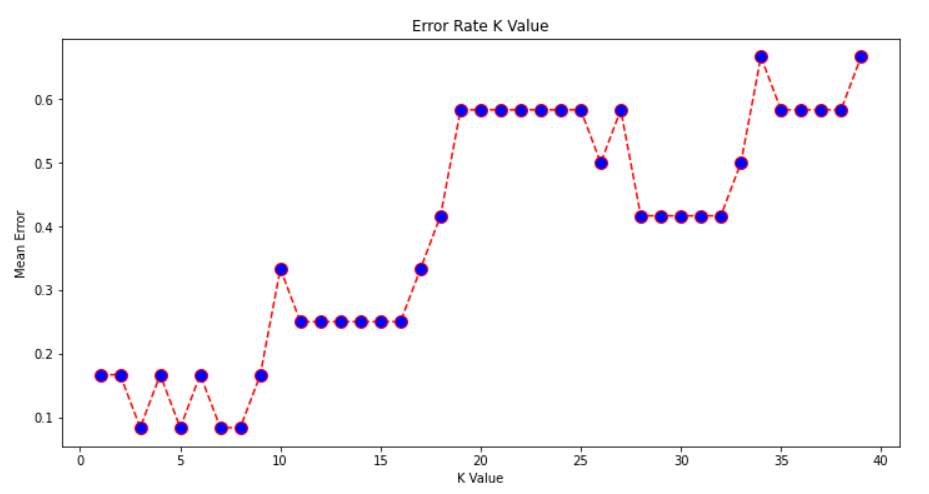


Also it was asked to perform atleast 3 test cases. It was programmed to perform a test case in every instances of a test set is re-iterated in the predict function. So it’s also done.

So it is clear that all the performed test cases were positive and the fruits were classified according to the parameters given in the arguments.

So the data distribution of the data set (i.e.) the splitting of train dataset and test dataset of the corresponding dataset leads to the accuracy of 91.6% in KNN Classifier. The fruits can be classified and predicted using the testcases.

The objective here is to find the best k value with Highesh R\_score. So the K- values with less mean errors were found using “for loop” with range of 1 and 40 to determine the number of errors in K Value.



**K value in x-axis and mean error in y-axis**

***GitHub Link for Code:*** https://github.com/anish0007/Curneu-Assessment/blob/main/fruitsdataset.ipynb