SD03Q07

K - Nearest Neighbour (KNN) algorithm

PROBLEM STATEMENT:

A dataset labelled based on fruit height, width, mass and colour score is given in fruits.xlsx. A classifier based on k Nearest Neighbour (KNN) algorithm is to be crafted for classification.

DATA SET:

	fruit_label	fruit_name	mass	width	height	color_score
0	1	apple	192	8.4	7.3	0.55
1	1	apple	180	8.0	6.8	0.59
2	1	apple	176	7.4	7.2	0.60
3	2	mandarin	86	6.2	4.7	0.80
4	2	mandarin	84	6.0	4.6	0.79
5	2	mandarin	80	5.8	4.3	0.77
6	2	mandarin	80	5.9	4.3	0.81
7	2	mandarin	76	5.8	4.0	0.81
8	1	apple	178	7.1	7.8	0.92
9	1	apple	172	7.4	7.0	0.89
10	1	apple	166	6.9	7.3	0.93
11	1	apple	172	7.1	7.6	0.92
12	1	apple	154	7.0	7.1	0.88

The given dataset consists of 6 columns and 59 rows. “Fruit_label, Fruit_name, mass, width, height, color_score” are the respective columns.

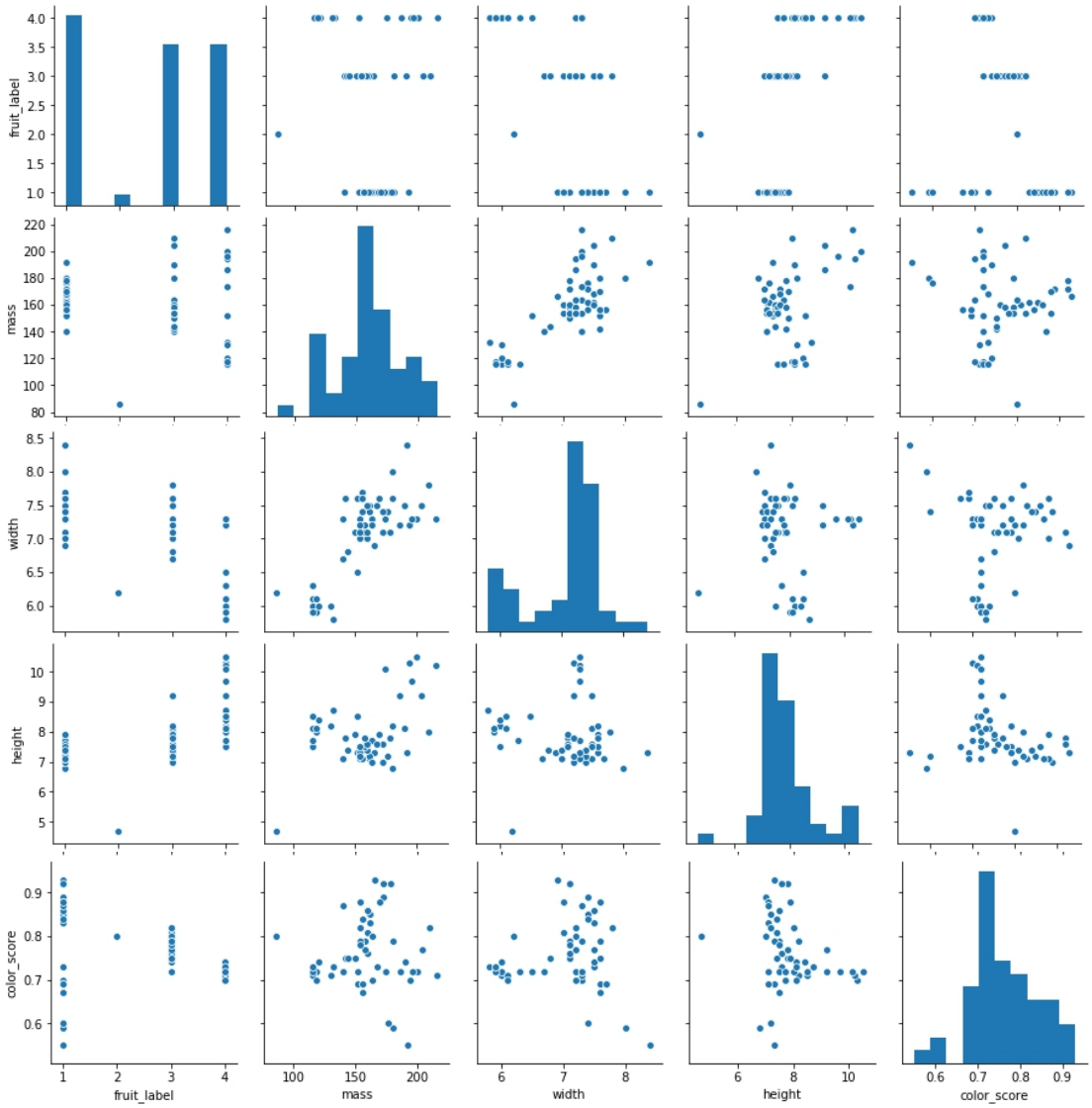
EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis is a process of investigating the data to find patterns and anomalies with the help of summary statistics and graphical representations.

	fruit_label	mass	width	height	color_score
count	59.000000	59.000000	59.000000	59.000000	59.000000
mean	2.542373	163.118644	7.105085	7.693220	0.762881
std	1.208048	55.018832	0.816938	1.361017	0.076857
min	1.000000	76.000000	5.800000	4.000000	0.550000
25%	1.000000	140.000000	6.600000	7.200000	0.720000
50%	3.000000	158.000000	7.200000	7.600000	0.750000
75%	4.000000	177.000000	7.500000	8.200000	0.810000
max	4.000000	362.000000	9.600000	10.500000	0.930000

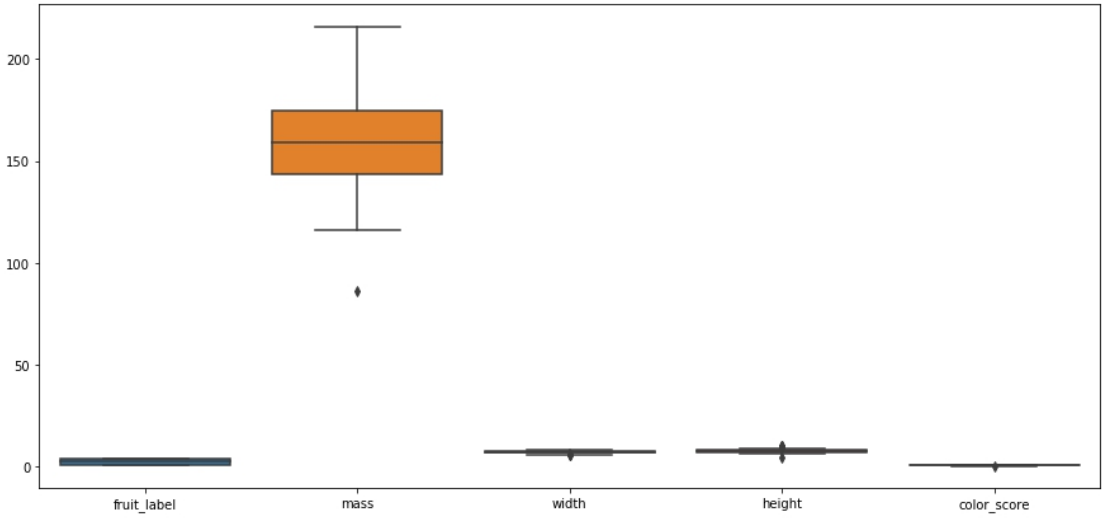
The summary statistics gives central tendency and other statistics etc.

PAIRPLOTS:



HANDLING THE OUTLIERS:

An outlier is an observation that lies an abnormal distance from other values which affect the statistics like mean, variance, etc.



Treating outliers is good for further analysis. Outliers are removed using IQR.The inter quartile range (IQR) is a measure of variability.

DATA PREPROCESSING:

Feature Scaling:

The data is scaled using Standard Scaler to avoid features with widely different range.

Data Splitting:

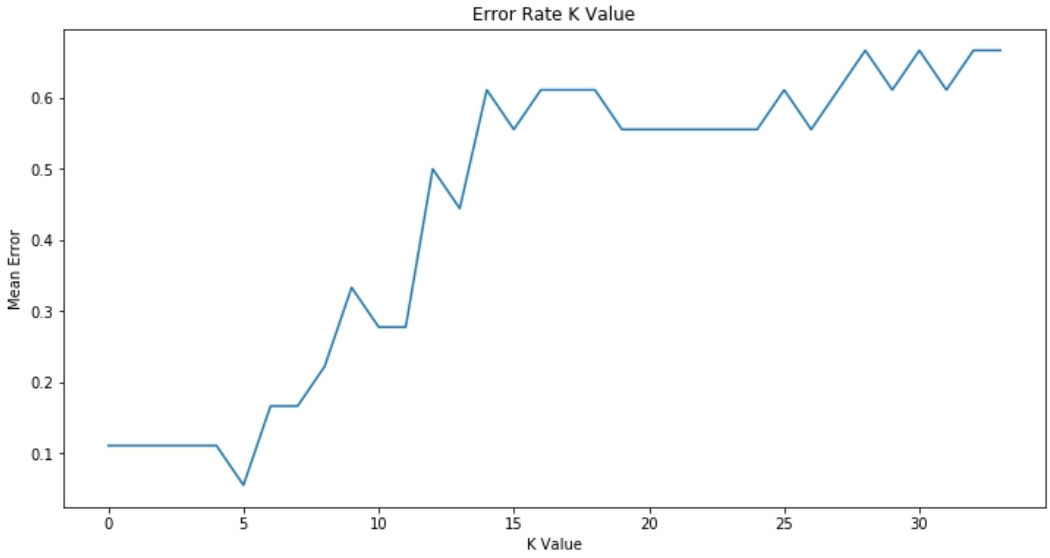
The data is split into training data and test data for model fitting. The data is split in the ratio 70:30.

MODEL FITTING AND EVALUATION:

After pre-processing, the data is classified using classification algorithms. This classification algorithm tries to draw some conclusions from the input values given for training. It will predict the class labels/categories for the new data.

```
Accuracy :   88.88888888888889
```

```
Text(0, 0.5, 'Mean Error')
```



The Fruit have been scaled using a Standard scaler to avoid features with widely different range. Since we did not considered the categorical variable it is not necessary to encode the data. The data is split in the ratio 70:30.

Further this data is being fitted with K Nearest Neighbour classifier models and their Accuracy 88.8%. This shown KNN model show better accuracy and it has classified each fruit correctly.