# Curneu MedTech Innovations Private Limited

**SD03Q03-1**

Problem Statement 1: - Mandatory

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.

• Generate scatter plots for various combination of parameters and do the feature engineering meaning thereby which parameters of best suited to build the classifier.
• Split the data into test and training split.
• Building a classifier using KNN from scratch.
• Figure out the best value of k with highest r_score.
• Run at least three test cases on the parameter and assess the fruit using the classifier.
• Only use python

## Dataset:

The given dataset consists of 6 columns and 59 rows. “Fruit_label, Fruit_name, mass, width, height, color_score” are columns where “fruit_name” column consists the name of the fruits and the rest have the details (in numbers) of the fruits.

| | 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 |

## Pre-processing:

The data Pre-processing has been done by plotting scatter plots for each and every combinations of the parameters. In those scatter plots the combination between “width” and “color_score” has been the best. “width” and “Color_score” had been taken as feature variables.“Fruit_label” has been taken as predictor variable. The dataset has been scaled by Scalar using Sklearn.metrics.

## Columns of data frame:

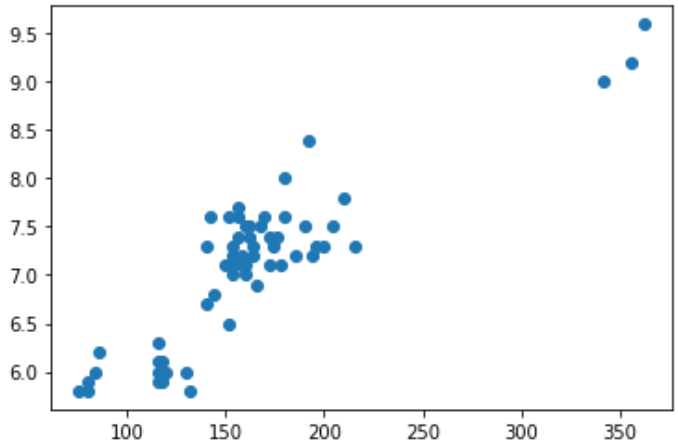
```
Index(['fruit_label', 'fruit_name', 'mass', 'width', 'height', 'color_s
core'], dtype='object')
```

## Describe dataframe:

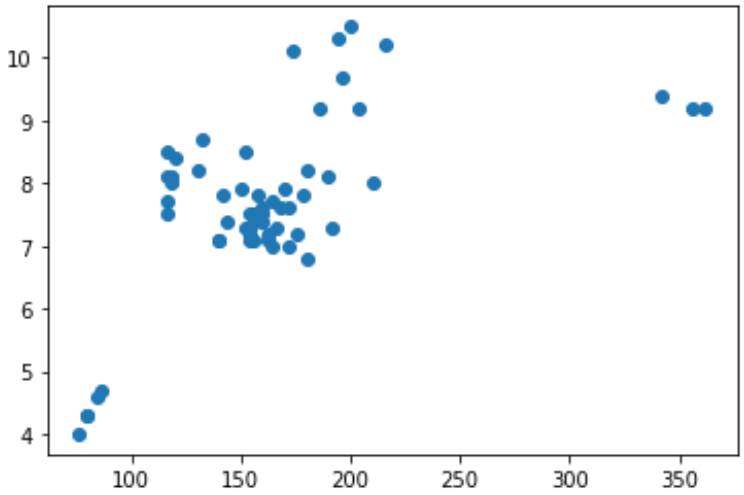
| | fruit_label | mass | width | height | color_score |
|---|---|---|---|---|---|
| **count** | 59.000000 | 59.000000 | 59.000000 | 59.000000 | 59.000000 |
| **mean** | 2.542373 | 163.118644 | 7.105085 | 7.693220 | 0.762881 |
| **std** | 1.208048 | 55.018832 | 0.816938 | 1.361017 | 0.076857 |
| **min** | 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 |

## Plots for various combination of parameters:

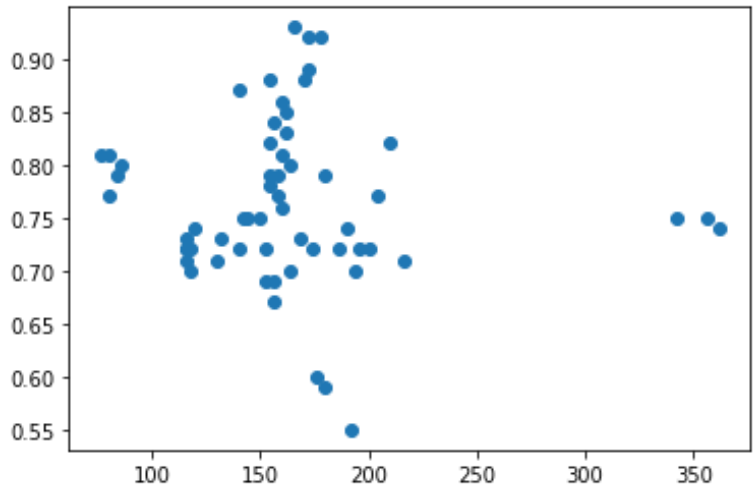
### Mass,Width:




### Mass,Height:



**Mass,color_score:**




**Correlation table:**

| | mass | width | height | color_score |
|---|---|---|---|---|
| **mass** | 1.000000 | 0.877687 | 0.609571 | -0.079794 |
| **width** | 0.877687 | 1.000000 | 0.396848 | -0.076576 |
| **height** | 0.609571 | 0.396848 | 1.000000 | -0.247047 |
| **color_score** | -0.079794 | -0.076576 | -0.247047 | 1.000000 |

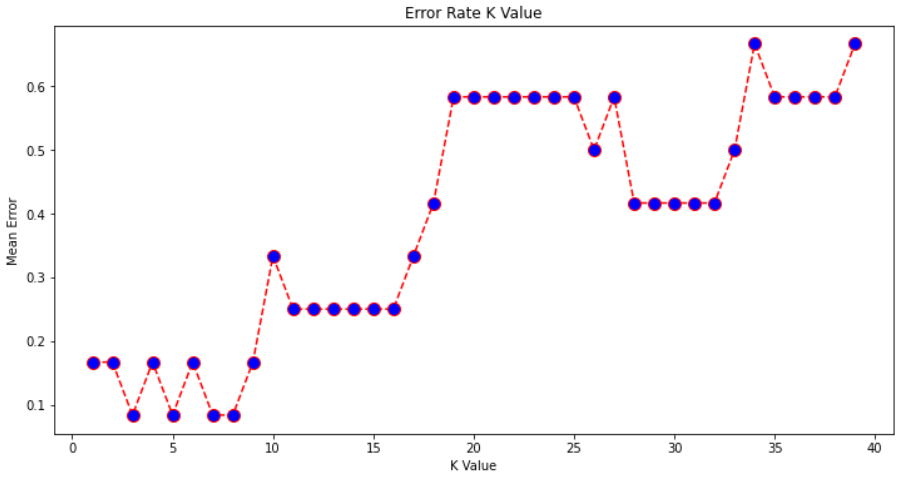
**KNN from scratch:**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique
The data is split in into train and test data and are scaled by using Scalar library.
The KNN model is fitted to train and test datasets.
The model is created with an accuracy of 0.833

```
Accuracy: 0.8333333333333334
```

The best value for k with highest R score has been figured out by finding the K-values with less mean errors.

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




Using the predict function the test cases have been performed on the train and test dataset.

```
[4, 1, 1, 4, 1, 3, 1, 3, 2, 4, 4, 1]
```

**Code link:**

https://github.com/shreetheerthaen/Curneu-Assessment/blob/main/fruits/Fruits_047.ipynb