**Curneu MedTech Innovations Private Limited**

**Shobana.S (1832046) , MSc Data Science, Department of computing , Coimbatore institute of technology.**

**Problem statement 1:**

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.

**INTRODUCTION:**

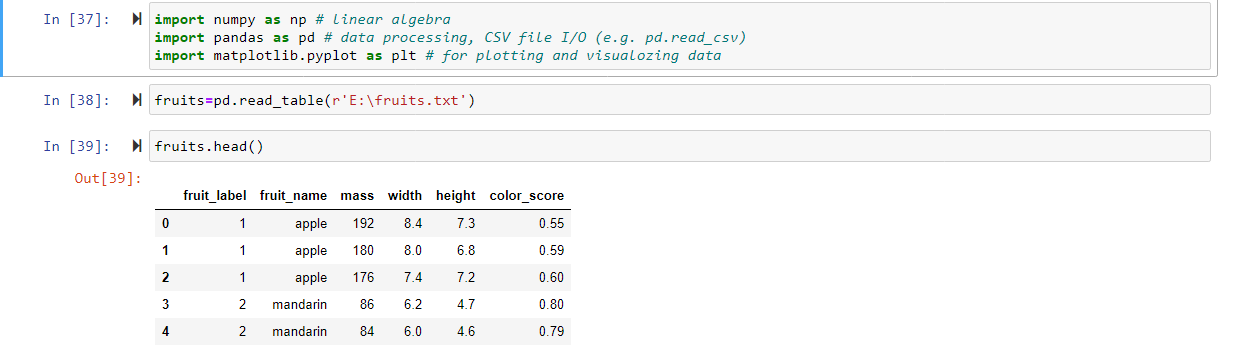
KNN falls in the **supervised learning** family of algorithms. Informally, this means that we are given a labelled dataset consisting of training observations (x,y)(x,y) and would like to capture the relationship between xx and yy. More formally, our goal is to learn a function h:X→Yh:X→Y so that given an unseen observation xx, h(x)h(x) can confidently predict the corresponding output yy.

The KNN classifier is also a **non parametric** and **instance-based** learning algorithm.

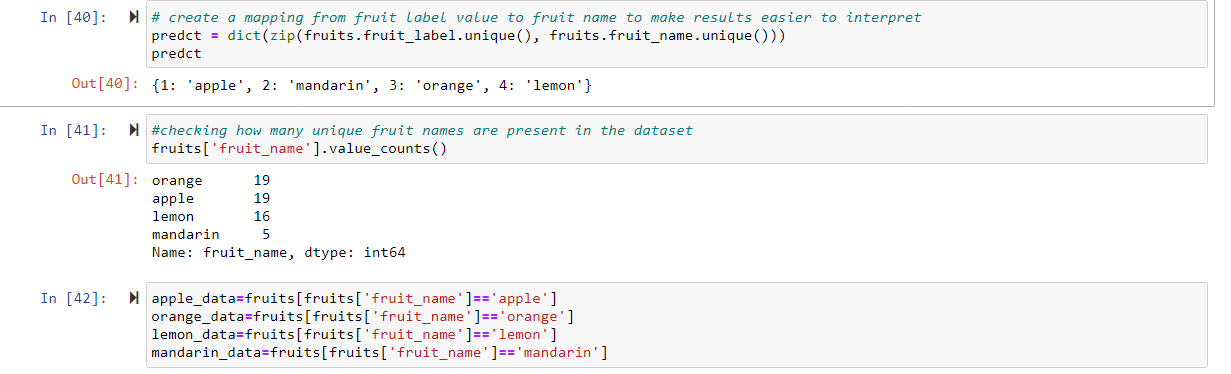
* **Non-parametric** means it makes no explicit assumptions about the functional form of h, avoiding the dangers of mismodeling the underlying distribution of the data. For example, suppose our data is highly non-Gaussian but the learning model we choose assumes a Gaussian form. In that case, our algorithm would make extremely poor predictions.
* **Instance-based** learning means that our algorithm doesn’t explicitly learn a model. Instead, it chooses to memorize the training instances which are subsequently used as “knowledge” for the prediction phase. Concretely, this means that only when a query to our database is made (i.e. when we ask it to predict a label given an input), will the algorithm use the training instances to spit out an answer.

**KNN ALGORITHM:**

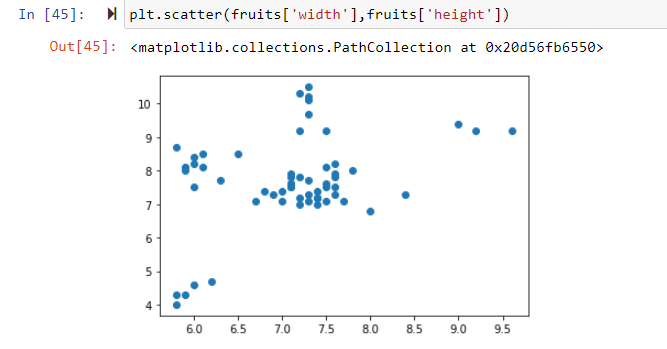
* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* Importing the libraries and dataset

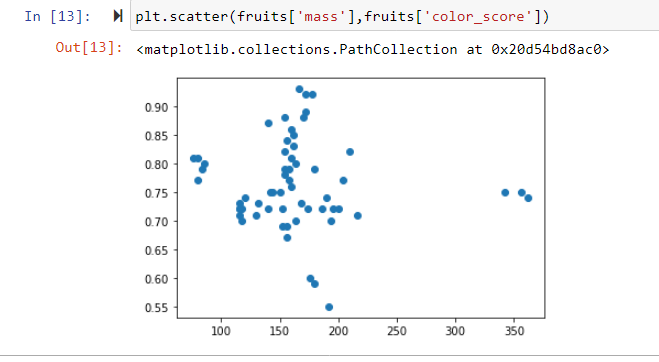


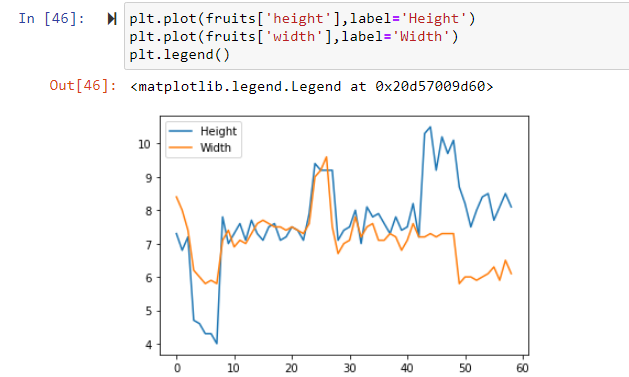
* Creating a mapping between attributes



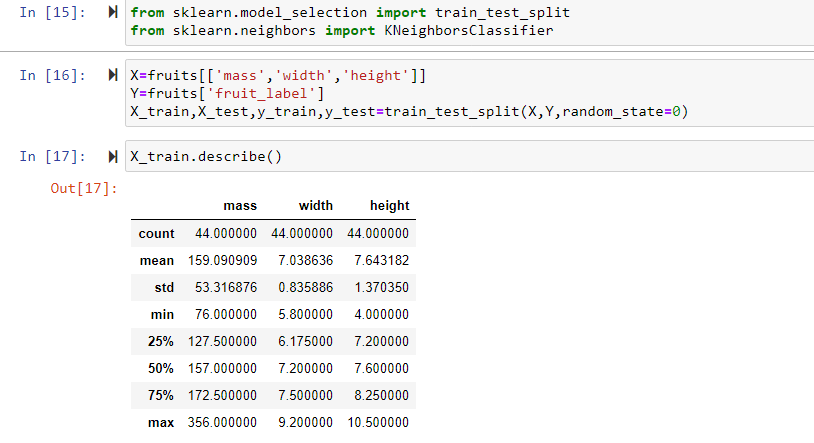
* Generate scatter plots for various combination of parameters and do the feature engineering meaning thereby which parameters of best suited to build the classifier.

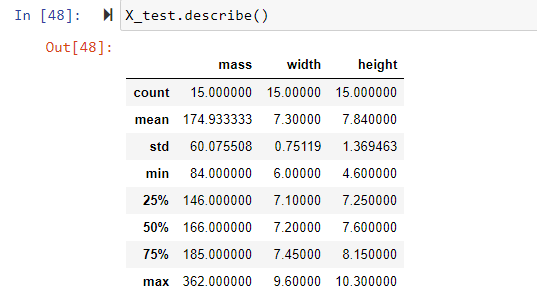




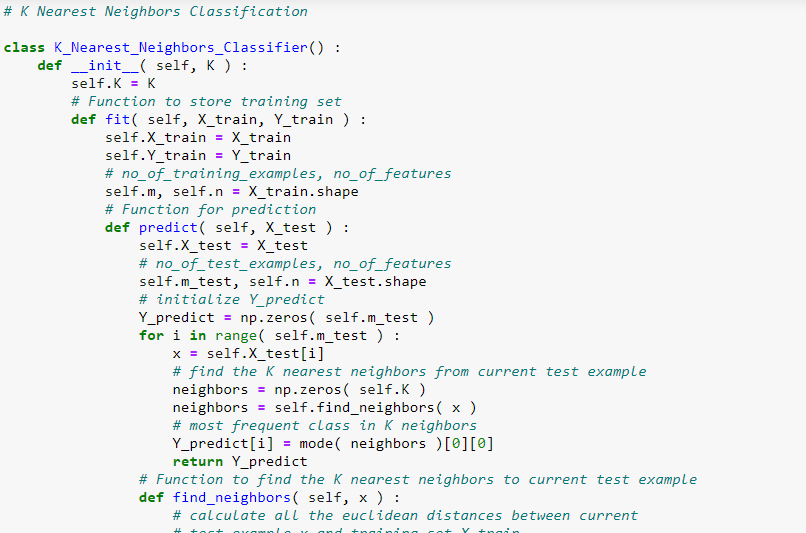


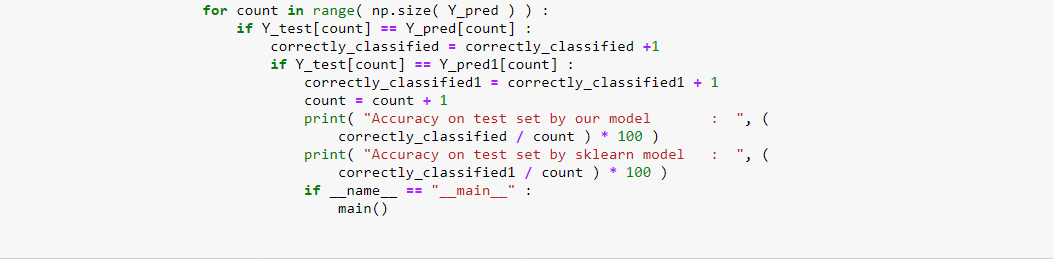
* Split the data into test and training split.



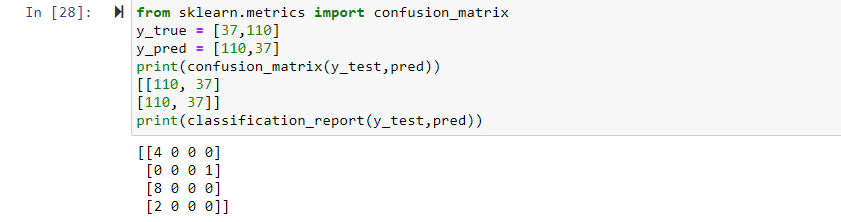


* Building a classifier using KNN from scratch

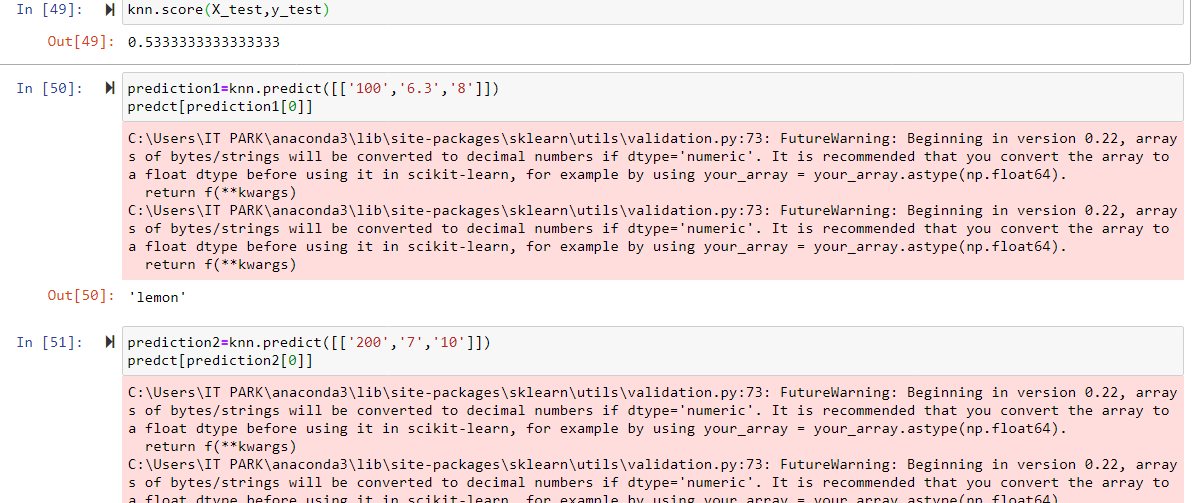


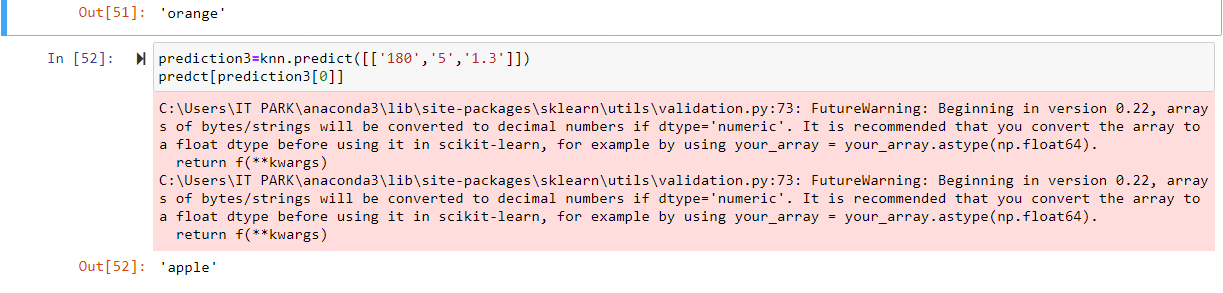


* Figure out the best value of k with highest r\_score.



* Run at least three test cases on the parameter and assess the fruit using the classifier.





**CONCLUSIONS:**

The summary of this problem is to generate K-Nearest neighbors algorithm from scratch, and to generate the scatter for two or more combination of parameters. Then the prediction of fruits by passing three or more test cases and access the fruit name.