**Curneu MedTech Innovations Assessment -Task 1**

**K-nearest neighbors classifier**

**Introduction:**

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithms.  is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

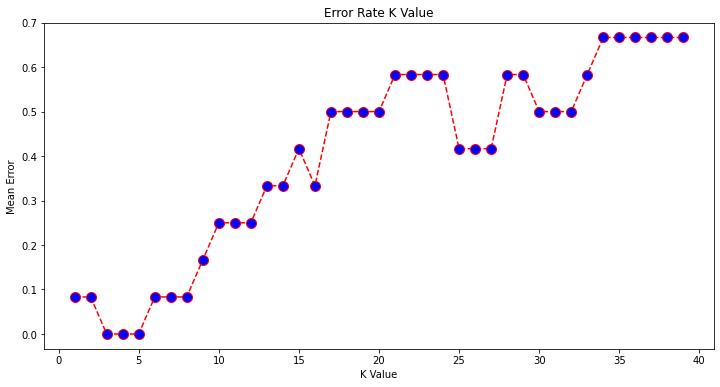
**Problem Statement:**

A dataset is given with labels based on fruits height, width, mass, colour score in a excel file. The task is to develop a KNN classifier machine learning model from scratch by analysing various combinations of parameters using scatter plot to find the best suited parameter combination to build the classifier model and to find out the best value of k with highest r2 score. Also, to run three test cases on the parameters and to access the fruit using classifier.

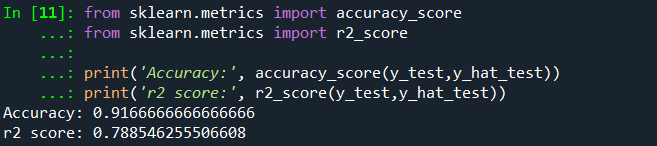
**Approach to the problem:**

Given dataset is read using pandas into a variable named dat. The description of the dataset and the first 5 rows of the dataset are printed. Then scatter plots are made for different combinations of features such as fruit height, width, mass and colour score. The correlation coefficient for every combination is also found to find the best possible combination of parameter. The best combination variables are separated from data as X and the target variable fruit\_label is chosen as y(target variable). X and y variables are then split into training and test variables. X\_train and X\_Test variables are then standardized to make the variables on the same scale. Then the functions for euclidean distance and knn is built from scratch. The developed knn function is then made to run under a loop to find the best value of k by spotting the least error giving k value. The model is passed with the best k value and the X\_test variable is used to predict the y value. The accuracy and r2 value of the predicted y variable is found.

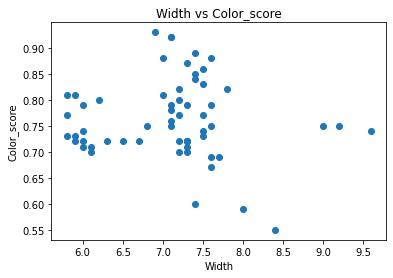
**Output:**

****

Interpreting the above graph, we find that minimum mean error occurs when the k values is between 3 to 5. Therefor the best suitable value of k is to be chosen between 3,4 and 5.

****

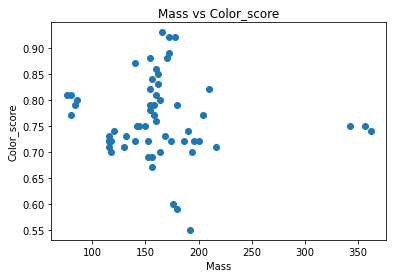
The accuracy score for the test data set is 0.916 which means that the model is 91% accurate to predict the target variable. The r2 value for the developed model is 0.78.

****

**Scatter plot for width and colour score**

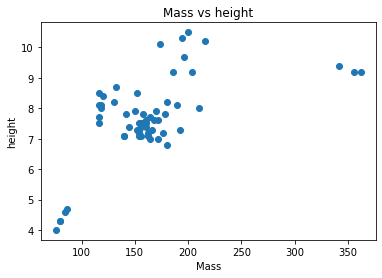
From this graph we could infer that the width and colour score is not very much correlated to

each other.

**Scatter plot for Mass and colour score**

From this graph we could infer that the mass and colour score is not very much correlated to

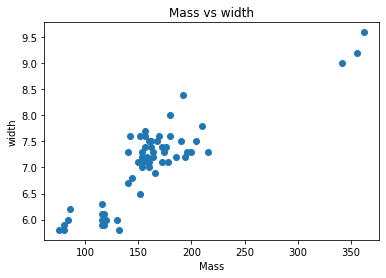
each other.

****

**Scatter plot for Mass and height**

From this graph we could infer that the mass and height is very strongly correlated to

each other.

****

**Scatter plot for Mass and width**

From this graph we could infer that the mass and width is very strongly correlated to

each other.