LUT
Lappeenranta
University of Technology

BM20A6100 Advanced Data Analysis and Machine Learning

ADAML PROJECT 2

Daniela Maldonado

Dilhara Liyanaaratchi

Shageerthana Sathiyamoorthy

PROCESS PLAN

Our goal is to build a naïve bayes structure where classify the fruits. Regarding the data set, as there were many sub-categories for apples, pears, grapes, nuts, and maroon, here we considered images of one sub-category as for one class data. And at the end, the final data set has nearly 490 observations for each fruit classes.

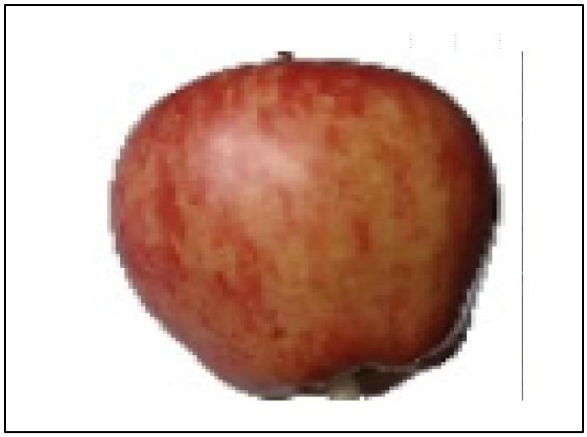
As far as for the features, at the moment we are considering [1]statistical parameters, [2]histogram count (peak) and the [3]edges of the image. Based on the accuracy of model that will be build, more features will add to the process.

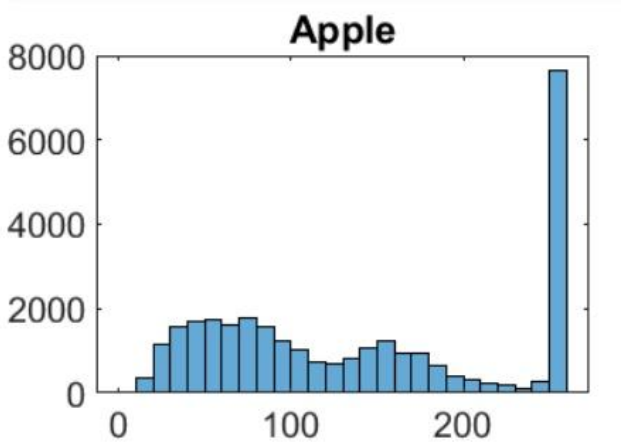
In order make sure the independence of the observed variables, the features will be normalized and will be using for the process.

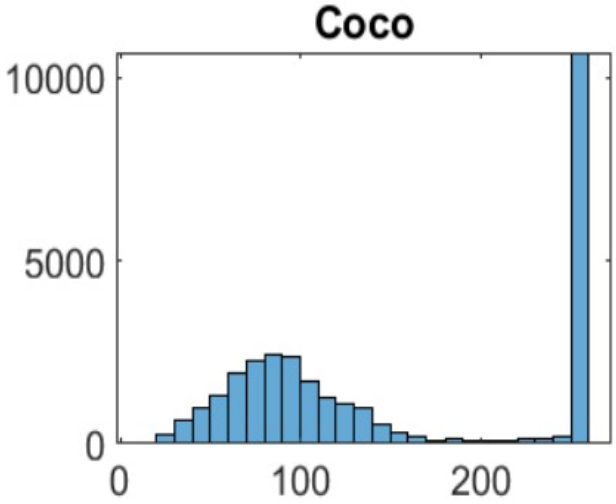
Histogram Count

As per the histogram and the following images of apples and coco, it is observed that the peak count is for the white area. Hence the that peak value will be removed, and the next peak value will be use for the model.

[1] Dilhara Liyanaaratchi
[2] Daniela Maldonado
[3] Shageerthana Sathiyamoorthy







Edges of the Images

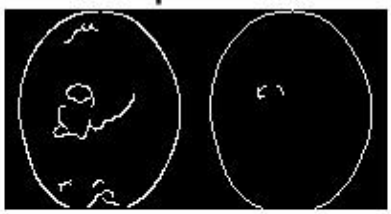
In order to take edges of the images, we used 2 methods, called canny and prewitt methods. For the method canny it gives more edges compared to the prewitt method. As per the following images,

Apple Red1

Cactus fruit

Cocos

Grape White



Hazel nut

Maracuja

PearMonster

Raspberry

Watermelon

Plum 3

As per the above images, the fruits apples, cocos, grape white, water melon and plum are seems to have the same shape. Therefore, the it could give the similar value for the edges.

Intermediary submission 2

After performing several trials of feature processing with the above-mentioned features, such as Histogram count and Edge detection. We decided to proceed with RGB feature extraction. Where our data is the images of fruits, we thought it would be more accurate to classify the images using the color.

Below you can find the Bayesian Network from bins 2 to 7. As per the figures you can see that there has been no change in the model.

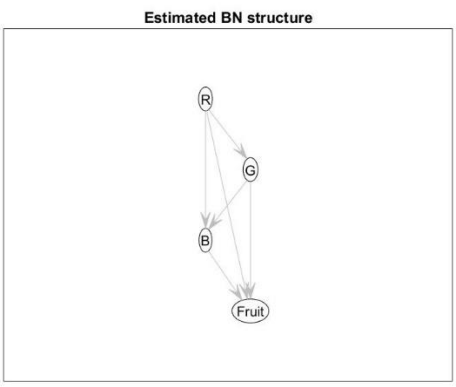


Figure 1: BN for 4 Bins

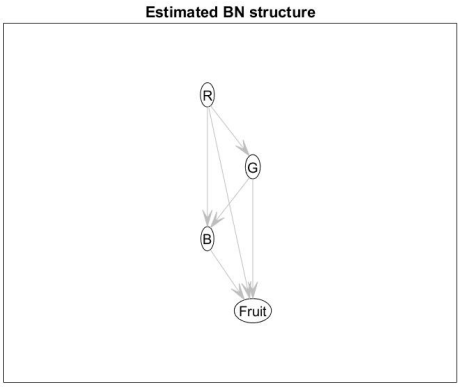


Figure 2: BN for 2 Bins

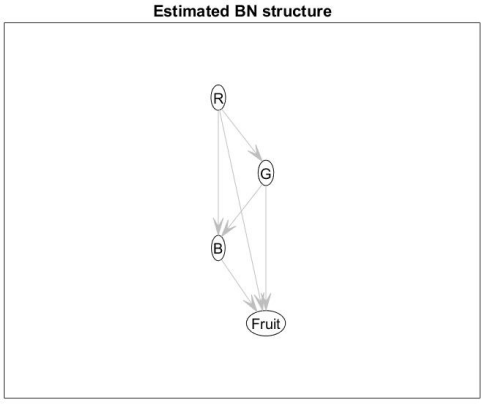


Figure 3: BN for 6 Bins

Confusion matrix

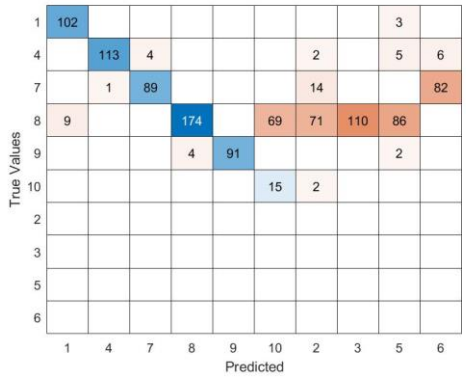


Figure 4: Confusion Matrix at 2 Bins

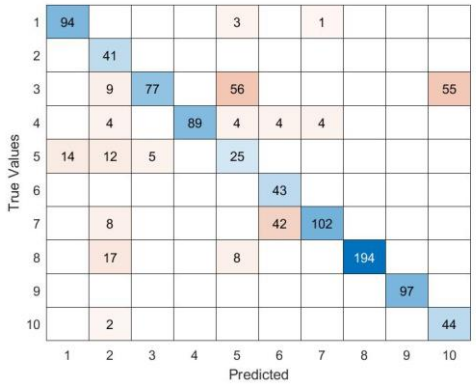


Figure 5: Confusion matrix at 4 Bins

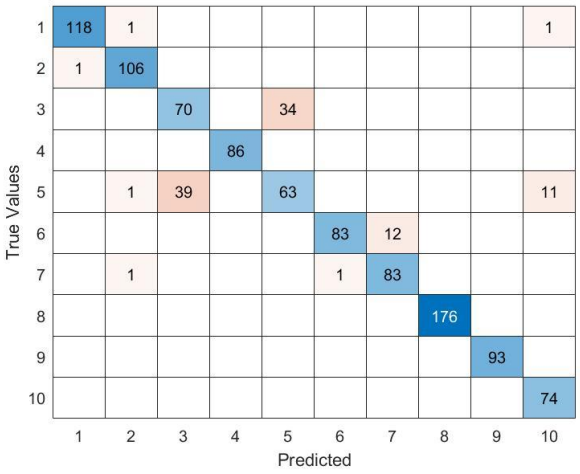


Figure 6: Confusion matrix at 6 Bins

According to the confusion matrix, we can say that Plum is the most accurately predicted fruit. While Cocos and Cactus fruit are the least accurately predicted fruits. But once the bins are increased, out of all the fruits Maracuja is considered to be the least accurately predicted fruit.

Accuracy of the model

No.of bins	2	3	4	5	6	7	8
Accuracy	55%	71%	76%	79%	90%	90%	90%

We finalized the number of bins to be 6 as the accuracy remained the same afterwards. Hence, we performed the Bayesian Network for the testing data set as well with the same number of bins and achieved 92% of accuracy level.