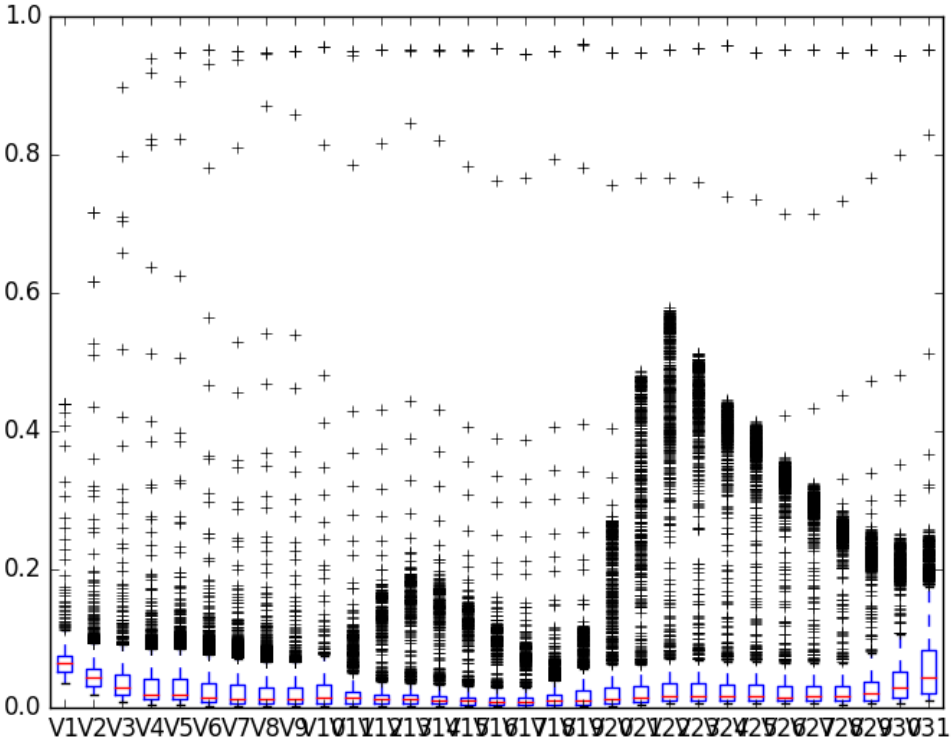
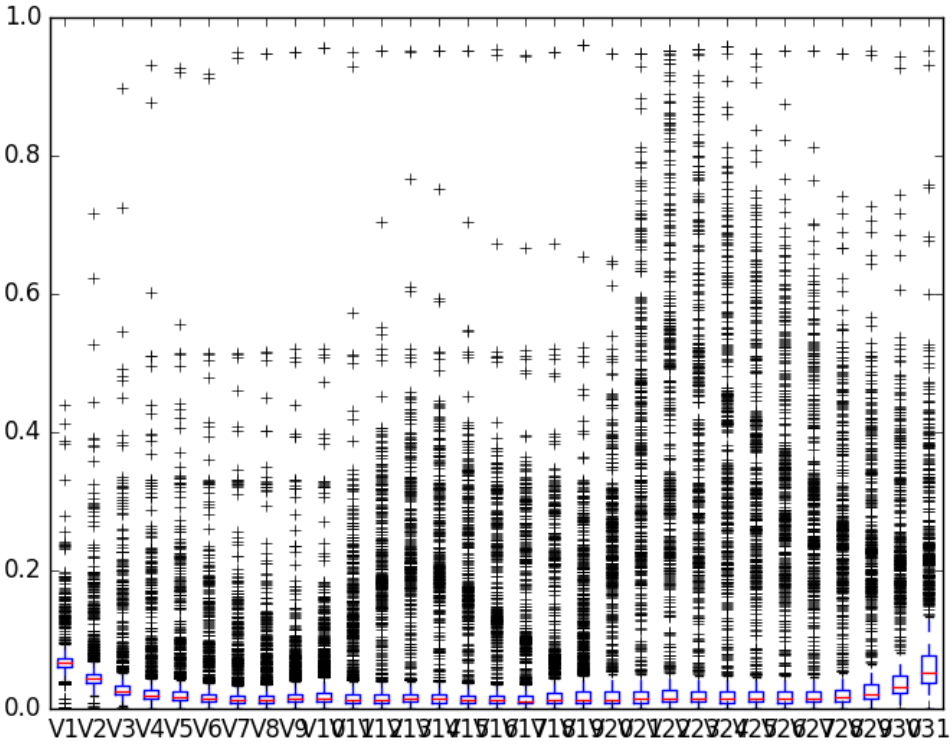
Question 1:
Distribution of the ‘Red’ and ‘Background’ data features.

Red:



Background:



Question 2:

trained LDA, 10 fold cross validation

in bag
```
Accuracy: 0.74 (+/- 0.04)
```

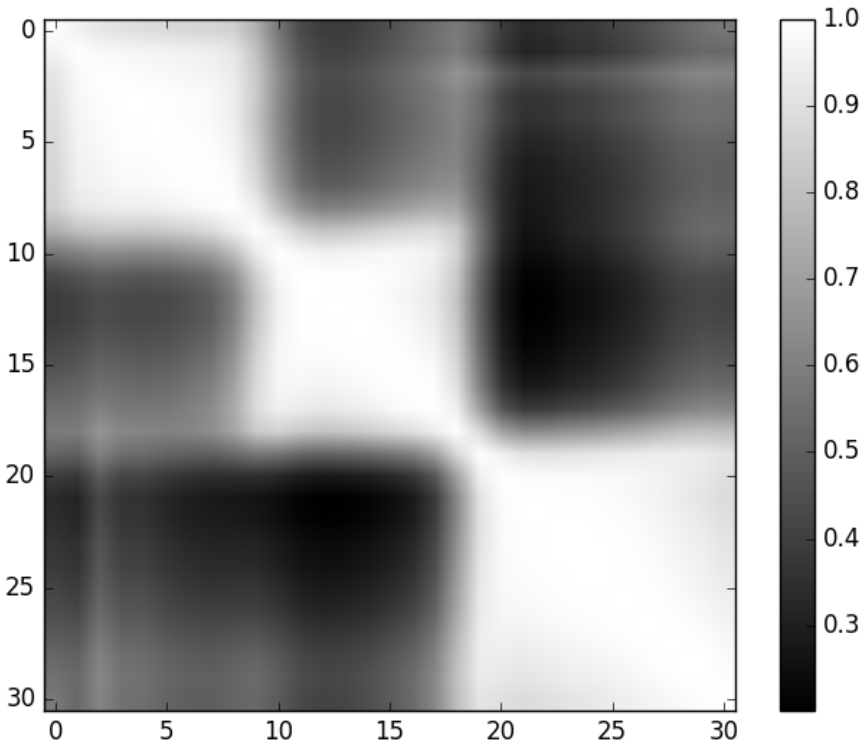
Question 3:

```
Test confusion matrix
[[1669  342]
 [ 666 1345]]
```

out of bag accuracy (test)
```
0.749378418697
```

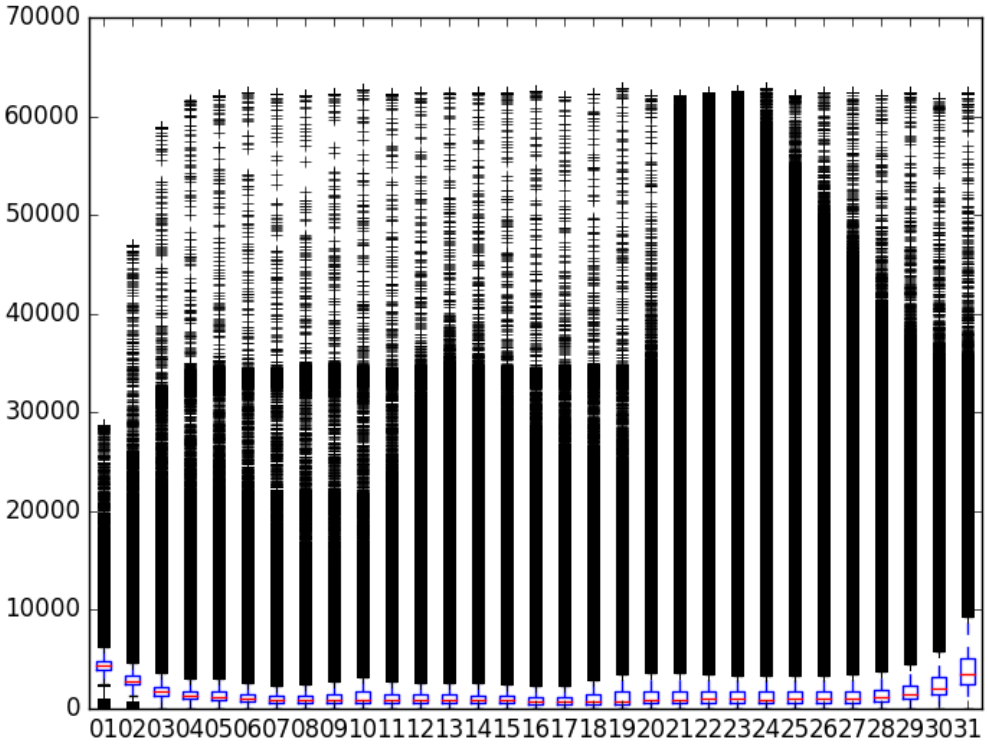
Question 4:

The correlation matrix of the training dataset is plotted as a 2-D image in gray scale below. Looks like adjacent features, e.g. V1 to V10 and V20 to V31, have pretty high correlations between each other.



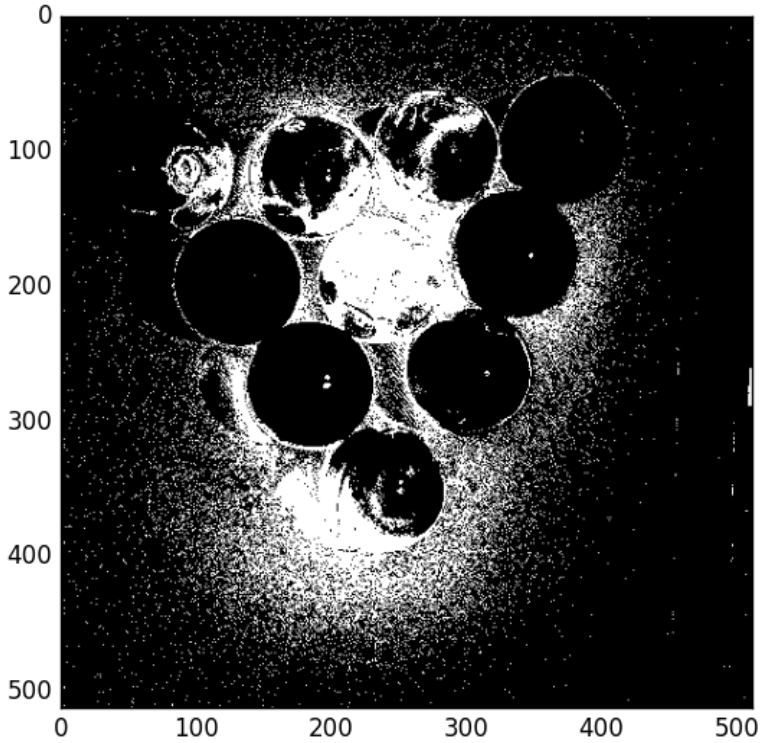
Question 4a:

Read the png files in to dataframe of 31 columns. The following data box plot shows the distribution of data for each column.

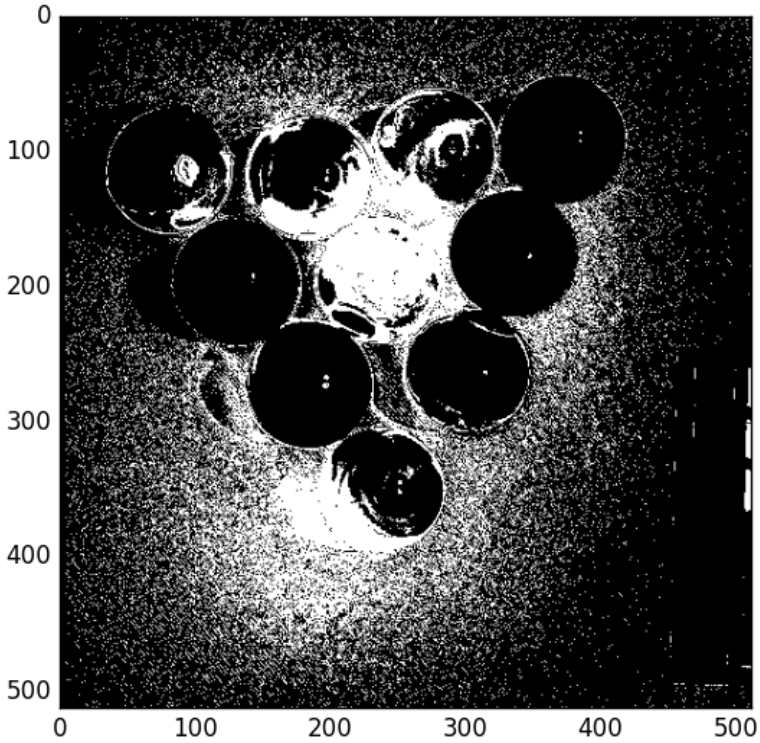


Question 5:

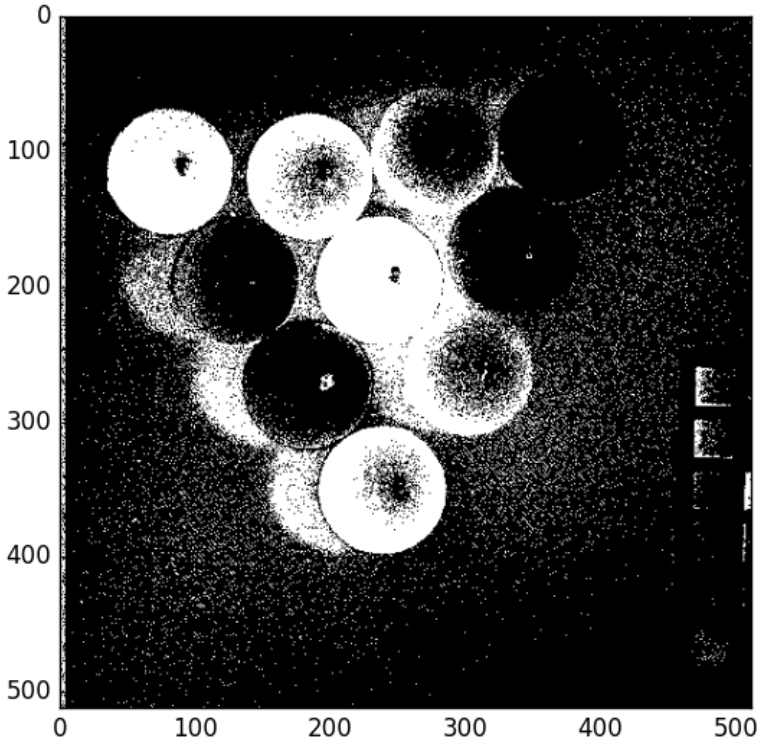
Prediction result from trained random forest model (10 estimators).



KNN



Prediction result from LDA



Question 6:

The accuracy score of the random forest model is
`Accuracy: 0.821212768555`

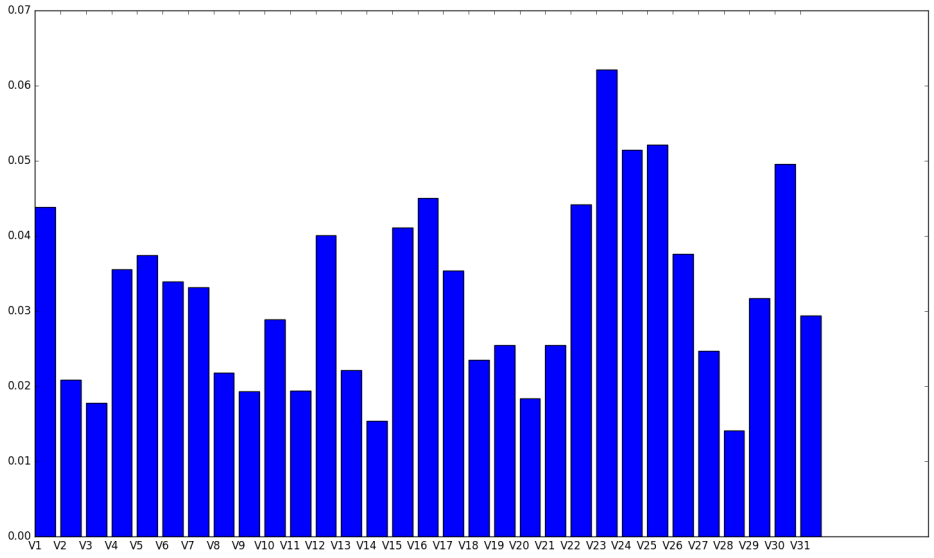
The accuracy score of KNN(k=1)
`0.775695800781`

The confusion matrix and accuracy score of LDA are

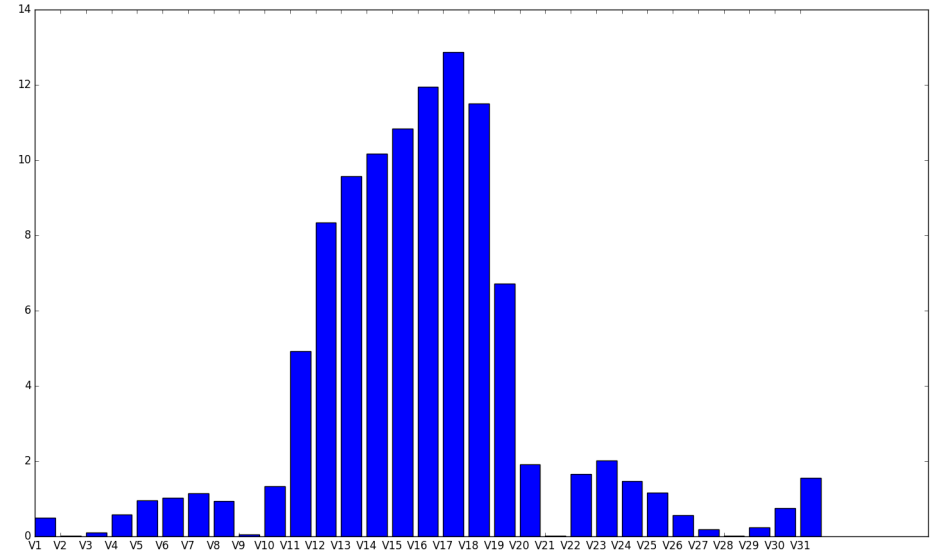
```
[[207622  41760]
 [  1570  11192]]
0.83470916748
```

Question 7:

The feature_importances_ attribute of random forest is plotted as a bar chart below. V23 is most important feature in the data for the random forest classifier.



The scores_ attribute from SelectKBest of sklearn package:



V17 is the most important feature.

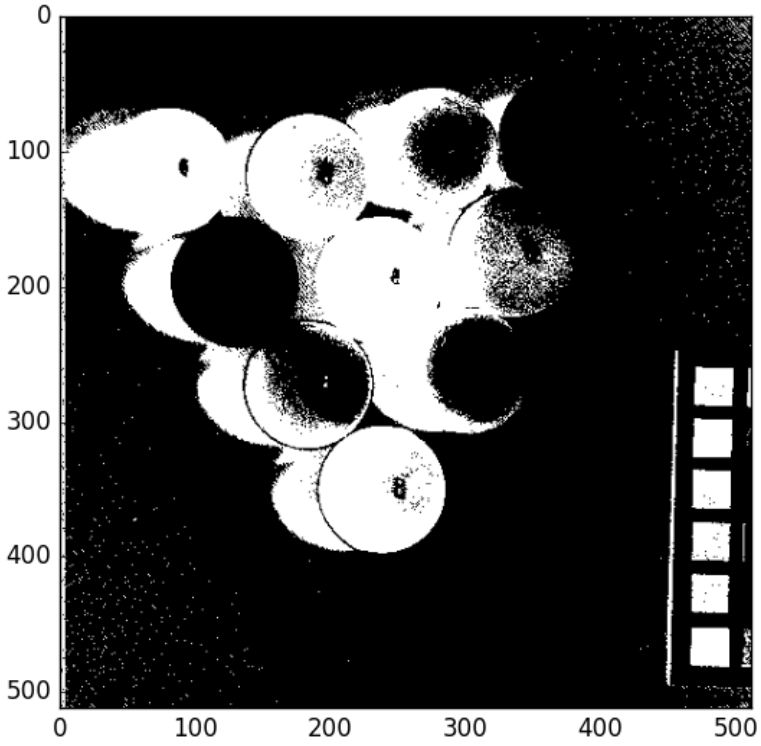
Question 8:

LDA seems to have the highest accuracy score, although there are many false positives.

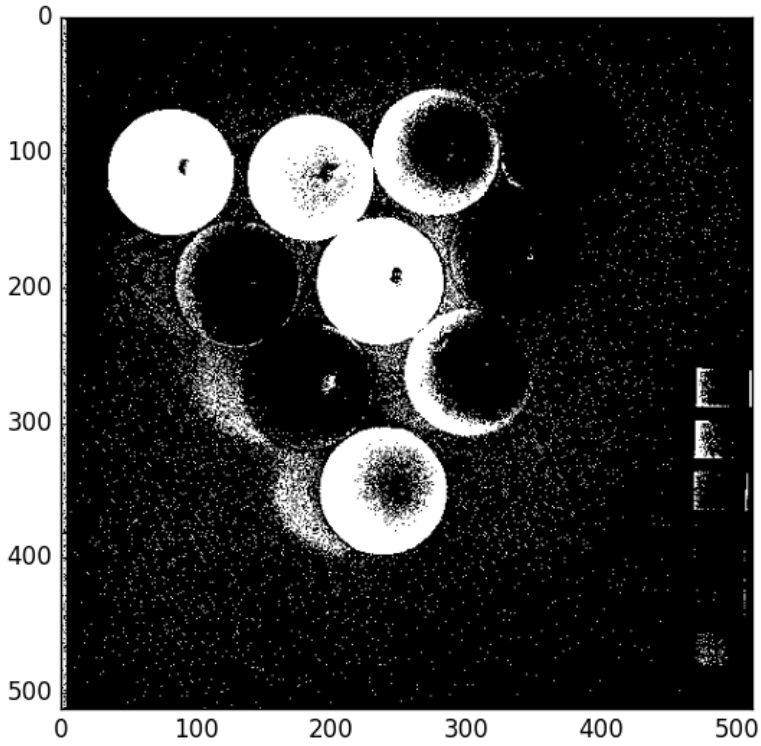
Question 9:

The ranges of the training data and the raw data are drastically different. The prediction of the raw data cannot be done without normalization.

Here is the result of preprocessing with PCA (10 components):

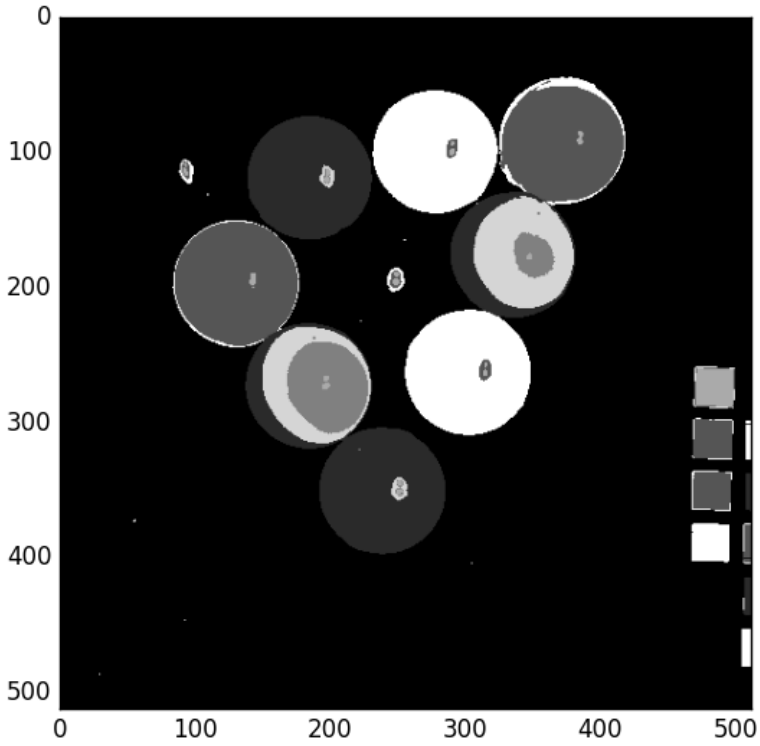


Here is the result of preprocessing with PCA (15 components):



Question 10:

```
Kmeans  n = 7
```



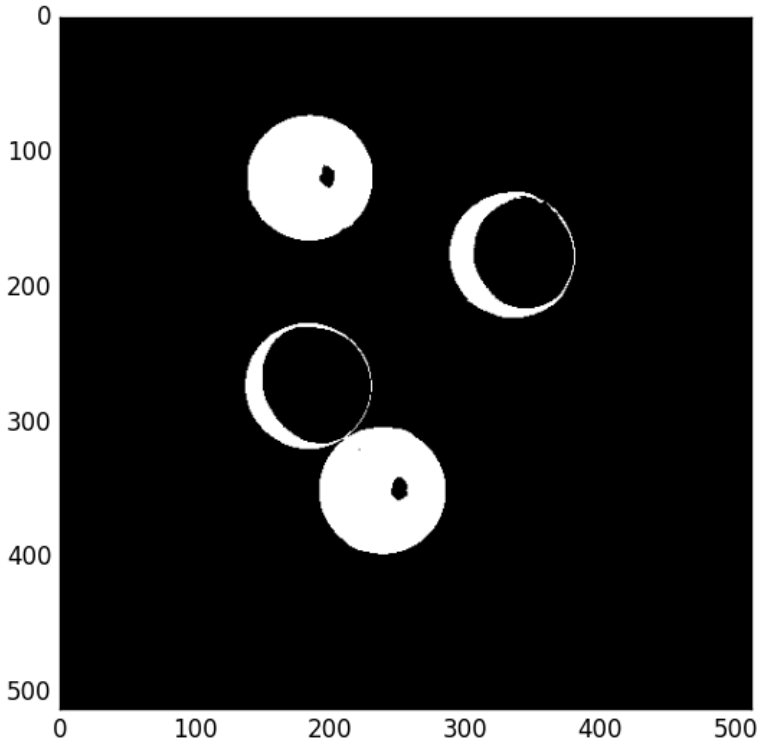

```
Spectral Angle Mapper
SAM(red_vec,n_vec)

Spectral angle between red and each of the clusters

n = 0
1.02058756124
n = 1   (red)
0.0427734158977
n = 2
0.786061228408
n = 3
0.228536872241
n = 4
0.817991347889
n = 5
0.192797740726
n = 6
1.19850945215
```

The spectral angle between red_vec (based on the mask) and the pink_vec (based on the training dataset) is `0.449854512666.` The training data are mislabeled?

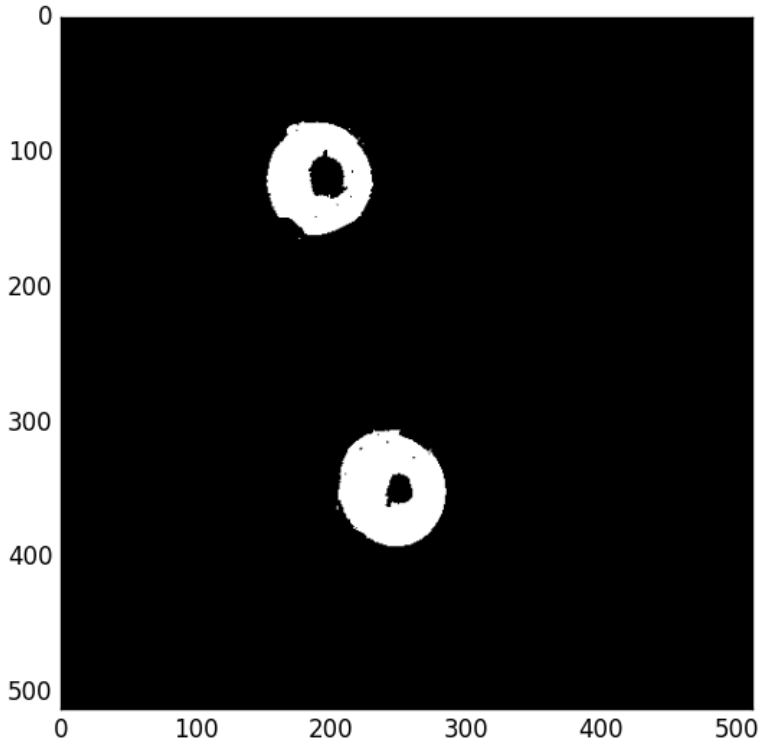
```
red_vec (based on Red_Mask.png)
01    0.060547
02    0.037405
03    0.036399
04    0.021189
05    0.020102
06    0.013204
07    0.010558
08    0.009856
09    0.009772
10    0.009054
11    0.008946
12    0.009455
13    0.010228
14    0.010061
15    0.009424
16    0.009482
17    0.013022
18    0.029068
19    0.076515
20    0.188466
21    0.362177
22    0.473328
23    0.444790
24    0.393083
25    0.364511
26    0.321195
27    0.283435
28    0.243242
29    0.213019
30    0.197569
31    0.202146
dtype: float32

pink_vec (based on training data)
V1     0.066603
V2     0.052576
V3     0.040367
V4     0.037603
V5     0.038276
V6     0.034310
V7     0.030936
V8     0.028983
V9     0.026211
V10    0.023236
V11    0.023527
V12    0.026143
V13    0.027817
V14    0.026157
```

```
V15    0.022949
V16    0.019576
V17    0.017490
V18    0.018588
V19    0.025081
V20    0.043082
V21    0.074443
V22    0.096931
V23    0.093682
V24    0.084976
V25    0.080331
V26    0.072609
V27    0.066135
V28    0.059909
V29    0.057324
V30    0.059901
V31    0.072827
dtype: float64
```

Randomly choosing 4000 points from the Red_Mask.png as labels, I created a new training dataset. The trained new model (SVM, Kmeans and SAM) yielded good classification result.

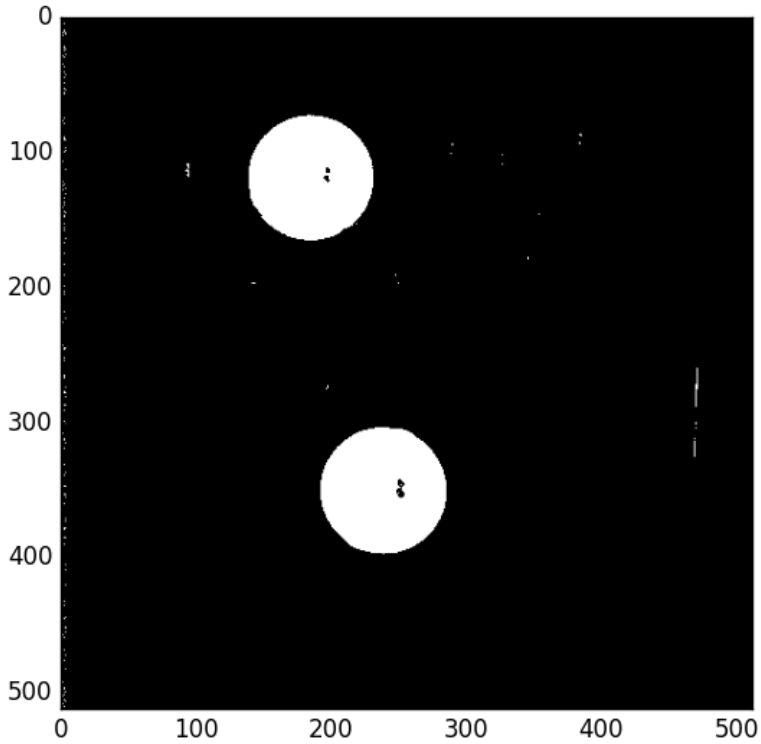


Lastly, I randomly chose 500 points from the Red_Mask.png as labels, creating a new training dataset. I used Kmeans, SAM, SID, Chebyshev and NormXCorr to create new features for pixels. The red vector (as reference for calculating the SAM etc.) was averaged from the red points in the 31 channel data files. The trained SVM model yielded good classification result.



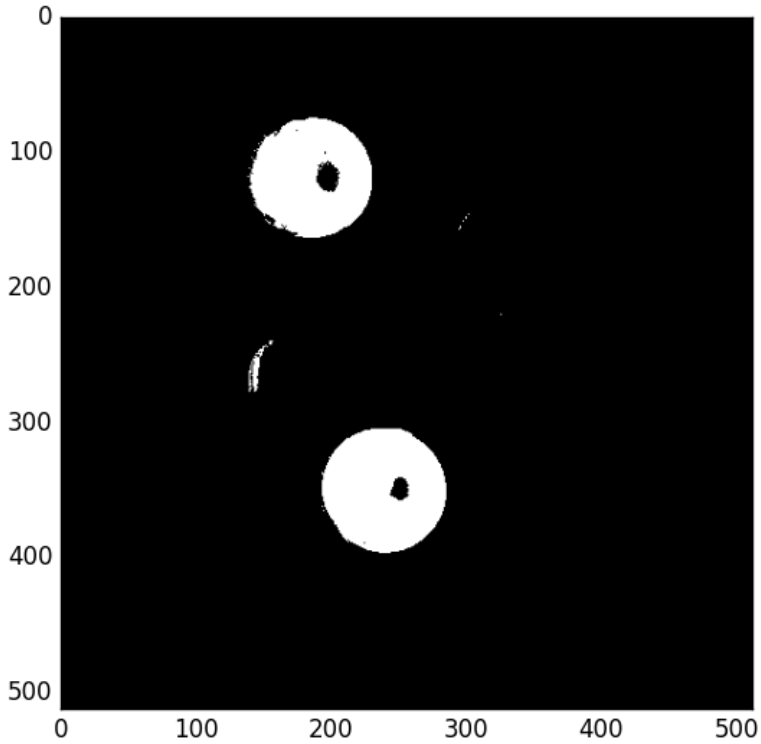
confusion matrix:

```
[[249280    102]
 [  3612   9150]]

accuracy:
0.985832214355
```



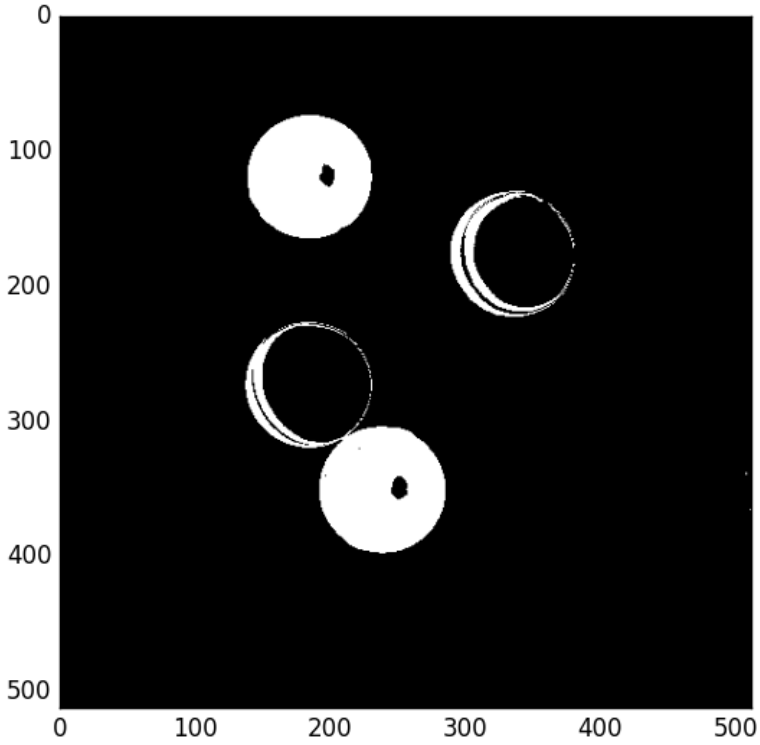
LDA

```
[[248261   1121]
 [   168  12594]]
0.995082855225
```



Random Forest Classifier

```
[[248911    471]
 [   673  12089]]
0.995635986328
```



KNN

```
[[245839   3543]
 [   206  12556]]
0.985698699951
```