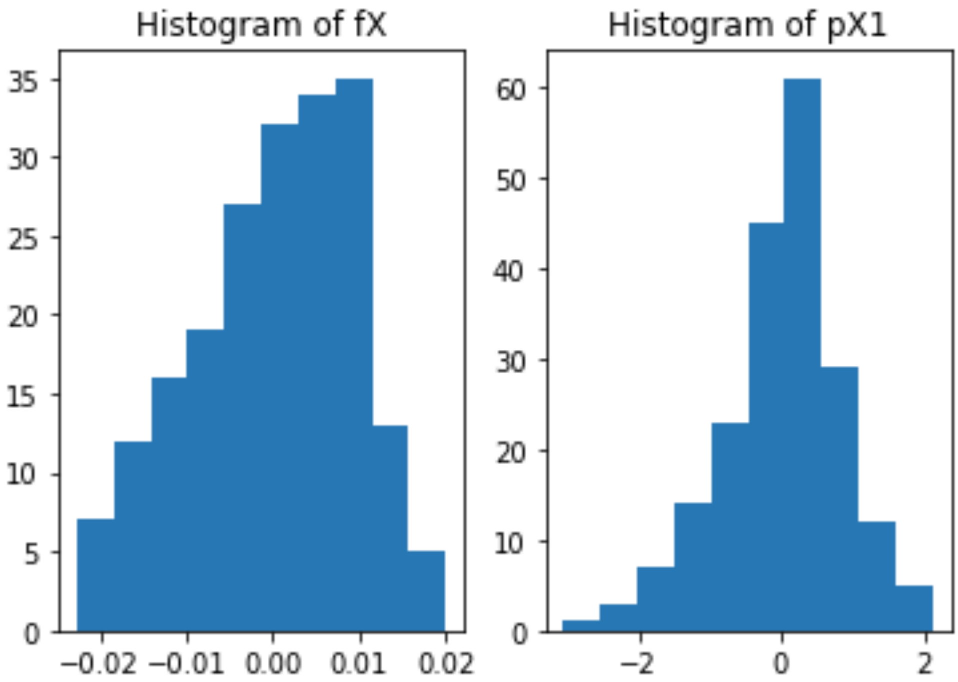
Harshvardhan
Student Id: 609162

Question 1

All the generated datasets are in the appended Python notebook. The histogram generated is appended below.



pX1 might look like a Normal distribution with very low kurtosis. However, fX is not Normal looking. It probably makes more sense to apply a non-parametric method like kNN in this case to the data.

	Overall Accuracy	Class-0 Accuracy	Class-1 Accuracy	Runtime
nX	0.710	0.650	0.834	0.029
fX	0.614	0.439	0.972	0.018
pX1	0.394	0.210	0.770	0.044
pX	0.328	0.0	1.0	0.009

Table 1: Results for Case 3 Classifier.

	Overall Accuracy	Class-0 Accuracy	Class-1 Accuracy	Runtime
nX	0.765	0.892	0.504	0.358
fX	0.759	0.883	0.504	0.388
pX1	0.647	0.816	0.302	0.527
pX	0.663	0.816	0.348	0.313

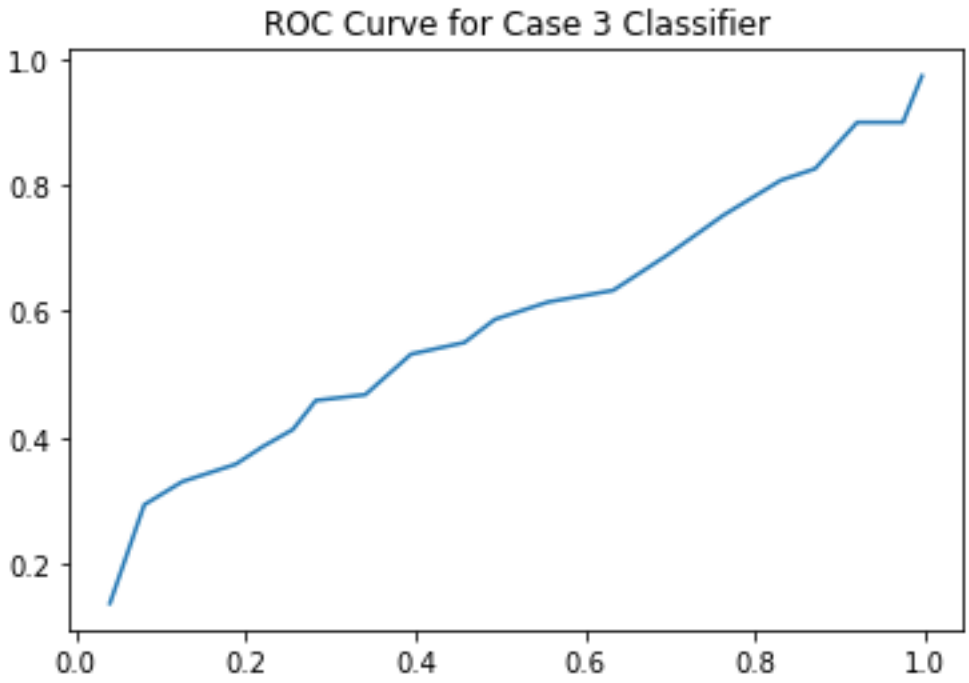
Table 2: Results for kNN classifier with k = 9.

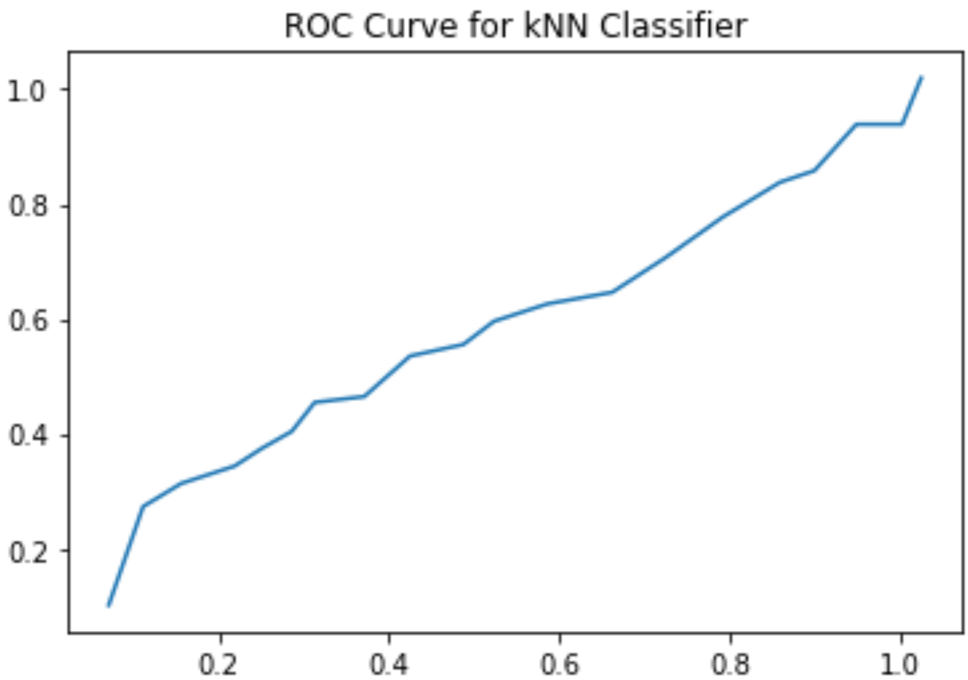
Confusion Matrices for Case 3 Classifier

```
[[145.  78.]
 [ 18.  91.]]
[[ 98. 125.]
 [  3. 106.]]
[[ 47. 176.]
 [ 25.  84.]]
[[  0. 223.]
 [  0. 109.]]
```

Confusion Matrices for kNN Classifier

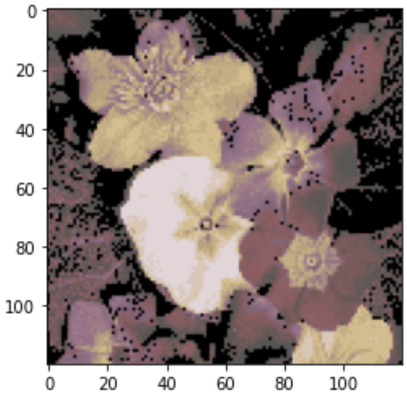
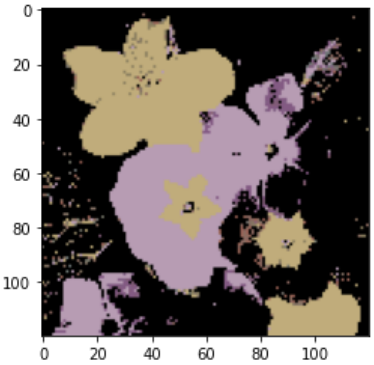
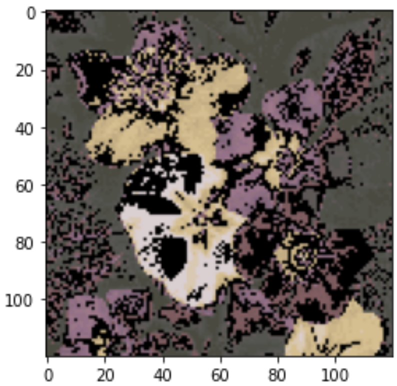
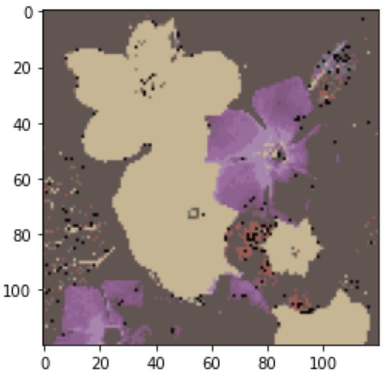
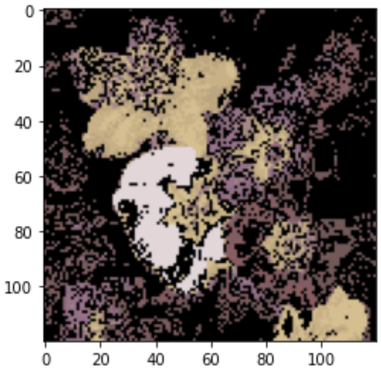
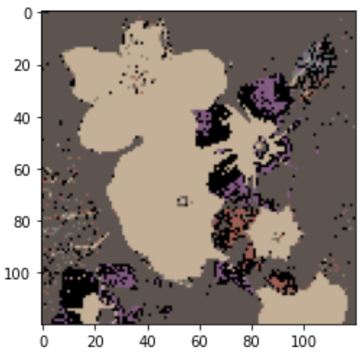
```
[[199.  24.]
 [ 54.  55.]]
[[197.  26.]
 [ 54.  55.]]
[[182.  41.]
 [ 71.  38.]]
[[182.  41.]
 [ 76.  33.]]
```

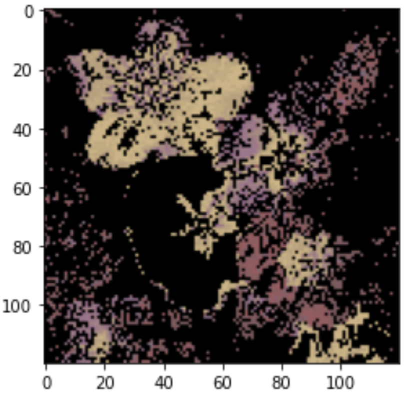
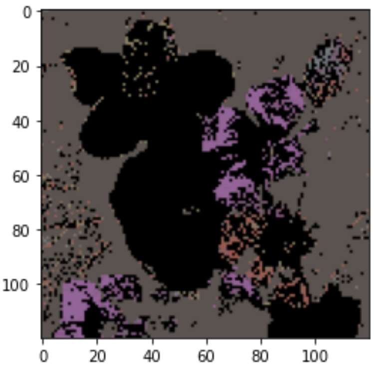




The codes to generate all the figures and results are in the attached Python notebook.

Question 2

	kmeans	wta
K = 32	 RMSE = 0.33	 RMSE = 0.48
K = 64	 RMSE = 0.49	 RMSE = 0.266
K = 128	 RMSE = 0.49	 RMSE = 0.33

K = 256	 RMSE = 0.577	 RMSE = 0.55

When we reduce the number of dimensions in the image, the image loses the lustre. The image has details which are progressively lost as we keep reducing the image size. However, it is beyond appreciation what details could be preserved with the 128x128 image over the original image. This loss is so low that it can be appropriately called loss-less compression.

There are a lot of points that are represented as black in the plot. I think they are because the values are really small at certain places. If I allowed my clustering to run longer, the colours could've improved but after running for an hour for each "k", this are the results obtained.

The images obtained from WTA_Kmeans is worse in general than normal k-means algorithm. This is possible because it is an image and different points — although close — will have very different colours and thus very different points.

Final Discussion

Through this project, we found that Principal Components and Fisher's Linear Discriminant can be used for reduction in dimensions. Thus, we can effectively represent or discriminate between classes better with fewer features than with the whole data. We will lose on some accuracy but we will gain a lot from the computational easiness.

With the image size reduction, we learn that kmeans and wta algorithms that have been previously used for clustering pima dataset, can also be used for clustering the pixels and thus help in image compression. With the analysis, we found that some degree of reduction can be done directly using fewer clusters. At k = 256, the computation to cluster 14400 points is huge and takes more than an hour. However, the result we get is amazing and is almost "loss-less".