Dimensionality Reduction Findings

I was given a dataset from a company (*Wine.csv*). The dataset contained 13 columns which represented different ingredients in their different wines. The 14th column represented customer segment. This indicated what particular wines customers would buy. They had separated their customers into three groups.

Our goal was to use dimensionality reduction in order to predict what group a customer would belong to so that the company can give them the wine(s) most attuned to them.

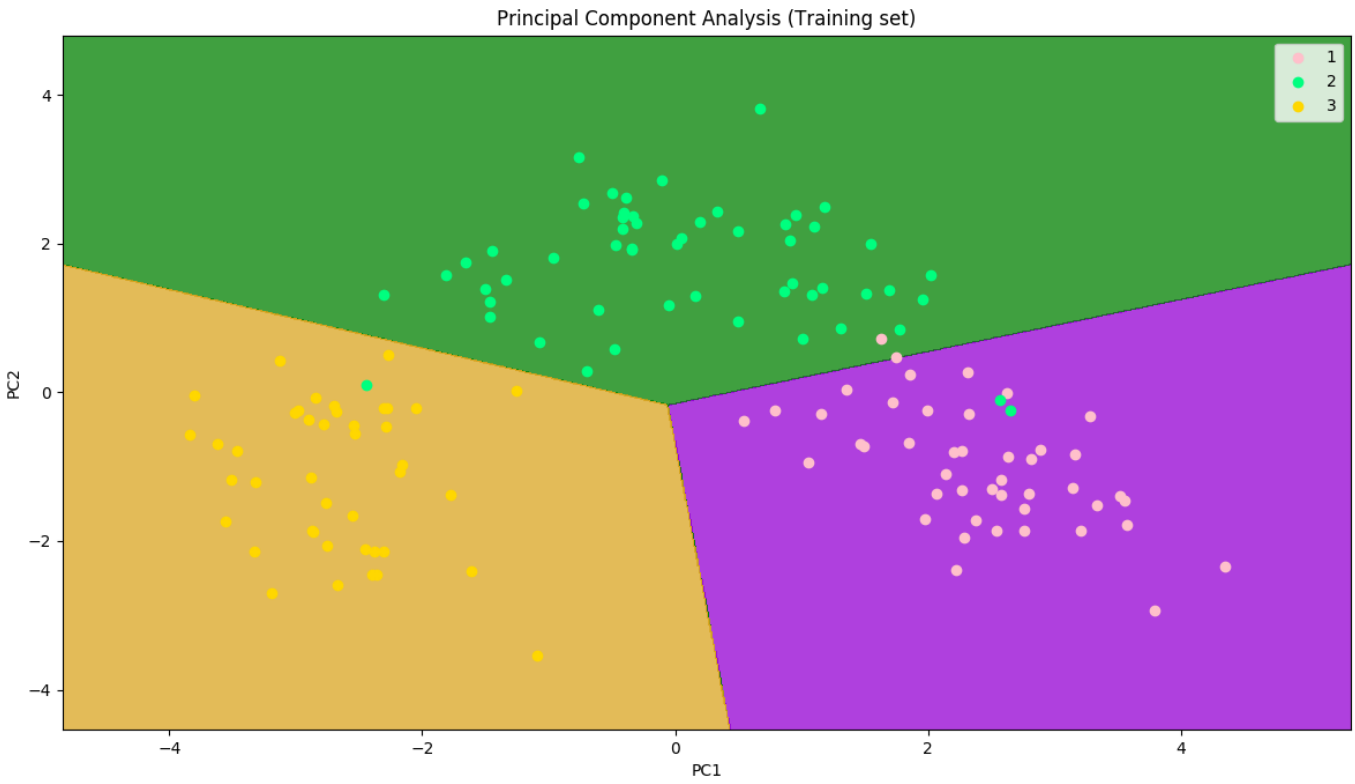
Principal Component Analysis (PCA) Findings

PCA is a feature extraction technique. This is an unsupervised model because we don't consider the dependent variable in the algorithm.
I extracted the top two most important components and applied PCA. After creating the confusion matrix I received the following output.

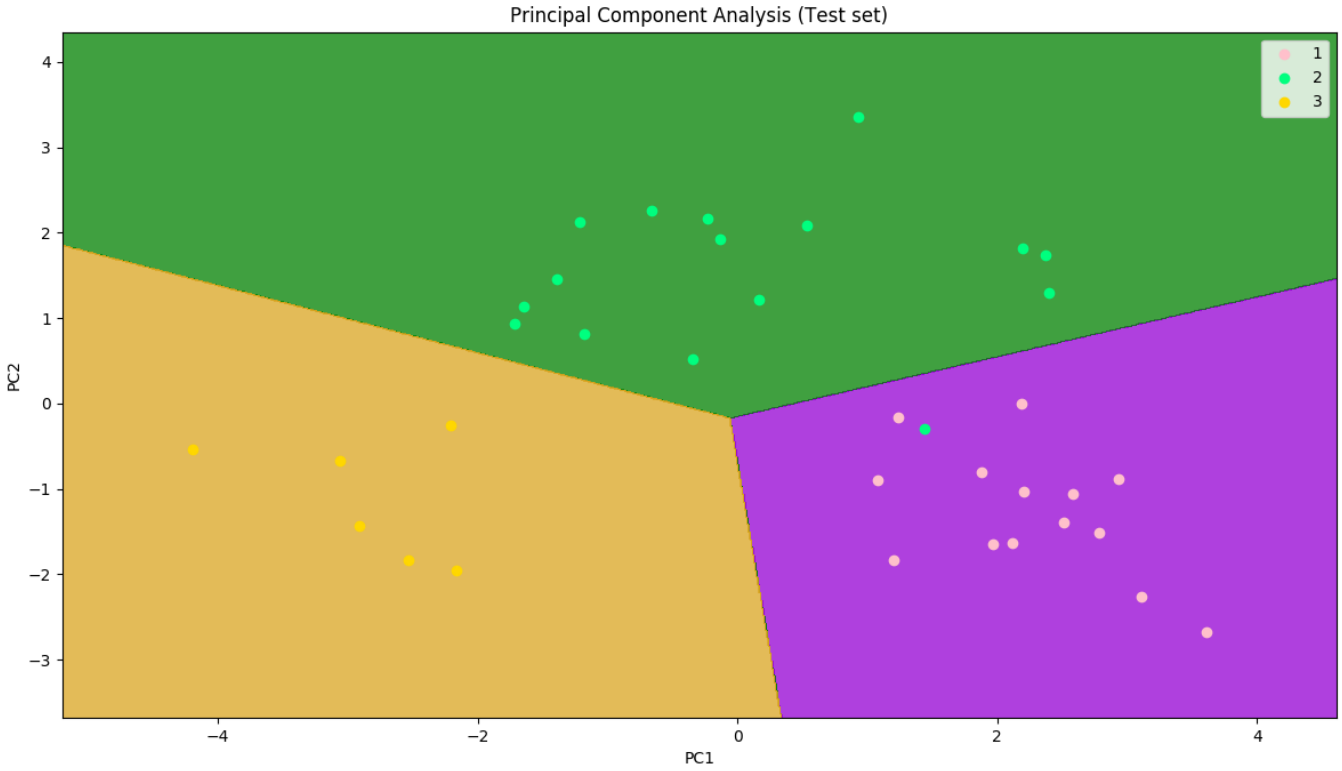
There are 35 our of 36 correct predictions (97.22% accuracy). Other than 100% this is the best accuracy we can obtain.

	0	1	2
0	14	0	0
1	1	15	0
2	0	0	6

Training Set



Test Set



Here you can see the one incorrect prediction; the green dot in the purple space.

Linear Discriminant Analysis

PCA is a feature extraction technique. This is a supervised model because we consider the dependent variable in the algorithm. I extracted the top two most important components and applied PCA. After creating the confusion matrix I received the following output.

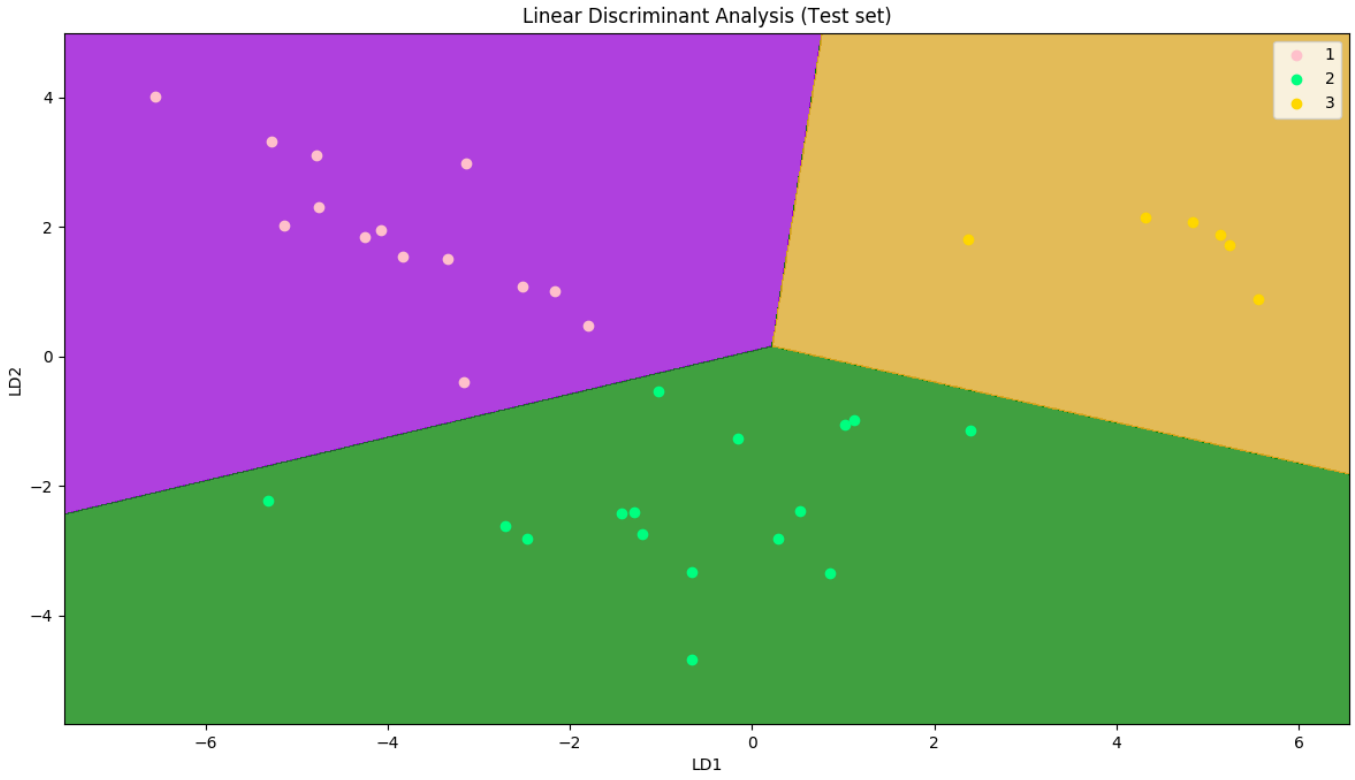
We received 36 out of 36 correct predictions (100%).

	0	1	2
0	14	0	0
1	0	16	0
2	0	0	6

Training Set

You will notice a green point in the gold section and several green and gold points relatively close to one another. These points are considered outliers and are therefore not counted by the algorithm.

Test Set



Kernel PCA

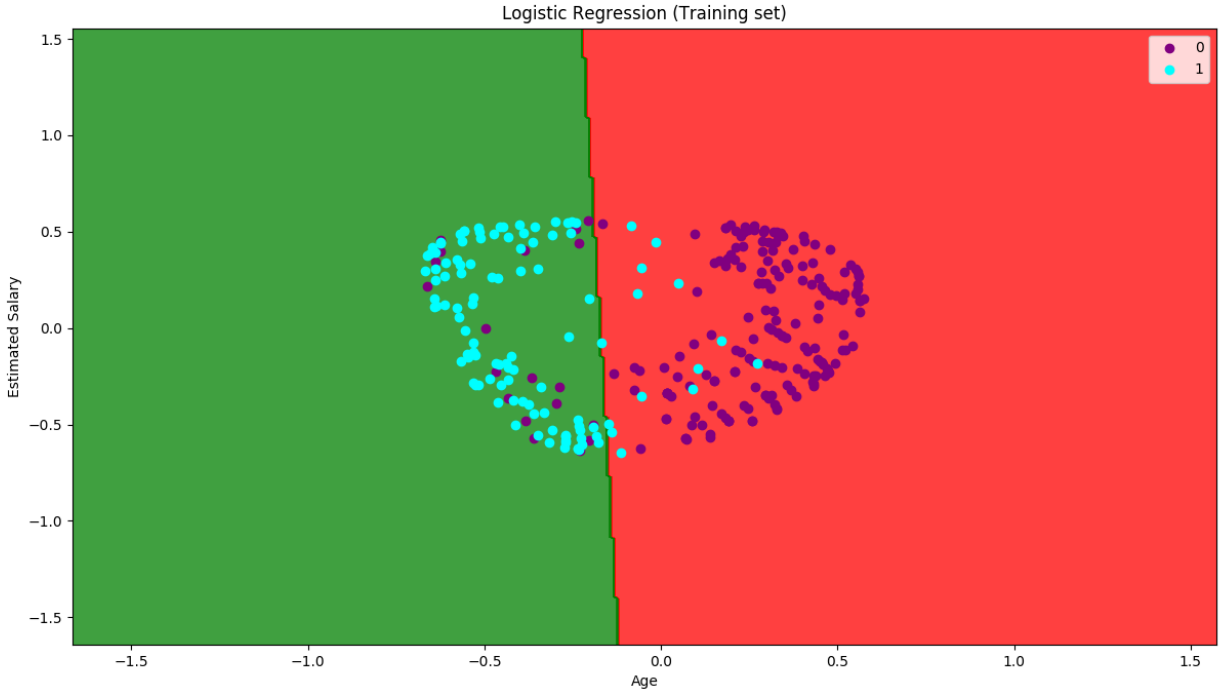
This is a non-linear algorithm. Because of this we cannot use the same dataset. The dataset we will be using is *Social_Network_Ads.csv*. Looking at the confusion matrix. We have 90 correct predictions with 10 incorrect predictions (90% accuracy).

	0	1
0	64	4
1	6	26

Taking a look at the graphs.

- The purple points are the customers who didn't click on the advertisement
- The cyan points are the customers who clicked on the advertisement
- The red region is the predicted region of customers who didn't click the advertisement
- The green region is the predicted region of customers who clicked the advertisement

Training Set



Test Set