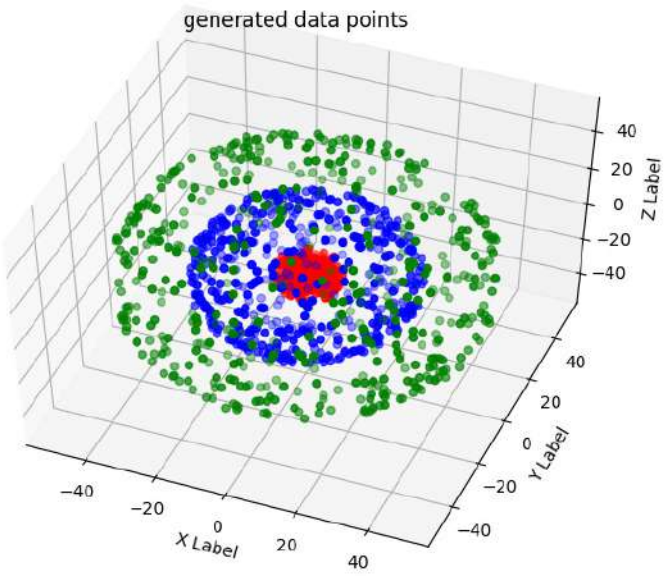
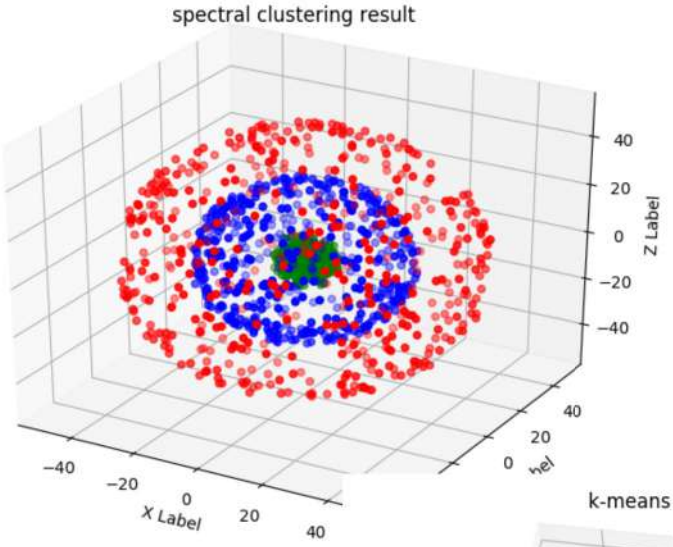
Machine Learning Assignment
vishal kumar chaudhary
111501030

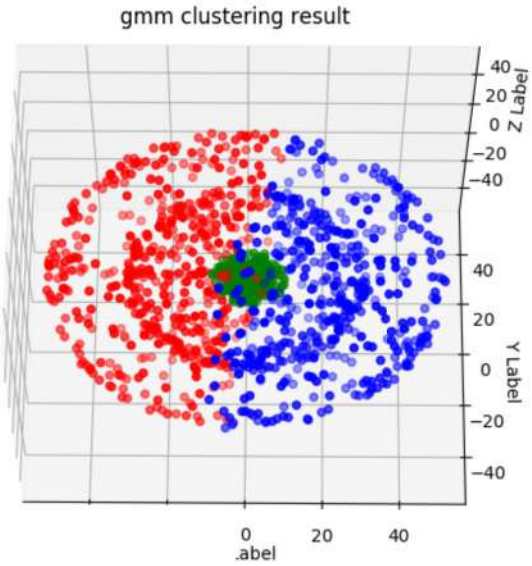
PART 1 :
Data generation : data is generated with uniform random function which are three spherically symmetrical balls
.

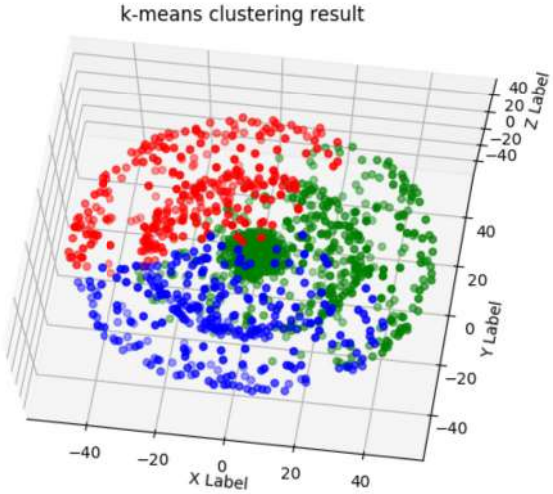


clustering algorithm apllied on data like k-means , spectral clustering ,gmm models

Performance of model are
Spectral clustering > gmm > k-means







Since the data was radially symmetric ,that is why spectral clustering model performs better than others

In k-means data , cluster is made one the basis of centre of cluster , so it did not perform good on radially symmetric data .

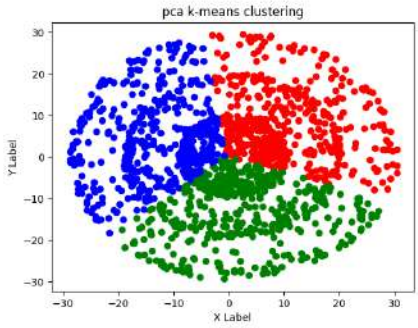
In GMM , It is able to figure out the underlying model of inner most cluster but not for the outer shells , so it performs better than k-means but not better than spherical model

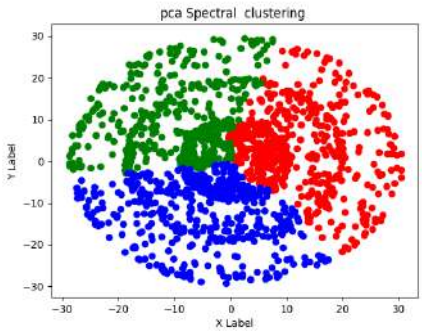
PART 2 :

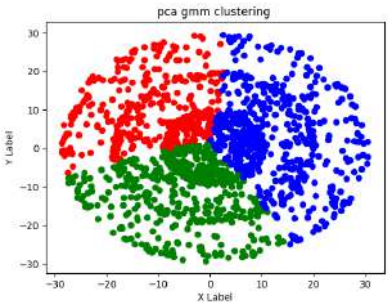
PCA :

3 dimension into 2 dimension :

Since we are lossing some information ,and principal component is same in all the diameters so the spectral clustering does perform similar to k-means clustering but GMM is able to recognise the inner most sphere but not outer so it divides the outer shell into two cluster .

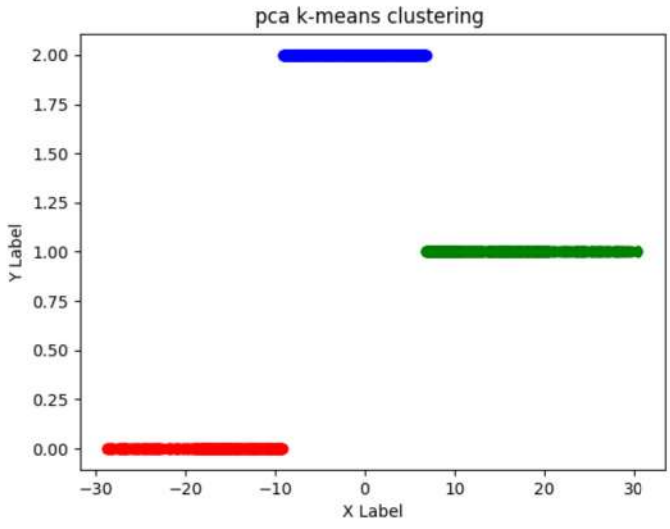


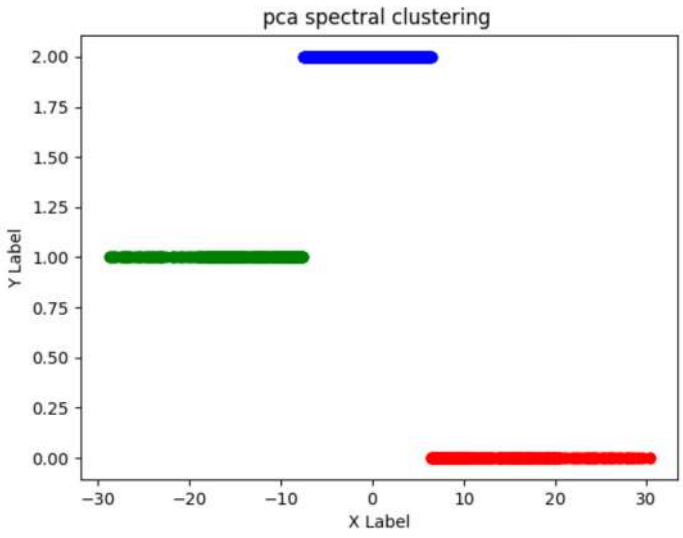


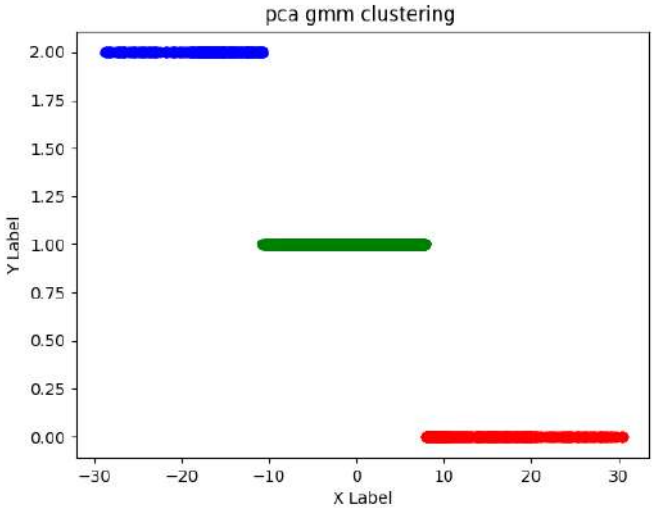


3 dimension into 1 dimension :

As we reduce the feature space ,we lose information in the data , it does contain more misscalssified points compare to pca (3-2)



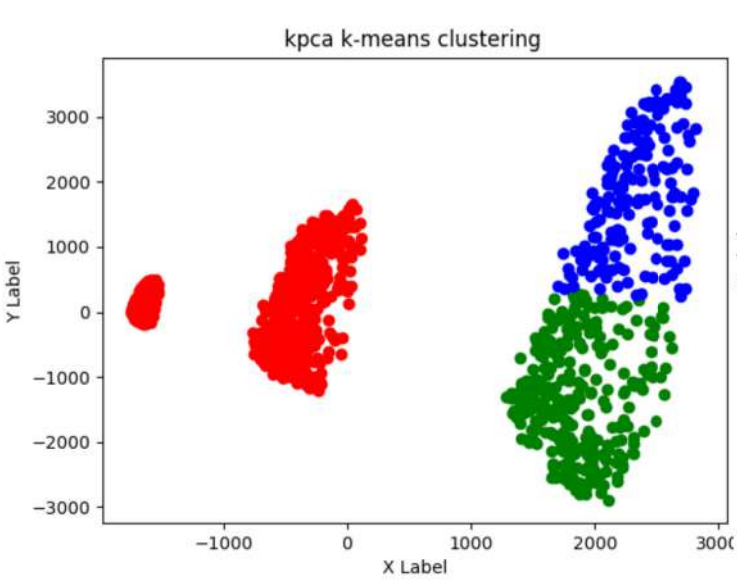


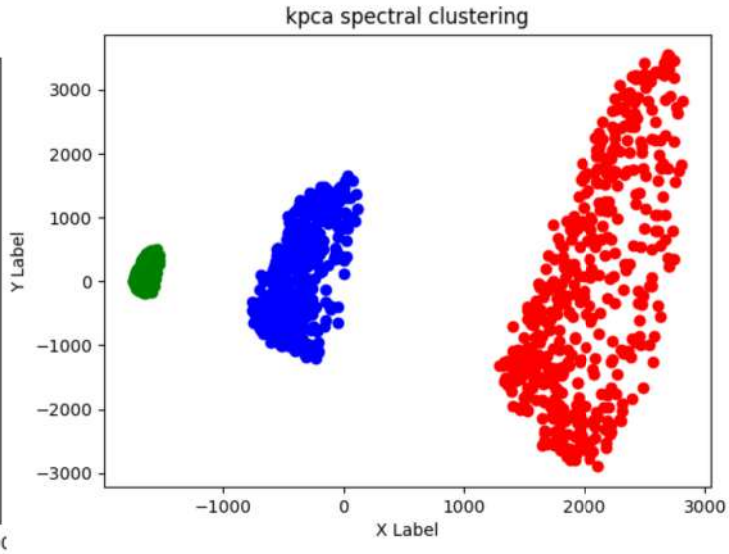


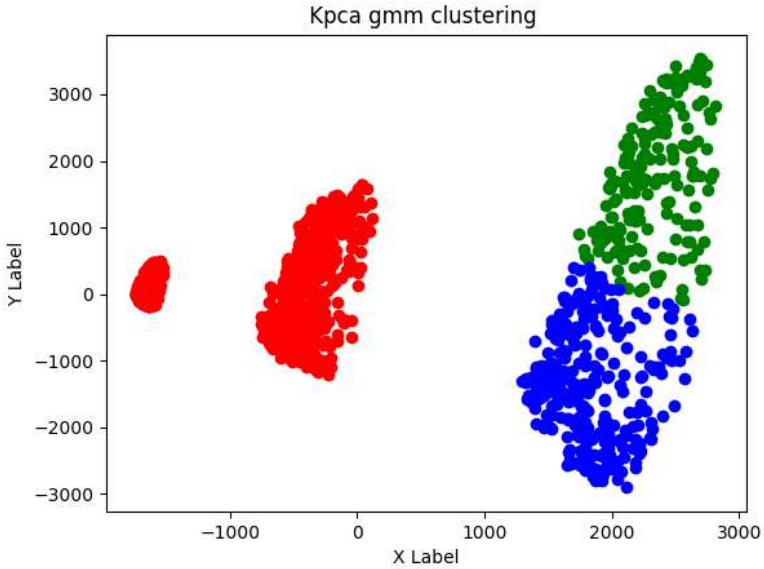
KPCA :

3 dimension to 2 dimension

since clustering is done with using kernel method so all models perform good except gmm , and it is able to recognise only some of them in the inner cluster .

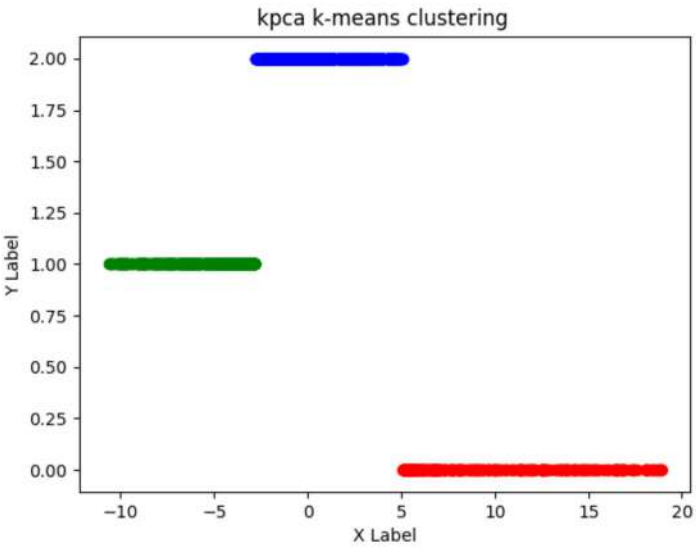


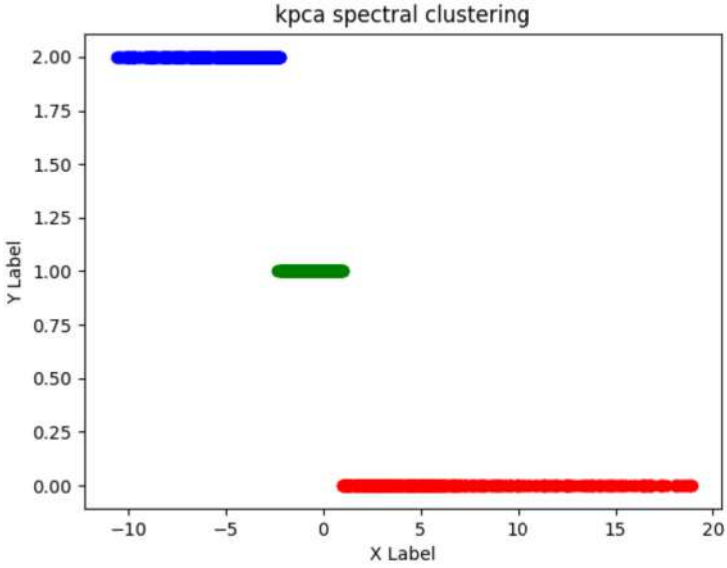


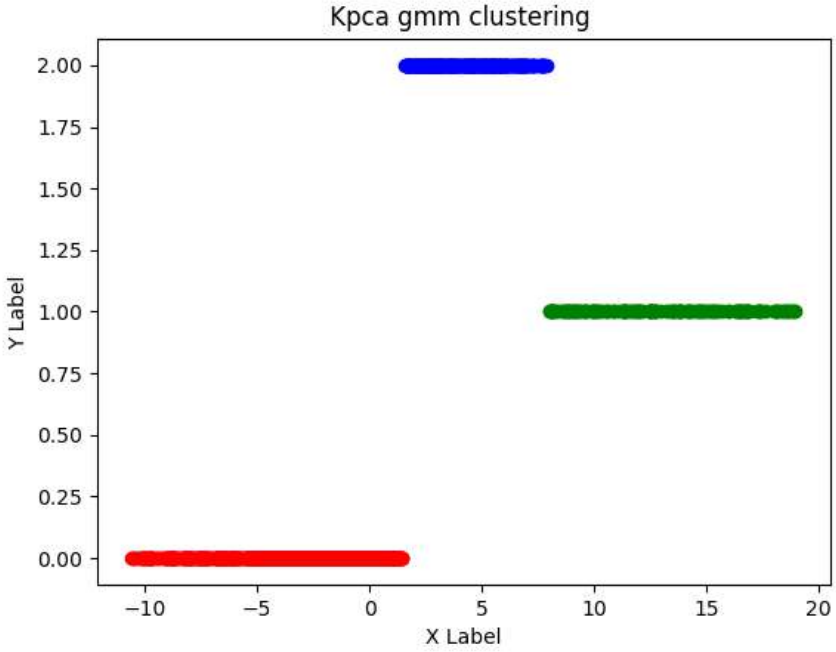


3 dimension to 1 dimension :

cluster will be same as above but there will be more error added because when we reduce the feature space , information in the data is lost and even after using kernel method that information can not be recovered .







PART 3 :

optimum depth ,number of trees is 10 , 19 respectively

the corresponding accuracy is 0.975235454

The indivisual decision tree which are in best random forest has the following accuracies on the same test dataset
[0.93571428571428572, 0.91428571428571426, 0.95357142857142863, 0.93928571428571428, 0.93571428571428572, 0.94999999999999996, 0.95714285714285718, 0.92142857142857137, 0.9107142857142857, 0.93571428571428572, 0.9285714285714286, 0.96071428571428574, 0.9285714285714286, 0.90357142857142858, 0.95714285714285718, 0.94285714285714284, 0.92142857142857137, 0.93571428571428572, 0.92500000000000004]

Conclusion :

The random forest performs with 97.5 % while its decision trees are performing less than this , none of the decision tree has performed better than the random forest .

This is because randomforest takes the decision based on the majority of the decision tree ,So if some of them does not predict the label correctly ,it does not matter because of the majority wins .