AN INTERNSHIP REPORT

ON

"CLASSIFICATION OF DATA BY SOME MODELS"

Submitted to

Dr. DILEEP A. D.

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

JUNE-JULY 2019

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER ENGINEERING 7TH SEM , 4th Yr.

Submitted by

MAYANK KUMAR

DEPARTMENT OF COMPUTER ENGINEERING
GOVT. ENGG. COLLEGE OF JHALAWAR
JHALAWAR, RAJASTHAN - 326001
2019-2020

Acknowledgements

We are profoundly grateful to **Dr. DILEEP A. D. and continuous encouragement throughout to see that this internship rights its target since its commencement to its completion.**

The internship opportunity I have with Indian Institute of Technology Mandi is a great chance for learning .

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped me directly or indirectly during this course of work.

MAYANK KUMAR

ABSTRACT

This internship report entitled to KNN Algorithm. The main objective of the study is to analyze the real time problem related to Data Science.

In this internship, I learned how to do work with numeric data and image data as well.

A K-Nearest Neighbor based classifier classifies a query instance based on the class labels of its neighbor instances. Although kNN has proved to be a ubiquitous classification/regression tool with good scalability, but it suffers from some drawbacks. Two of its major drawbacks are: (1) The existing kNN algorithm is equivalent to using only local prior probabilities to predict instance labels, and hence it does not take into account the class distribution around the wider neighborhood of the query instance, which results into undesirable performance on imbalanced data. (2) It uses all the training data at the runtime and hence is slow.

Bayesian Classification and decision making is based on probability and the principle of choosing the most probable or the lowest risk.

There are a variety of models and algorithms that solves classification problems. Among these models, Maximum Gaussian Mixture Model (MGMM) is a model we proposed earlier that describes data using the maximum value of Gaussians. Expectation Maximization (EM) algorithm can be used to solve this model. In this paper, we propose a multi-EM approach to solve MGMM and to train MGMM based classifiers. This approach combines multiple MGMMs solved by EM into a classifier. The classifiers trained with this approach on both artificial and real life datasets were tested to have good performance with 10-fold cross validation.

In this internship, I learned how to do work with numeric data and image data as well.

Contents

List of Figures

Chapter 1

CLASSIFICATION OF NON-SEPREBALE DATA

1.1 Problem Statement

We have given Non-Linear Seperable Dataset , in this dataset we have 2 classes and 2 fetures using knn alogrithm we have to build model for classification of this dataset.

After building this model, also calculate-
1-Accuracy of the model
2-Ploting of training data
3-Confusion matrix for best value of k
4-Precesion , Recall , F1-score
5-K vs Accuracy graph
6-Decision Bounding ploting for best value k
7-Mean error value vs k graph

1.2 Problem Description

Given 2D-Dataset

This figure contain head of 2D-data. In this 2D-data, we have 2 Classes(0,1) and 2 Features.

1.3 Model Evalution

1.3.1 Accuracy for best value of k

For given datas the best value of k is 23 and Accuracy on this k is 72 percentage.

1.3.2 Confusion Matrix for best value of K=23

Confusion Matrix		
PREDICTION/ACTUAL	CLASS 1	CLASS 2
CLASS 1	1182	573
CLASS 2	420	1250

1.3.3 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.74	0.67	0.70
CLASS 2	0.69	0.75	0.72

1.4 Graph

1.4.1 K vs Accuracy

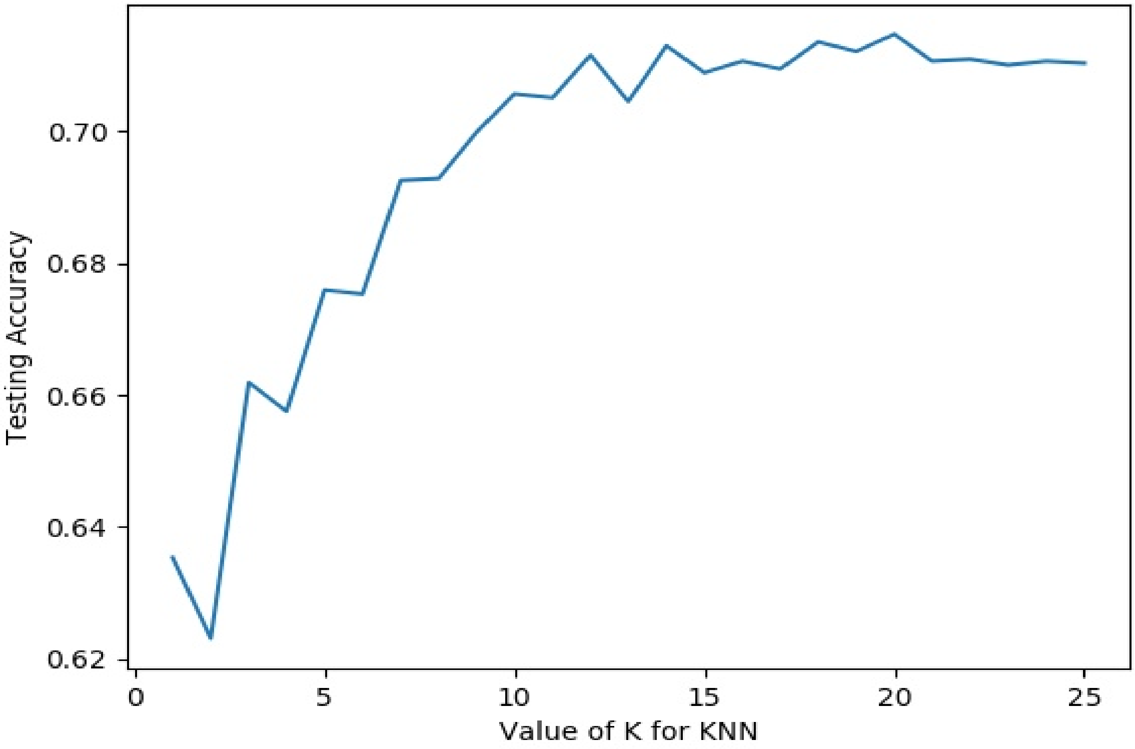

Figure 1.1: K VS Accuracy

1.4.2 K vs Mean error value

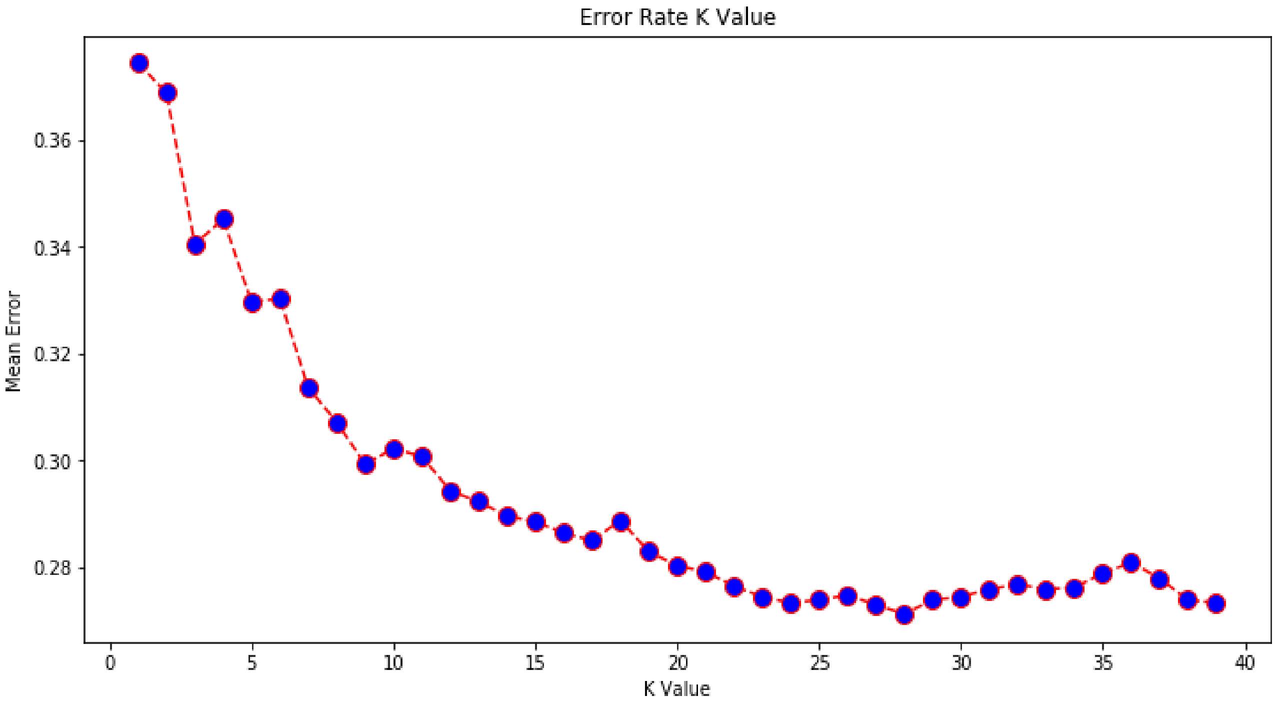

Figure 1.2: K VS mean

1.4.3 Data plotting

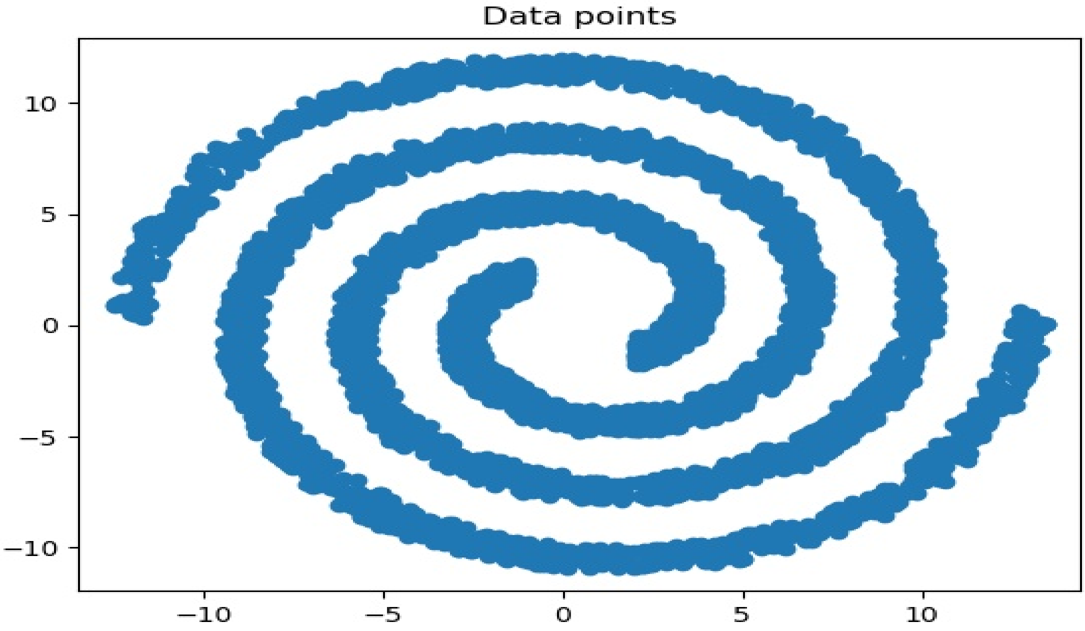

Figure 1.3: Data Plotting

1.4.4 Decision boundary

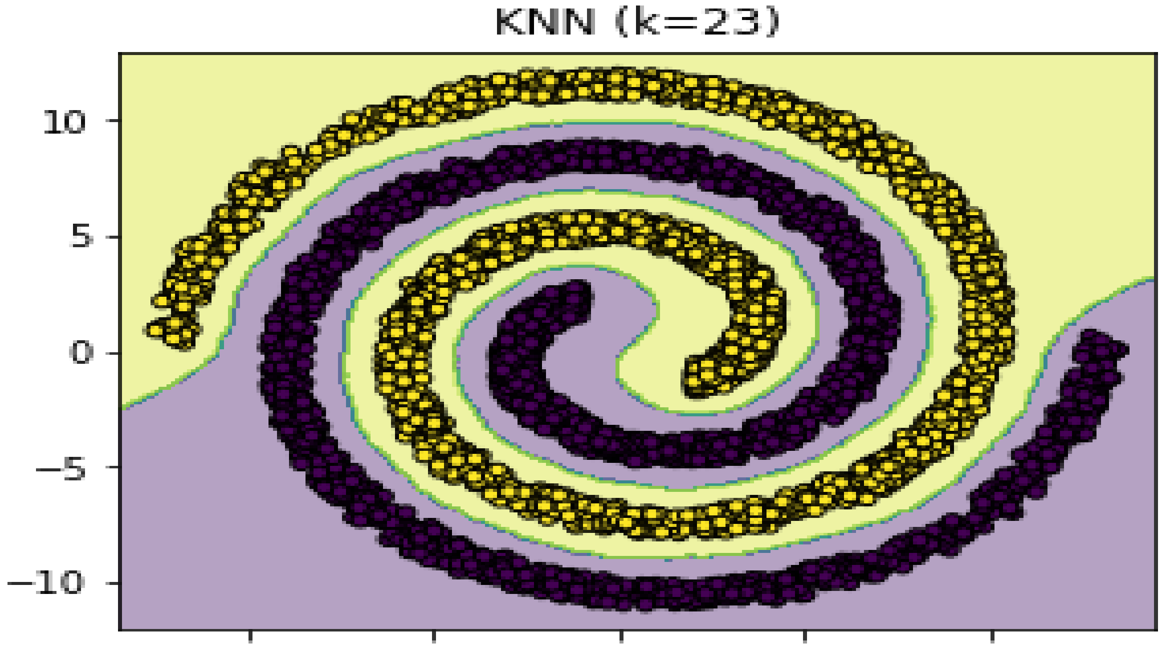

Figure 1.4: Decision Boundary

1.5 Euclidean distance use with mean

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

1.5.1 Model Evalution

1.5.2 Accuracy

Accuracy : 0.6013

1.5.3 Confusion Matrix

Confusion Matrix		
PREDICTION/ACTUAL	CLASS 1	CLASS 2
CLASS 1	141	88
CLASS 2	95	135

1.5.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.60	0.62	0.61
CLASS 2	0.60	0.59	0.59

1.5.5 Decision boundary

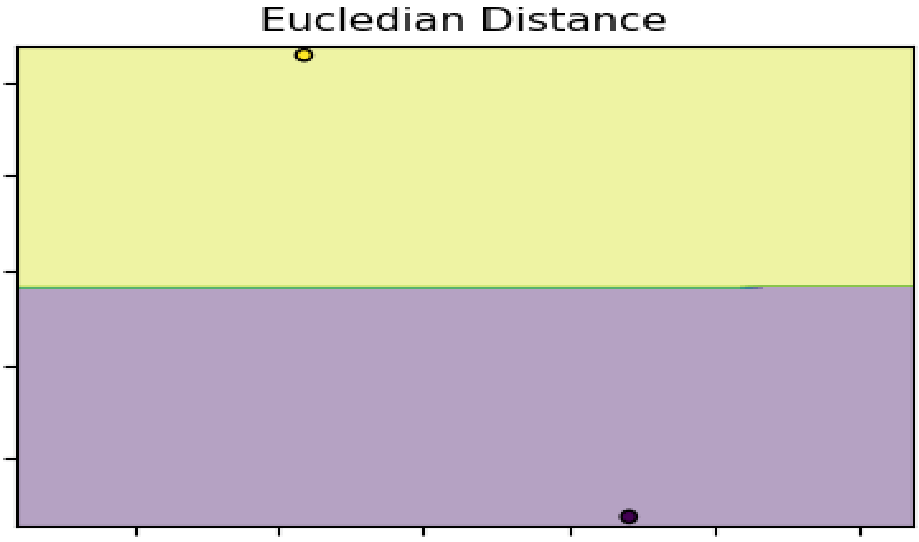

Figure 1.5: Decision Boundary

1.6 Using Variance for classification

After building this model, also calculate-

1-Accuracy of the model

2-Confusion matrix for best value of k

3-Precesion , Recall , F1-score

1.6.1 Model Evalution

1.6.2 Accuracy

Accuracy : 0.6013

1.6.3 Confusion Matrix

Confusion Matrix		
PREDICTION/ACTUAL	CLASS 1	CLASS 2
CLASS 1	141	88
CLASS 2	95	135

1.6.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.60	0.62	0.61
CLASS 2	0.60	0.59	0.59

1.6.5 Decision boundary

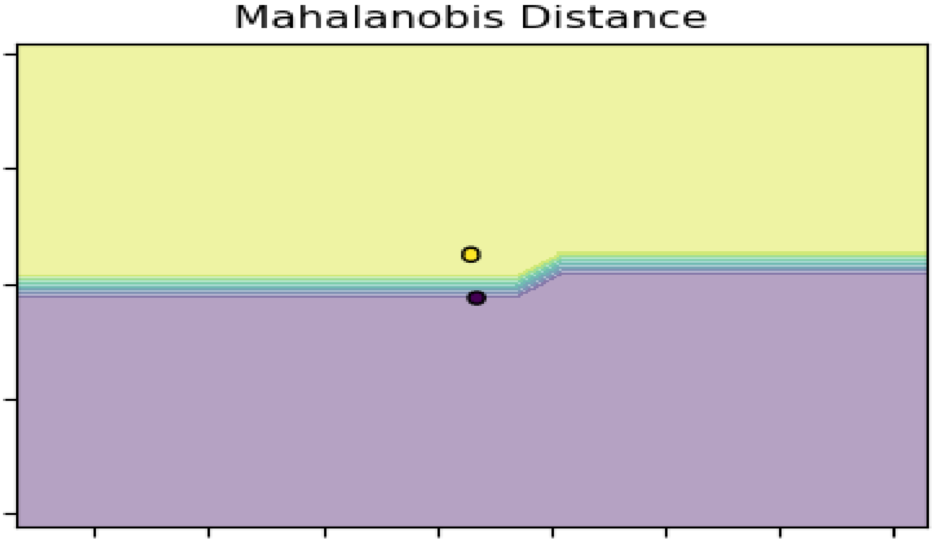

Figure 1.6: Decision Boundary

1.7 Using BayesClassifier For Classification

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

1.7.1 Model Evalution

1.7.2 Accuracy

Accuracy : 0.629

1.7.3 Confusion Matrix

Confusion Matrix		
PREDICTION/ACTUAL	CLASS 1	CLASS 2
CLASS 1	467	254
CLASS 2	282	465

1.7.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.62	0.65	0.64
CLASS 2	0.65	0.62	0.63

1.8 Using GMM For Classification

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

1.8.1 Model Evalution

1.8.2 Accuracy

Accuracy : 0.100

1.8.3 Confusion Matrix

Confusion Matrix		
PREDICTION/ACTUAL	CLASS 1	CLASS 2
CLASS 1	242	0
CLASS 2	0	230

1.8.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	1.0	1.0	1.0
CLASS 2	1.0	1.0	1.0

1.8.5 Clustering of data from Gmm

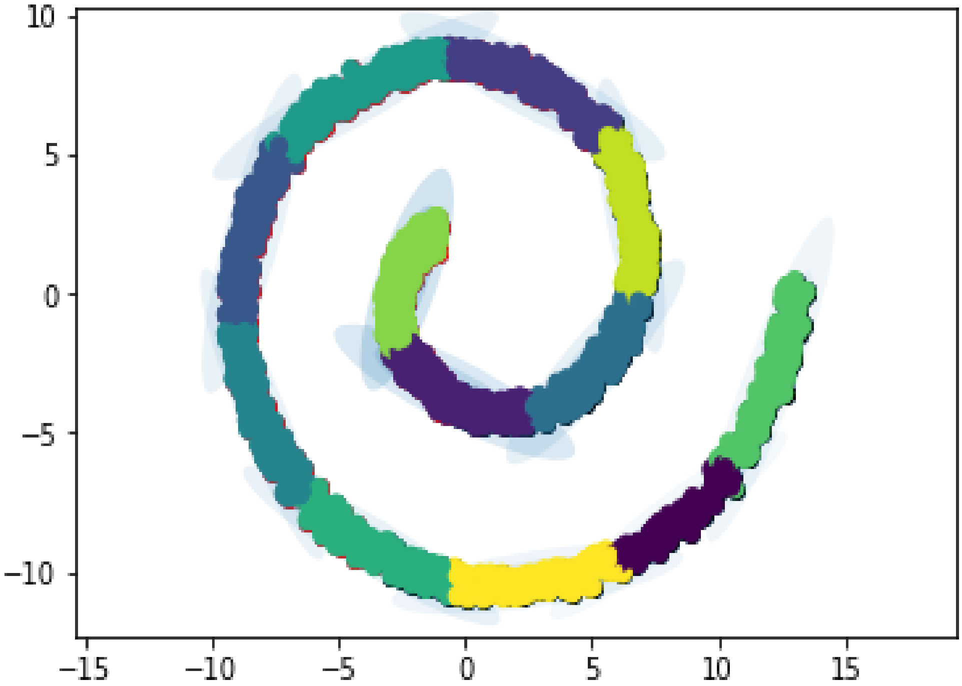

Figure 1.7: Cluster for class1 data

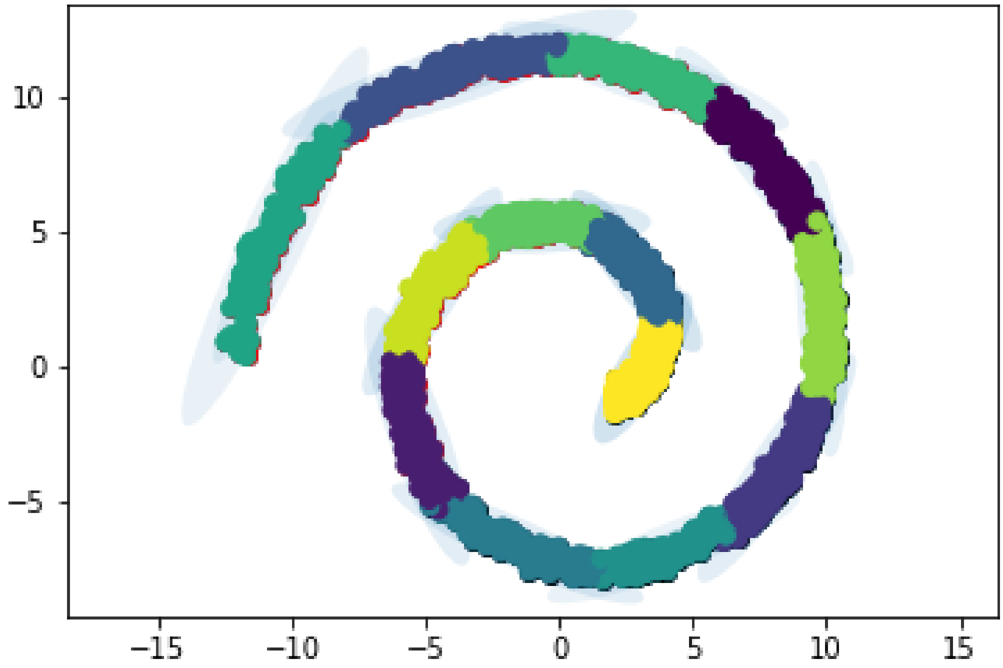

Figure 1.8: Cluster for class2 data

1.8.6 Decision boundary

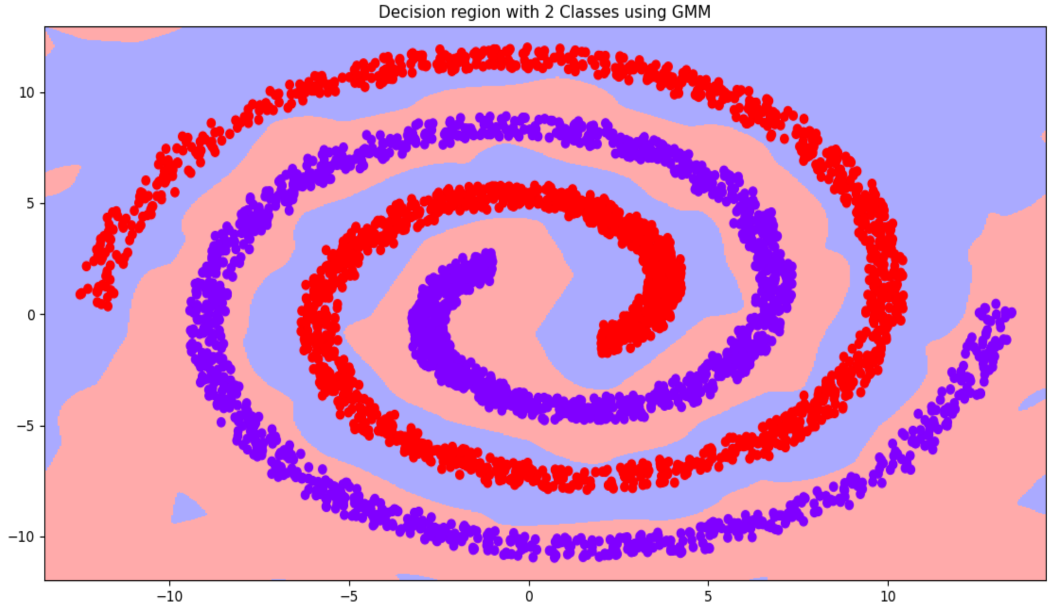

Figure 1.9: Decision Boundary

1.8.7 Density Ploting

density.png

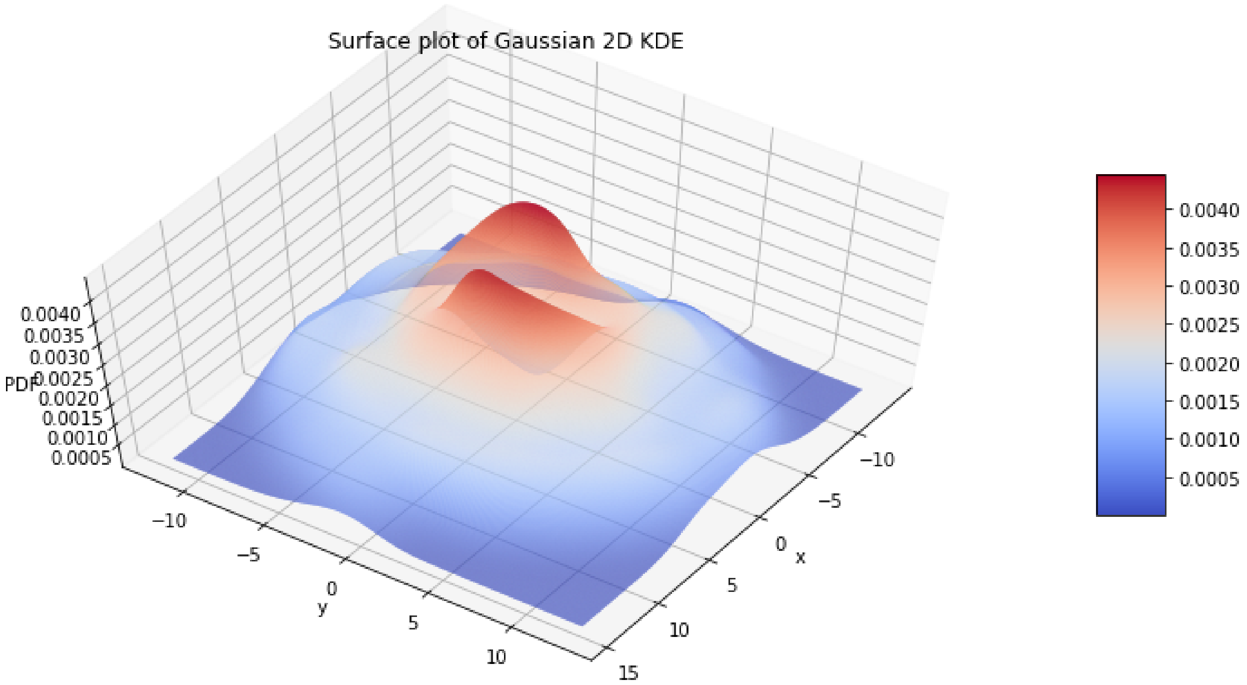

Figure 1.10: Density Plotting

1.9 Conclusion

We perform KNN on Nonseperable datset The best value of k for this model is 23 and accuracy is 72%

We perform first order statistics on nonsep dataset accuracy for that is 60.13%

We perform second order statistics nonsep dataset so the accuracy for that is 60.13%

We perform BayesClassifier on nonseperable datset so the accuracy is 62.9%

We perform GMMClassifier on nonseperable data so the accuracy for that is 100%

Chapter 2

CLASSIFICATION OF IMAGE DATA

2.1 Problem Statement

We have given IMAGE dataset, and in this dataset we have 3 classes only. So for this dataset using knn algorithm we have to build model for classification of this dataset.

After building this model, also calculate-
1-Features of image data
2-Accuracy of the model
3-Ploting of training data
4-Confusion matrix for best value of k
5-Precesion , Recall , F1-score
6-K vs Accuracy graph
7-Mean error value vs k graph

2.2 Problem Description

2.2.1 Feautre vector of image data

In this problem statement we have data on the form of images of 3 different classes.
class1-bayou
class2-desertvegetation
class3-musicstore
All three classes contain 50-50 data images.so now we create model for extract feautres of this images. and all the images have 1*24 dimension features are ex-

tract , so from this data we can calculate all above details.

2.3 Model Evalution

2.3.1 Accuracy for best k

for given data accuracy max for k = 5.

Accuracy : 0.6533

2.3.2 Confusion Matrix

Confusion Matrix			
PRED./ACTUAL	BAYOU	DESERT VEGETATION	MUSIC STORE
BAYOU	32	10	7
DESERT VEGET.	17	29	4
MUSIC STORE	11	4	35

2.3.3 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
BAYOU	0.53	0.65	0.59
DESERT VEGET.	0.67	0.58	0.62
MUSIC STORE	0.76	0.70	0.73

2.4 Graph

2.4.1 K VS Accuracy

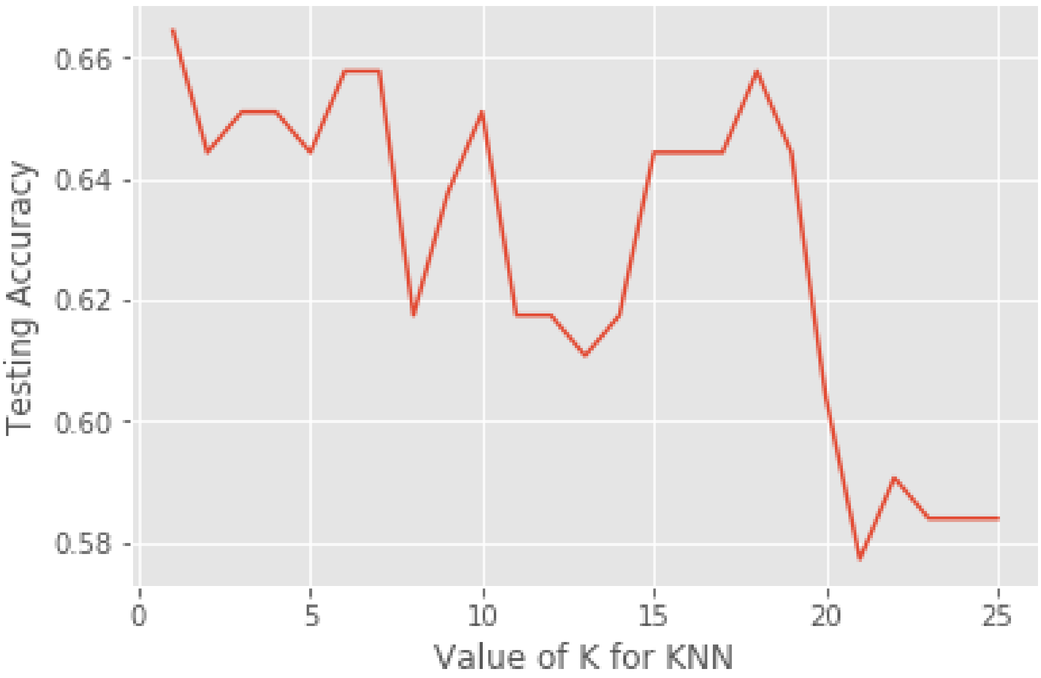

Figure 2.1: K VS Accuracy

2.4.2 K VS Mean error

errror vs k image.jpg

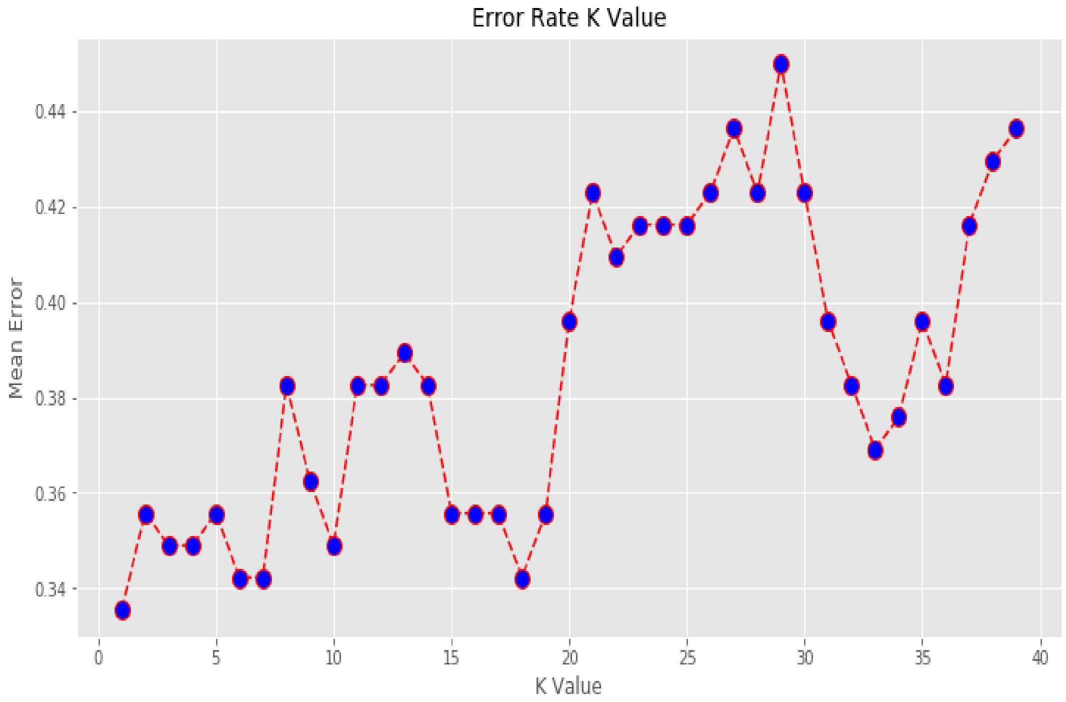

Figure 2.2: K VS Mean error

2.5 Euclidean distance use with mean

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

2.5.1 Model Evalution

2.5.2 Accuracy

Accuracy : 0.4333

2.5.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	BAYOU	DESERT VEGETATION	MUSIC STORE
BAYOU	22	14	14
DESERT VEGET.	9	8	33
MUSIC STORE	9	6	35

2.5.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
BAYOU	0.55	0.44	0.49
DESERT VEGET.	0.29	0.16	0.21
MUSIC STORE	0.55	0.44	0.49

2.6 Using Variance for classification

After building this model, also calculate-

1-Accuracy of the model

2-Confusion matrix for best value of k

3-Precesion , Recall , F1-score

2.6.1 Model Evalution

2.6.2 Accuracy

Accuracy : 0.46666

2.6.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	BAYOU	DESERT VEGETATION	MUSIC STORE
BAYOU	23	12	15
DESERT VEGET.	8	10	32
MUSIC STORE	3	10	37

2.6.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
BAYOU	0.68	0.46	0.55
DESERT VEGET.	0.31	0.20	0.24
MUSIC STORE	0.44	0.74	0.55

2.7 Using BayesClassifier for Classification

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

2.7.1 Model Evalution

2.7.2 Accuracy

Accuracy : 0.363

2.7.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	BAYOU	DESERT VEGETATION	MUSIC STORE
BAYOU	11	20	19
DESERT VEGET.	8	9	33
MUSIC STORE	6	9	35

2.7.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
BAYOU	0.44	0.22	0.29
DESERT VEGET.	0.24	0.18	0.20
MUSIC STORE	0.40	0.70	0.51

2.8 Using GMM for Classification with PCA

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

2.8.1 Model Evalution

2.8.2 Accuracy

Accuracy : 0.406

2.8.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	BAYOU	DESERT VEGETATION	MUSIC STORE
BAYOU	27	21	2
DESERT VEGET.	14	33	3
MUSIC STORE	25	24	1

2.8.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
BAYOU	0.41	0.54	0.47
DESERT VEGET.	0.42	0.66	0.52
MUSIC STORE	0.17	0.02	0.04

2.9 Conclusion

We perform KNN on image datset also extract features (i.e. color bins) for images. The best value of k for this model is.5 and accuracy is 68%

We perform first order statistics on features of images so the accuracy for that is 43%

We perform second order statistics on features of images so the accuracy for that is 46%

We perform BayesClassifier on IMAGE datset and accuracy is 36.3%

We perform GMMClassifier on Seperable data so the accuracy for that is 37.3%

Chapter 3

CLASSIFICATION OF SEPREBALE DATA

3.1 Problem Statement

We have given 3 Classes dataset, and using knn algorithm and we have to build model for classification of this dataset.

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score
4-K vs Accuracy graph
5-Mean error value vs k graph
6-Ploting of training data
7-plotting boundary region

3.2 Model Evalution

3.2.1 Accuracy

Accuracy : 0.905

3.2.2 Confusion Matrix

Confusion Matrix			
PRED./ACTUAL	CLASS 1	CLASS 2	CLASS 3
CLASS 1	273	73	0
CLASS 2	26	317	0
CLASS 3	0	0	361

3.2.3 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.91	0.75	0.85
CLASS 2	0.81	0.92	0.86
CLASS 3	1.00	1.00	1.00

3.3 Graph

3.3.1 Decision Region

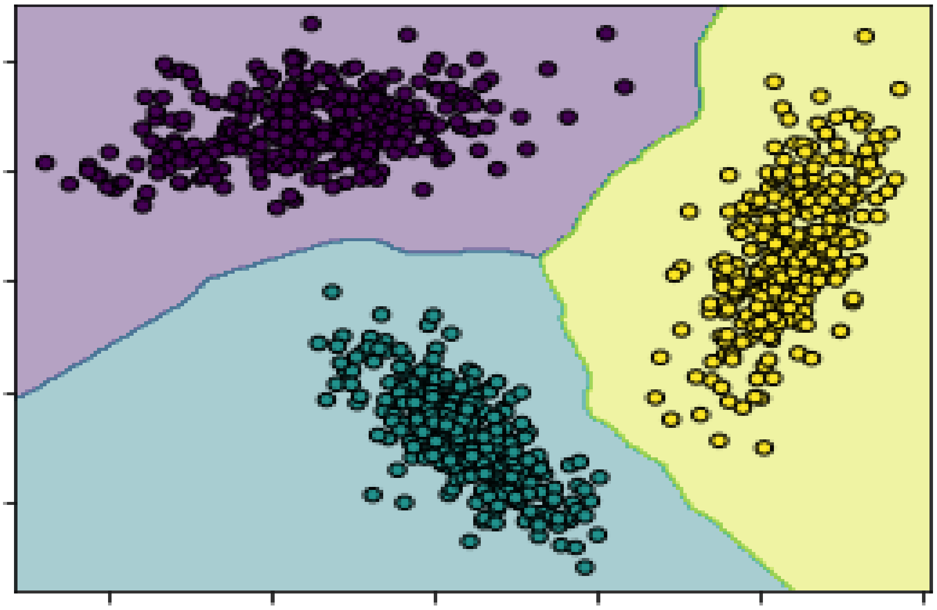

Figure 3.1: Decision Region

3.3.2 Data Plot

plotting.jpg

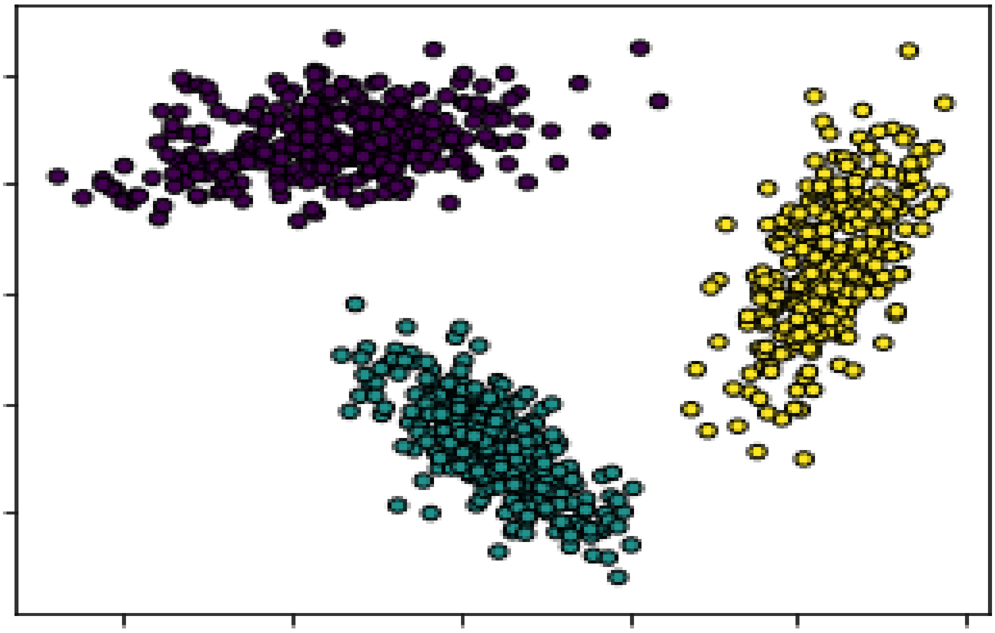

Figure 3.2: Data plotting

3.4 Euclidean distance use with mean

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

3.4.1 Model Evalution

3.4.2 Accuracy

Accuracy : 0.987

3.4.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	CLASS 1	CLASS 2	CLASS 3
CLASS 1	194	0	1
CLASS 2	0	211	0
CLASS 3	0	7	207

3.4.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	1.00	0.99	1.00
CLASS 2	0.97	1.00	0.98
CLASS 3	1.00	0.97	0.98

3.4.5 Decision boundary

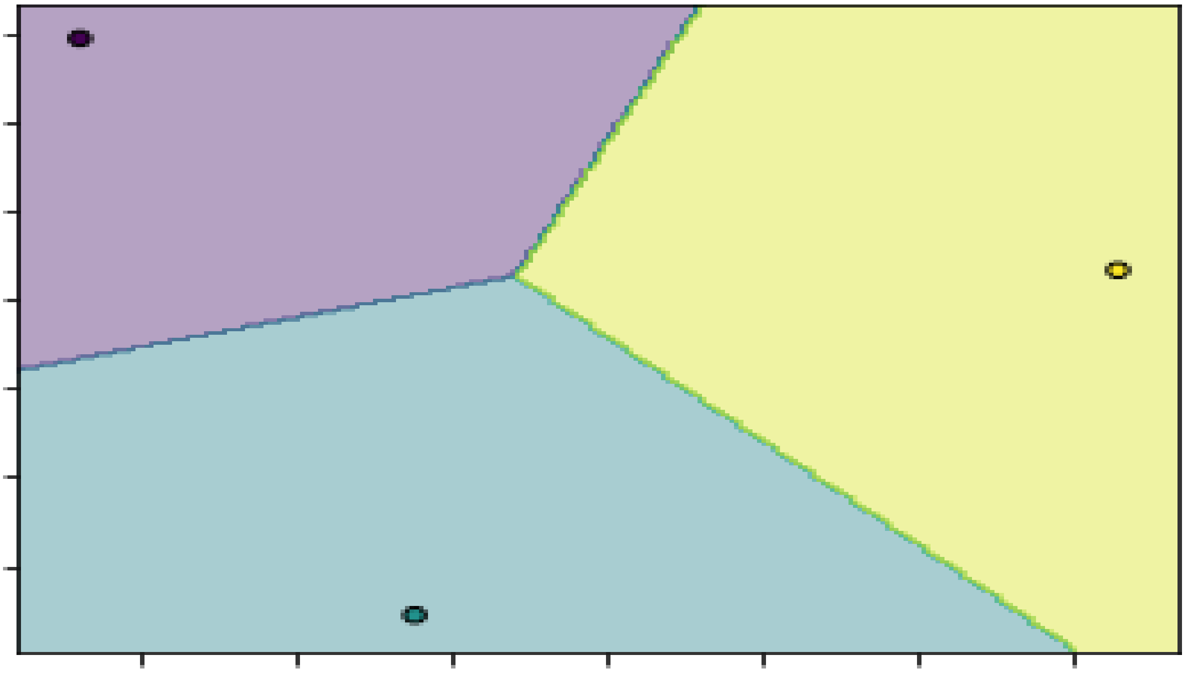

Figure 3.3: Decision Boundary

3.5 Using Variance for classification

After building this model, also calculate-
1-Accuracy of the model

2-Confusion matrix for best value of k

3-Precesion , Recall , F1-score

3.5.1 Model Evalution

3.5.2 Accuracy

Accuracy : 0.985

3.5.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	CLASS 1	CLASS 2	CLASS 3
CLASS 1	195	0	0
CLASS 2	6	205	0
CLASS 3	0	3	211

3.5.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	0.97	1.00	0.98
CLASS 2	0.99	0.97	0.98
CLASS 3	1.00	0.99	0.99

3.5.5 Decision boundary

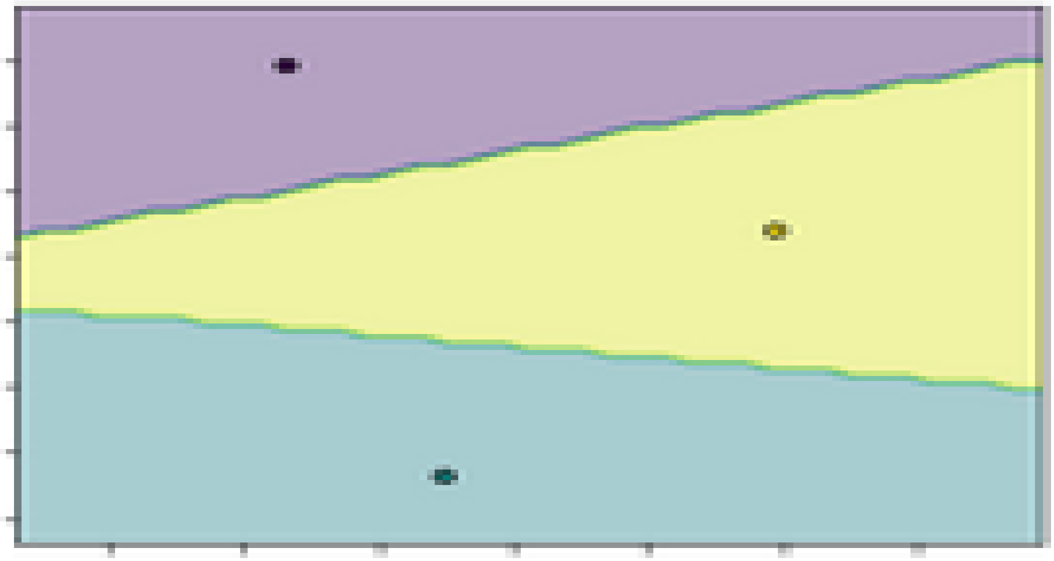

Figure 3.4: Decision Boundary

3.6 Using BayesClassifier for Classification

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

3.6.1 Model Evalution

3.6.2 Accuracy

Accuracy : 0.100

3.6.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	CLASS 1	CLASS 2	CLASS 3
CLASS 1	167	0	0
CLASS 2	0	141	0
CLASS 3	0	7	142

3.6.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	1.00	1.00	1.00
CLASS 2	1.00	1.00	1.00
CLASS 3	1.00	1.00	1.00

3.6.5 Decision boundary

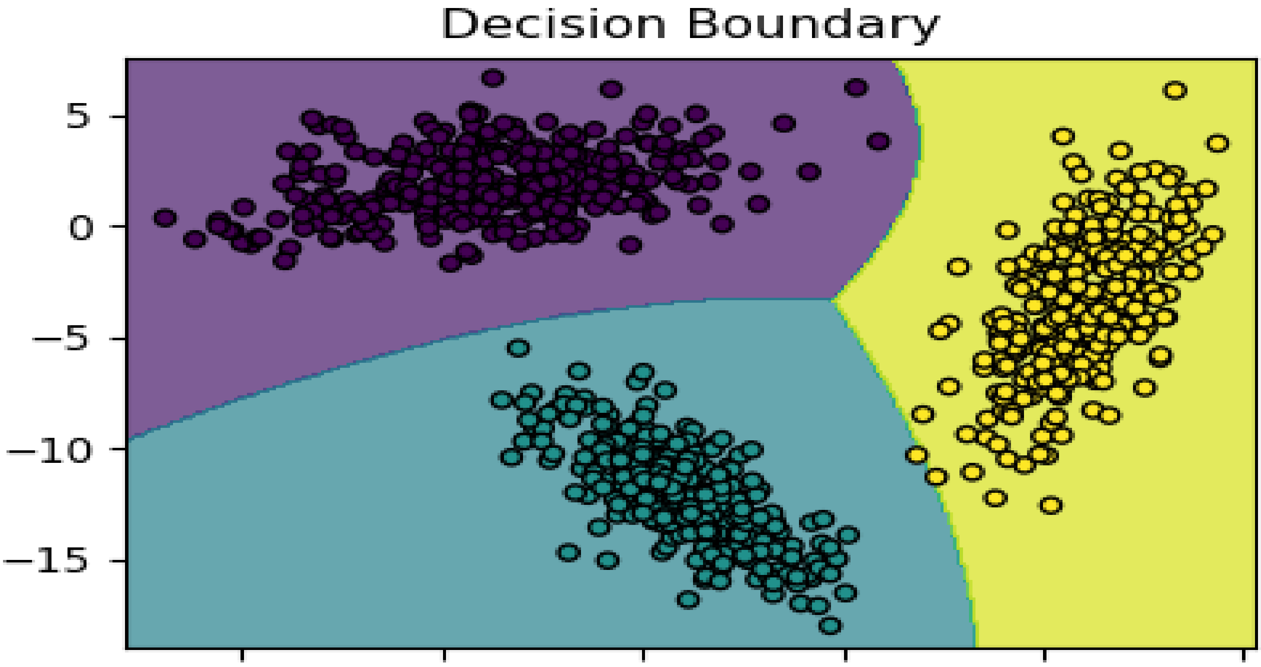

Figure 3.5: Decision Boundary

3.7 Using GMM For Classification

After building this model, also calculate-
1-Accuracy of the model
2-Confusion matrix for best value of k
3-Precesion , Recall , F1-score

3.7.1 Model Evalution

3.7.2 Accuracy

Accuracy : 0.100

3.7.3 Confusion Matrix

Confusion Matrix			
PREDICTION/ACTUAL	CLASS 1	CLASS 2	CLASS 3
CLASS 1	195	0	0
CLASS 2	0	211	0
CLASS 3	0	0	286

3.7.4 Precision, Recall, F1-score

P,R,H			
	PRECISION	RECALL	F1-SCORE
CLASS 1	1.00	1.00	1.00
CLASS 2	1.00	1.00	1.00
CLASS 3	1.00	1.00	1.00

3.7.5 Clustering of data from Gmm

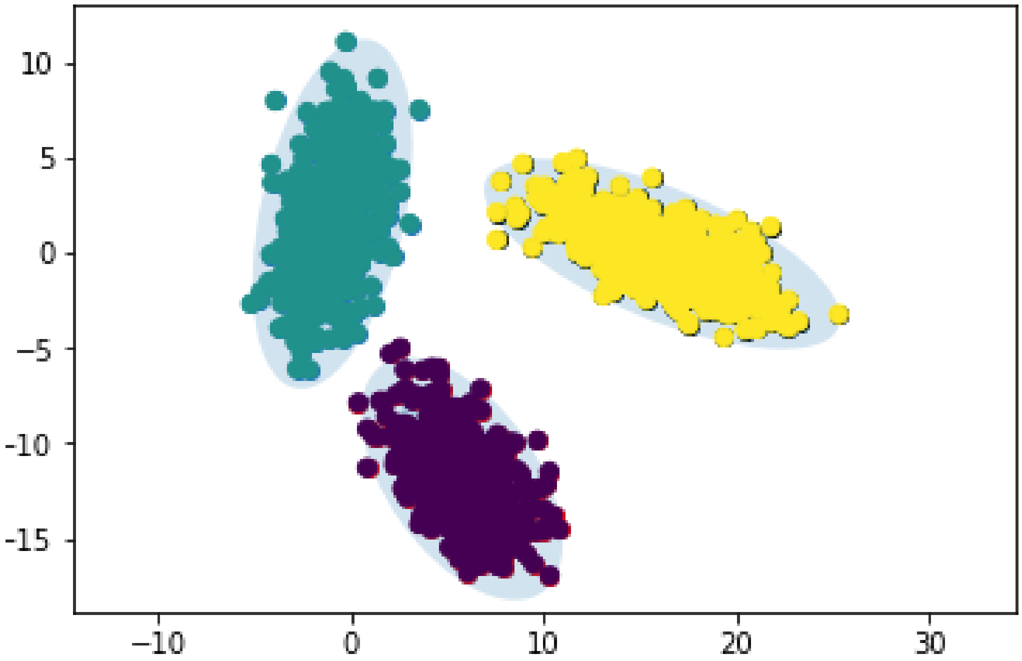

Figure 3.6: Clusters Of Data

3.7.6 Decision Boundary

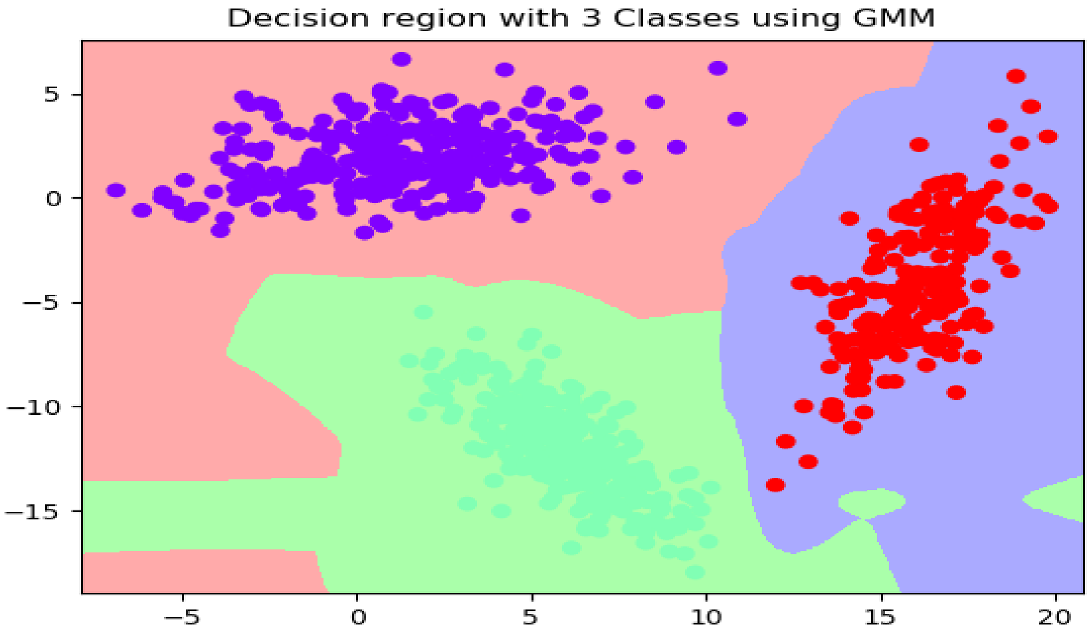

Figure 3.7: Decision Boundary

3.7.7 Density Plotting

plotsep.png

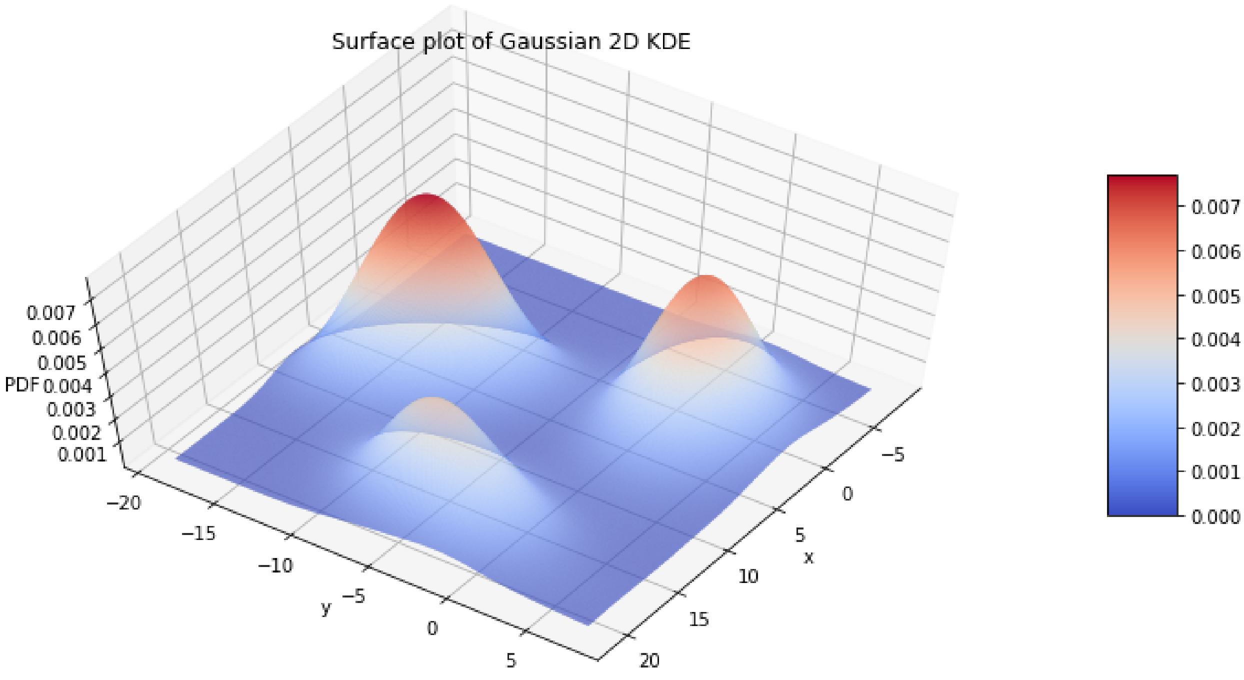

Figure 3.8: Density Plotting

3.8 Conclusion

We perform KNN on seperable datset and accuracy is 90%

We perform first order statistics on Seperable data so the accuracy for that is 98%

We perform second order statistics on Seperable data so the accuracy for that is 98%

We perform BayesClassifier on seperable datset and accuracy is 100%

We perform GMMClassifier on Seperable data so the accuracy for that is 100%

Chapter 4

COMPARISON BETWEEN GIVEN MODELS

4.1 Comparison of Accuracy

ACCURACY OF DATA			
CLASSIFIER	NONSEPERABLE	IMAGE	SEPERABLE
Knn	72%	68%	90%
Euclidean mean	60.3%	43%	98%
Mahalanobis	60.13%	46%	98%
Bayes Unimodel	62.9%	36.3%	100%
Bayes Multimodel	100%	-	100%
GMM with PCA	-	40.66%	-

4.2 Comparision of Boundary Region

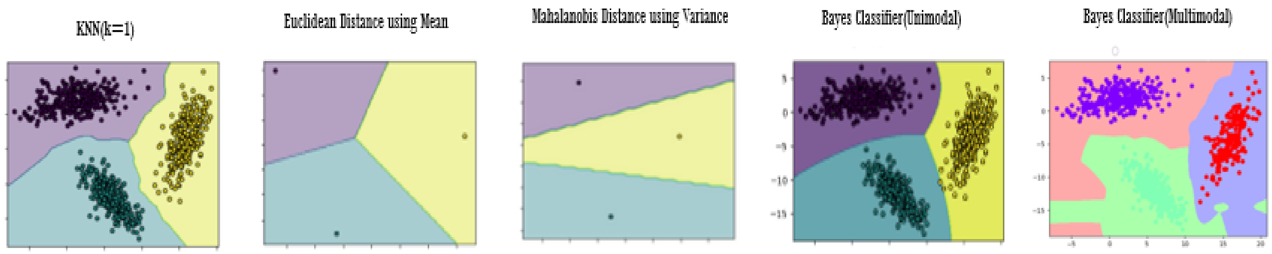

Figure 4.1: Seperable Data

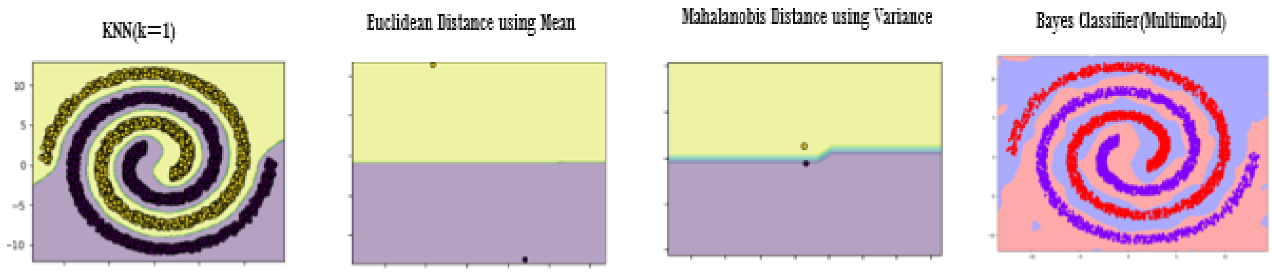

Figure 4.2: Non Seperable Data