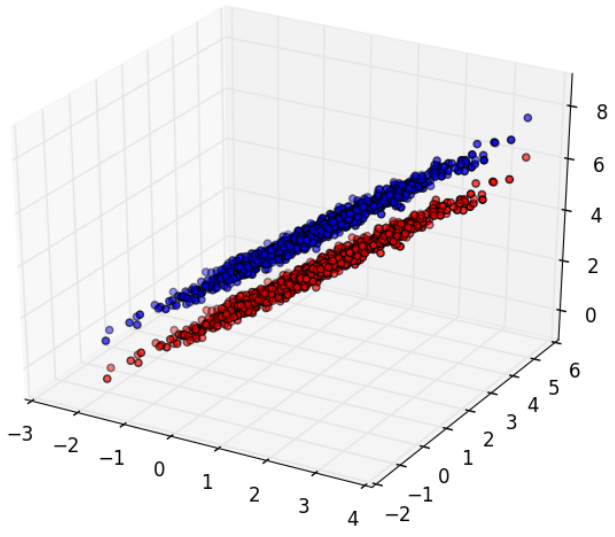
CS4011-Assignment 2

S Aakash CS15B060

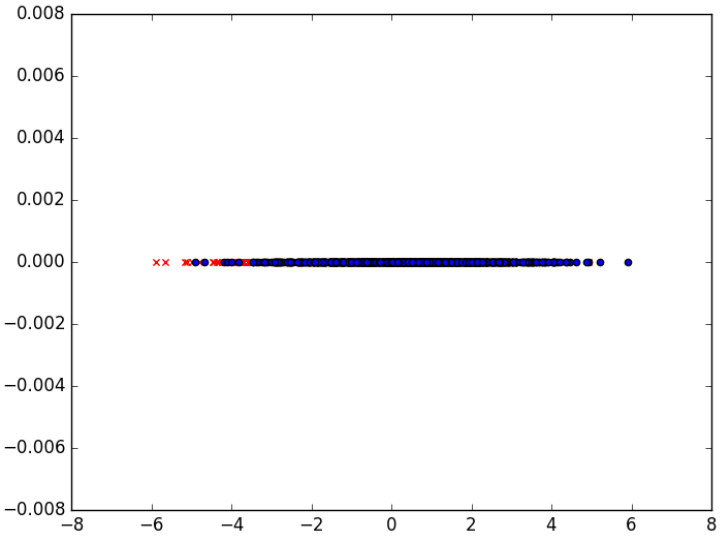
October 10, 2017

0.1 Question 1

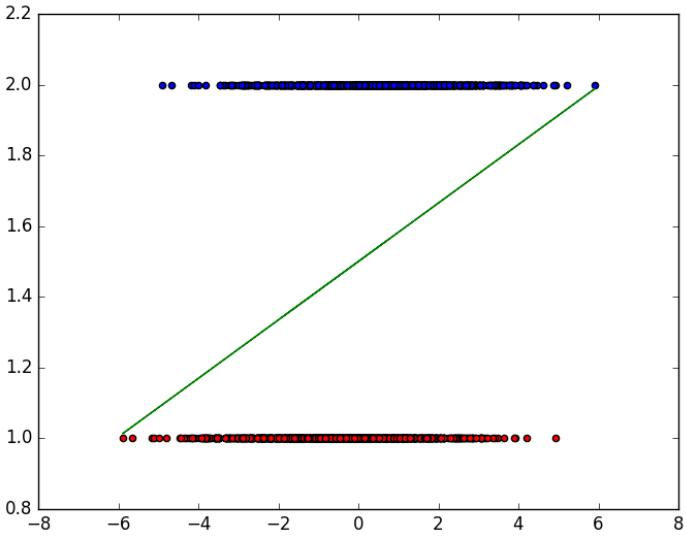
Assume Red-Class 1 and Blue-Class2
First we visualise the given Dataset(Figure 1).



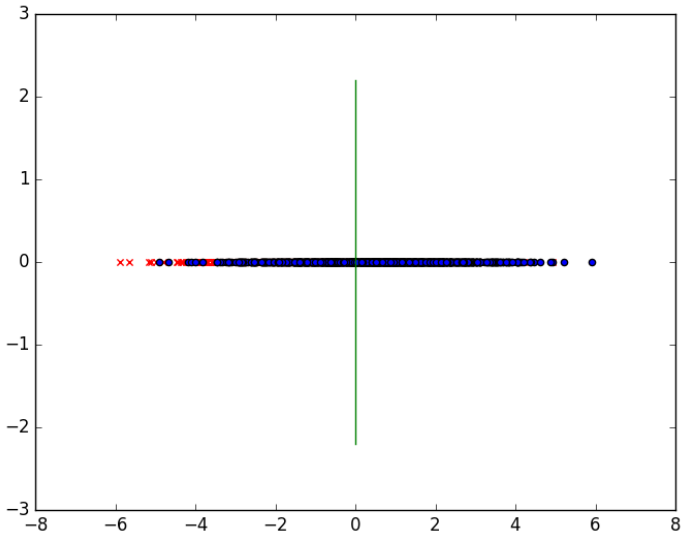
So Principal Component Analysis finds direction(s) in which there is high variance and minimum covariance,between components.Hence we find the directions corresponding to highest eigen value(in this case we find only a single direction). After visualising the data,we can inherently see that there is a direction along which all the data points are aligned(ie maximum variance).Hence considering this direction alone and neglecting other directions seems to make sense.



Now we want transform the data points in original space to the new dimensional space generated by PCA(In this question,it is a single direction).And we perform linear regression on this space.Following is the regression line.(Equation of regression line obtained is y=0.0826x+1.5)



Now,we can get the seperating line as (x=0) in this case,if we fix the threshold as 1.5(ie $\geq$1.5=>class=2).Figure 4 shows the decision boundry



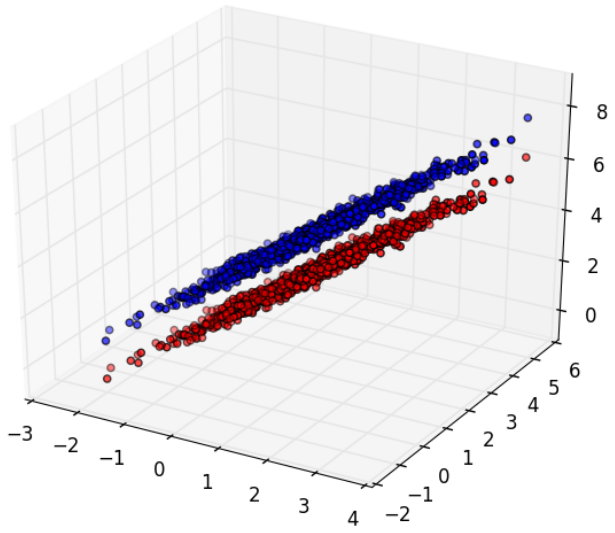
Now,we evaluate on test set,the regression line learnt and the following result is obtained:

Class	Precision	Recall	F1-Score	Support
1	0.61	0.61	0.61	200
2	0.61	0.61	0.61	200
avg/total	0.61	0.61	0.61	400

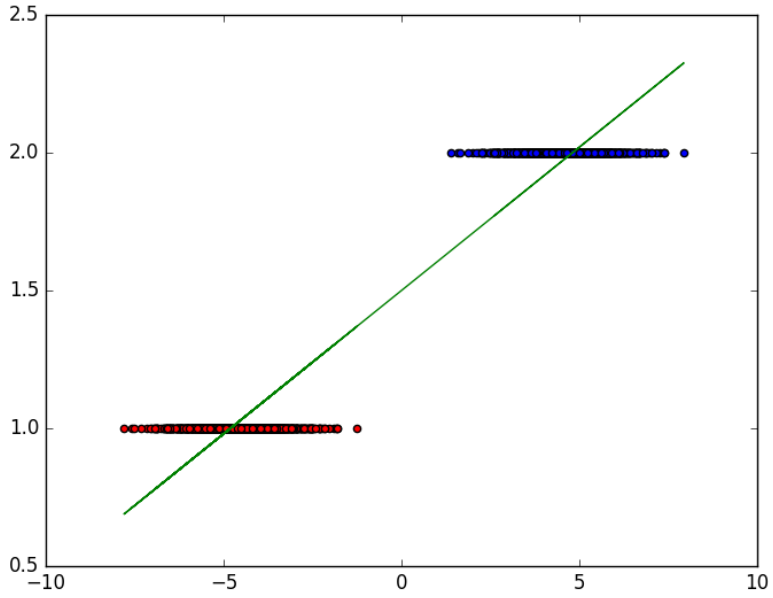
Clearly it is only slightly better than the randomised way of determining the class in 2-class classification problem.The reason comes from the Figure 3-Where we can observe a lot of overlap between the two classes in the projected axis.The reason for this is that,PCA doesnt take into account the value corresponding to datapoint(ie the y coordinate.).All it looks for is a direction with maximum variance and minimum covariance.So,there is a possibility for two different class points to overlap,in which case decision boundary may be poor.

0.2 Question 2

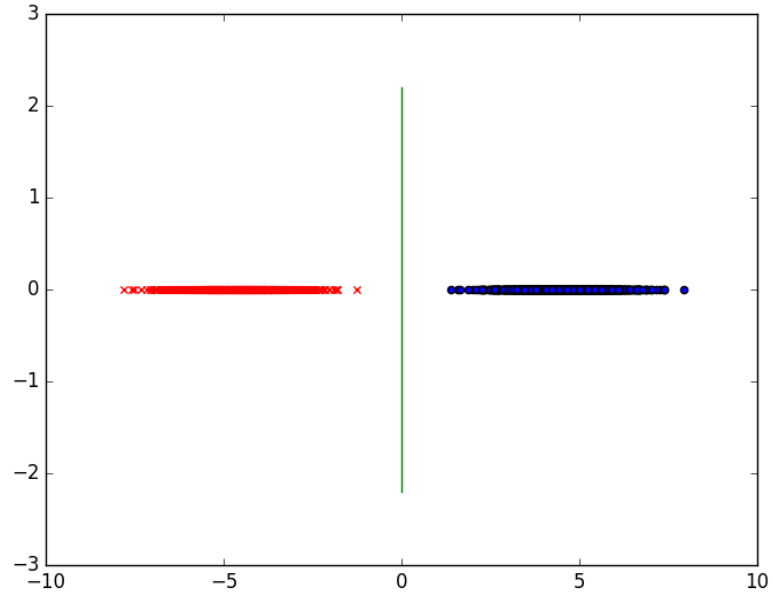
First we visualise the given Dataset(Figure 1).



So Linear Discriminant Analysis(as feature selection technique) finds direction(s) that maximise the between class variance and minimise within class variance,ie maximise Fischer ratio.We also know the equation of line(we derived in class,assuming class conditioned probablity to be guassian).We use inbuilt sklearn function to compute,then transform dataset to new dimension and we perform regression,like we did in first question.Equation of line is (Equation of regression line obtained is y=0.1016x+1.5)



Considering the threshold as 1.5,we again get decision boundary as x=0



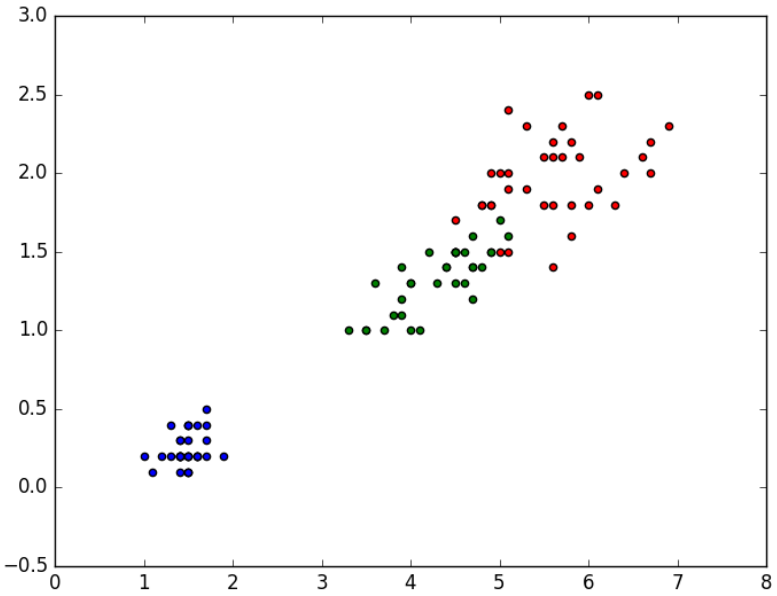
The following is the performance on test set.

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	200
2	1.00	1.00	1.00	200
avg/total	1.00	1.00	1.00	400

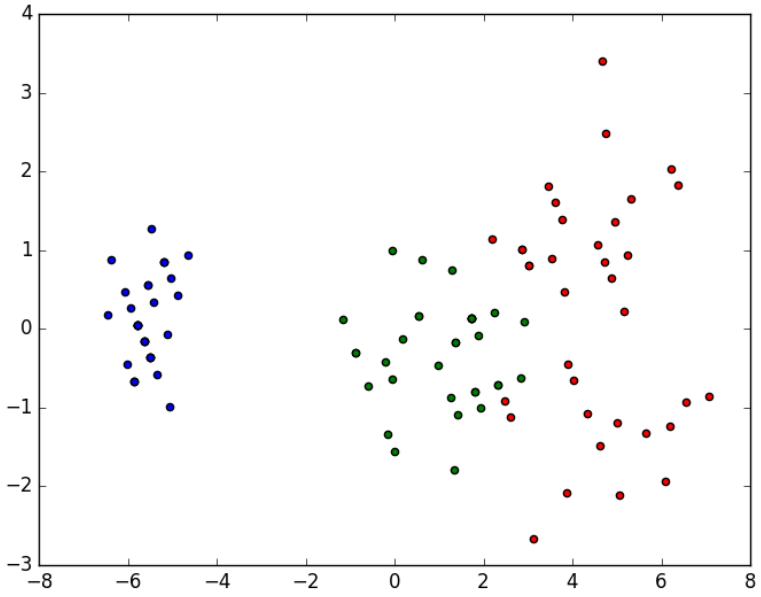
Note that because LDA takes class labels into account,we get a very good precision and recall,but PCA fails to do that and gives us low accuracy.So hence LDA is a better feature extraction mechanism as PCA generates lot of overlappings of disjoint class label points.

0.3 Question 3

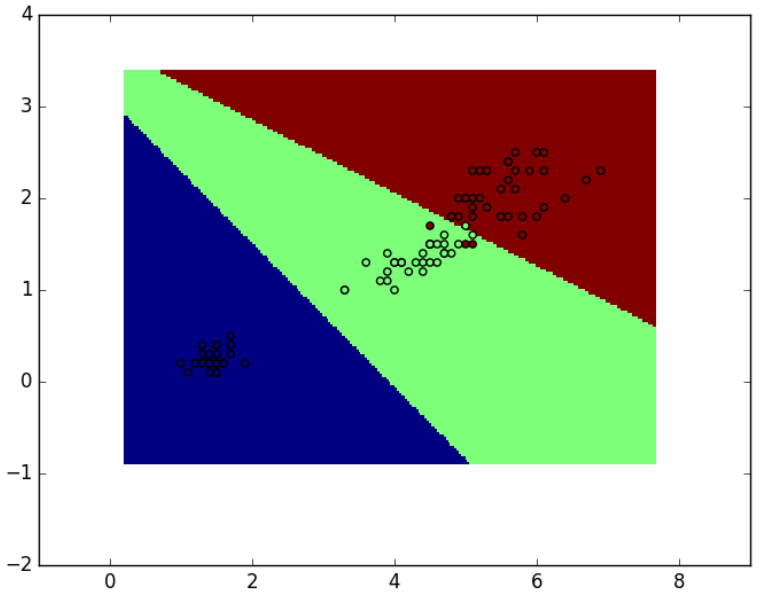
Let us first visualise the original data-set.Note that Blue-Class1(Iris Sethosa),Green-Class2(Iris Versicolour),Red-Class3(Iris Virginca)



Clearly blue class can be linearly seperated from other two,but red and green seems to be not linearly seperable.Now lets perform LDA inorder to obtain two new sets of features,using sklearn function and we will project the data points into a this dimension.



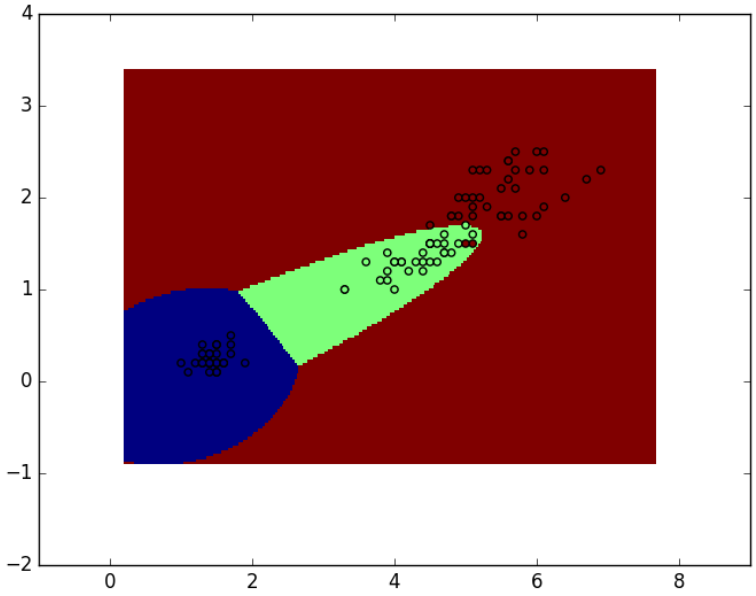
Now,Following is the decision boundry obtained on LDA(Original Space).



Now we analyse performance of LDA on test set.

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	15
2	0.94	0.94	0.94	15
3	0.92	0.92	0.92	15
avg/total	0.97	0.97	0.97	45

B)Now,in QDA we no longer take pooled variances of two classes(ie we no-longer make assumption that the two classes have same covariance matrix) and hence the quadratic terms donot cancel out(the equation sir derived in class,if the two classes have different covariances,we qet QDA).So using this equation,we can plot QDA decision boundry.

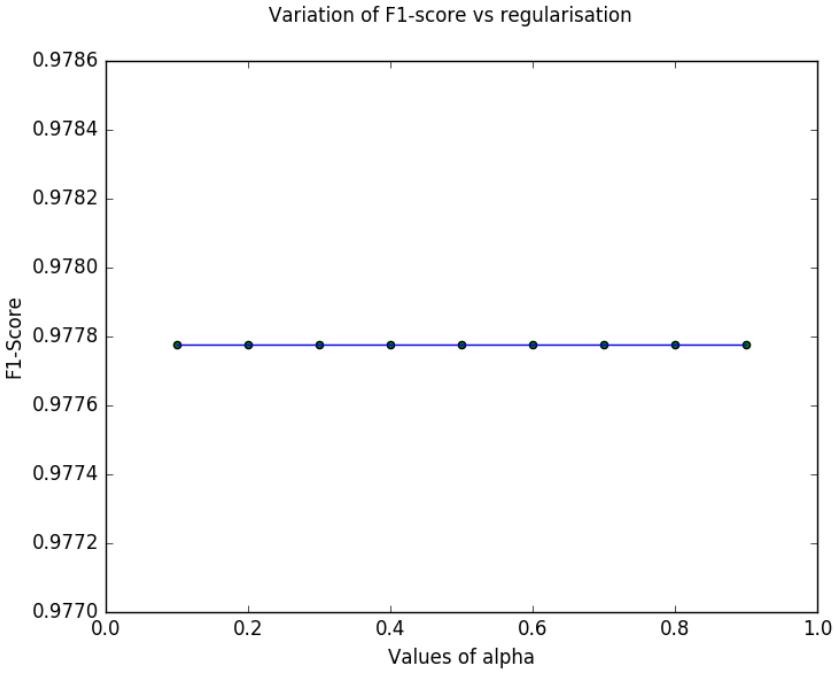


Now we analyse performance of QDA on test set.

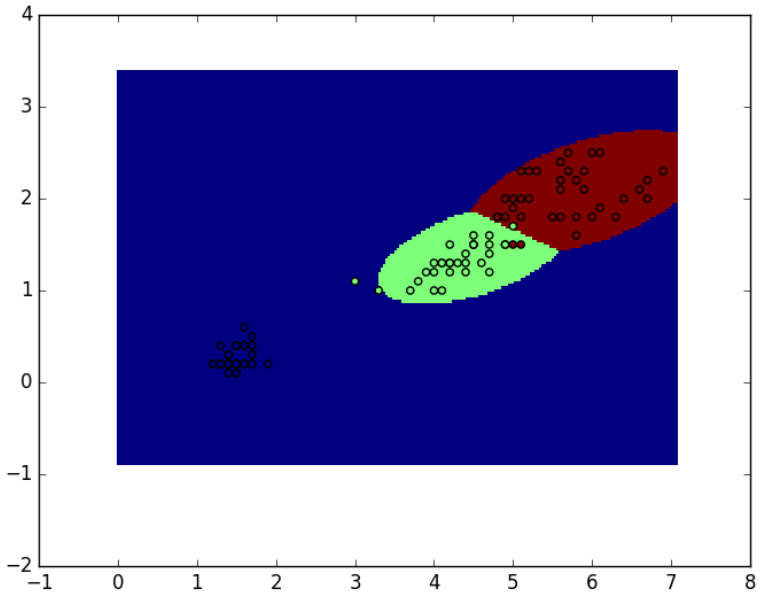
Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	15
2	1.00	0.94	0.97	15
3	0.93	1.00	0.96	15
avg/total	0.98	0.98	0.98	45

Hence LDA and QDA perform quite good equally on this Dataset.
Now we gote RDA(Regularised discriminant Analysis),where we have a regularisation term α present.So essentially,the problem of LDA is that because we are

making assumption that every class has same covariance matrix,and because we approximate it to Pooled factor,there is a chance of high bias,as we are assuming the behaviour of class conditioned probablity to be same for all classes,which may not be true.Also QDA,results in high variance when the training set is so small that computing per class covariance involves only a few datapoints.As a thumb rule if no of datapoints is a linear order of number of features QDA leads to high variance.
In order to balance both,we have RDA.Here we make weights to pooled covariance and perclass covariance.And new class covariance is computed as:
Σ_{kRDA}=α.Σ_{pooled}+(1-α.)Σ_k
Now Σ_{pooled} and Σ_k follow the same emperical estimation technique,as for LDA,QDA.
Clearly α=1 =>LDA,α=0 =>QDA,
Intutively RDA can be thought as weighted mean of LDA,QDA and Discriminant score corresponding to each data point takes similar form compared to QDA.
δ_k(x)=-log $|\Sigma_{kRDA}|$+(x-μ_k)^t.Σ^{-1}_{kRDA}.(x-μ_k)+log π_k Hence We can now compute the discriminant scores classwise per data point and classify the point to the one resulting in highest score. Various values of α were tested.Performance was found to be almost similar,as LDA and QDA both perform equally well in this case.



Following is the Decision boundry for RDA.



Now lets take α=0.7 and evaluate on test set.Following is the result obtained.

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	15
2	0.94	0.94	0.94	15
3	0.92	0.92	0.92	15
avg/total	0.97	0.97	0.97	45

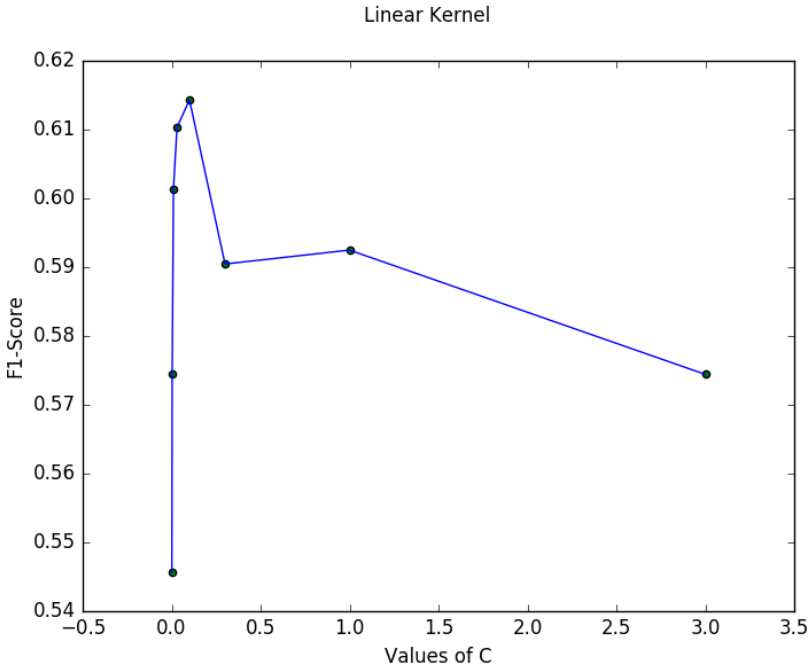
Since Both LDA and QDA,interesting provides a good result,RDA,also provides a good result.

0.4 Question 4

K-fold Cross Validation.Value of k=5.Assume Class1,2,3,4-coast,forest,insidecity,mountain
(1)Linear Kernel:Parameters under consideration:C So here,C was varied between 0.001,0.003.....1,3 and the F-score for each classifier was obtained using K-fold cross validation and optimal C was obtained.

C	Avg. F1 Score(over 5 folds)
0.003	0.54561344
0.001	0.57457035
0.01	0.60132214
0.03	0.61031303
0.1	0.61431776
0.3	0.59046627
1	0.59248617
3	0.57442109

Lets visualise this by plotting the graph of variation of F1-score vs C.

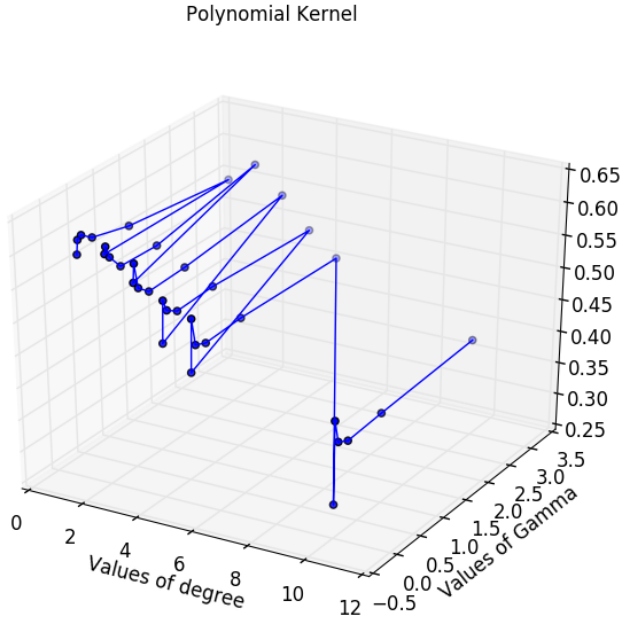


Clearly,Best value of C: 0.1 and its corresponding f1-score=0.614317758934.
Now,we will Look at the performance of this model at test set.

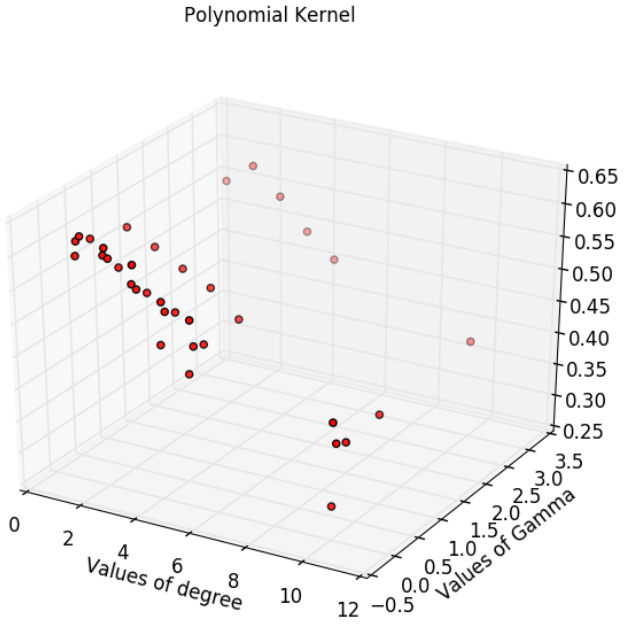
Class	Precision	Recall	F1-Score	Support
0	0.53	0.50	0.51	20
1	0.76	0.65	0.70	20
2	0.56	0.70	0.62	20
3	0.58	0.55	0.56	20
avg/total	0.61	0.60	0.60	80

So,essentially we used Crossvalidation to choose the better classifier.Number of Support vectors [217 103 169 214]

(2)Polynomial Kernel:
Parameters under consideration:C,gamma,coeff0,degree.
Ofcourse,in my code,I have taken all these criterion into account and have done a generalised validation,but for sake of visualisation we will fix C and coeff0 to default values and visualise the variation of F-scores with respect to degree and gamma values,which are currently important for us.
let gamma $\in \{0.01,0.03,0.1,0.3,1,3\}$ and degree $\in \{1,2,3,4,5,10\}$
Now K-fold cross validation is performed for each possible combination of gamma and degree and following result is obtained.



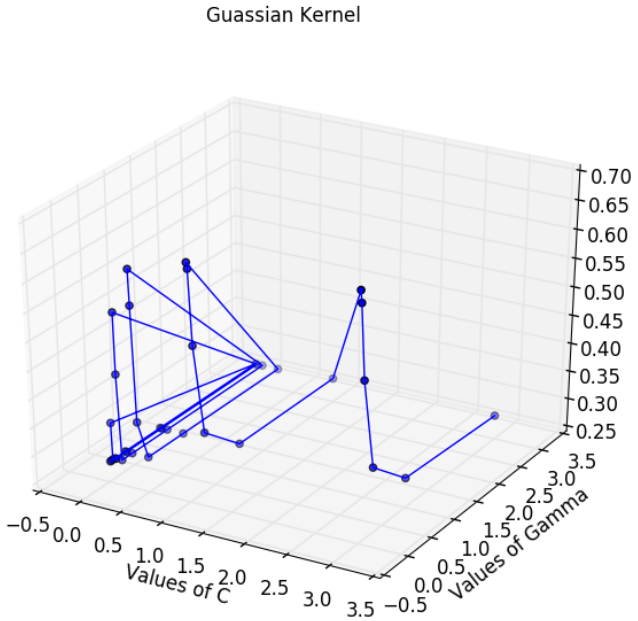
Degree	Gamma	avg-F1-Score
1	0.01	0.58646602
1	0.03	0.60743707
1	0.1	0.61137298
1	0.3	0.5985214
1	1	0.58341587
1	3	0.56377882
2	0.01	0.59640224
2	0.03	0.61646145
2	0.1	0.58745642
2	0.3	0.56461459
2	1	0.56262961
2	3	0.59542662
3	0.01	0.56280902
3	0.03	0.59047135
3	0.1	0.55071827
3	0.3	0.53585246
3	1	0.53867334
3	3	0.55653438
4	0.01	0.48102386
4	0.03	0.54463394
4	0.1	0.526692
4	0.3	0.51581195
4	1	0.51875181
4	3	0.51088072
5	0.01	0.44642054
5	0.03	0.5268426
5	0.1	0.48402878
5	0.3	0.47694435
5	1	0.48009898
5	3	0.47601425
6	0.01	0.2941958
6	0.03	0.42236024
6	0.1	0.38663234
6	0.3	0.37777667
6	1	0.38267239
6	3	0.39351781

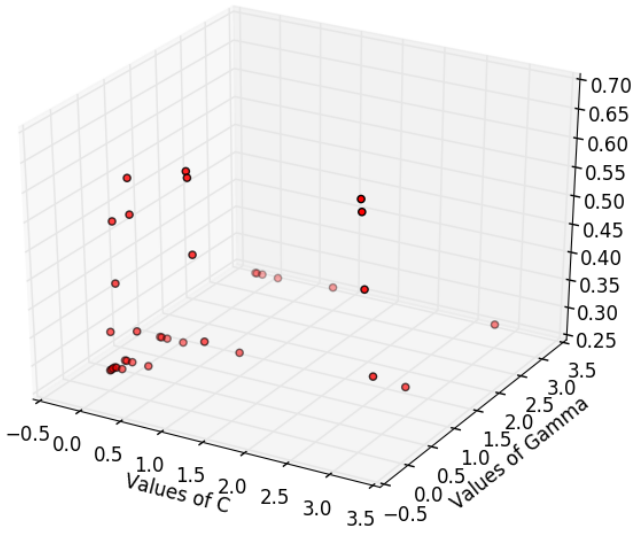


Best value of degree,gamma: 2 ,0.03 respectively and the corresponding avg.F1-score on k-fold cross validation 0.61646145.
The Performance of this model on test data:

Class	Precision	Recall	F1-Score	Support
0	0.53	0.50	0.51	20
1	0.76	0.65	0.70	20
2	0.56	0.70	0.62	20
3	0.58	0.55	0.56	20
avg/total	0.61	0.60	0.60	80

.Number of support vectors [201 135 192 233] (3)Guassian Kernel:
Parameters under consideration:C,gamma,coeff0.
Ofcourse,in my code,I have taken all these criterion into account and have done a generalised validation,but for sake of visualisation we will fix coeff0 to default values and visualise the variation of F-scores with respect to C and gamma values,which are currently important for us.
let gamma $\in \{0.01,0.03,0.1,0.3,1,3\}$ and C $\in \{0.01,0.03,0.1,0.3,1,3\}$
Now K-fold cross validation is performed for each possible combination of gamma and degree and following result is obtained.


Guassian Kernel
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Values of C
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Values of Gamma

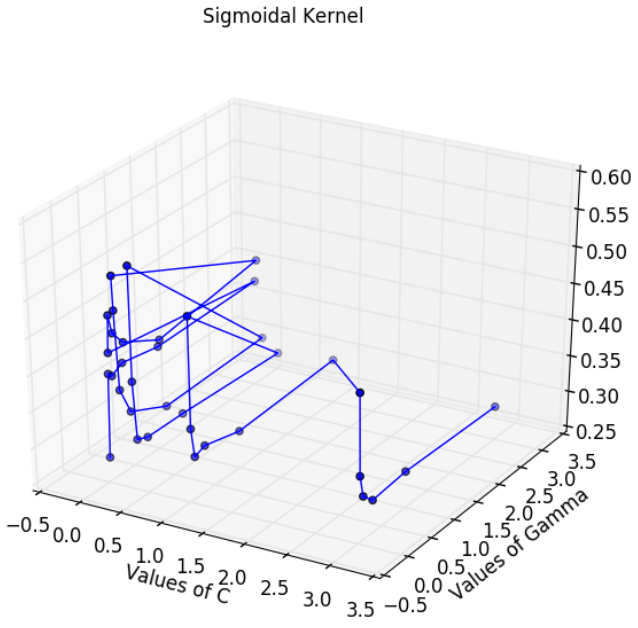

Guassian Kernel
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Values of C
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Values of Gamma

C	Gamma	avg-F1-Score
0.01	0.01	0.27733441
0.01	0.03	0.27733441
0.01	0.1	0.27733441
0.01	0.3	0.27733441
0.01	1	0.27733441
0.01	3	0.27733441
0.03	0.01	0.3469139
0.03	0.03	0.28030471
0.03	0.1	0.27733441
0.03	0.3	0.27733441
0.03	1	0.27733441
0.03	3	0.27733441
0.1	0.01	0.54183246
0.1	0.03	0.43347
0.1	0.1	0.27733441
0.1	0.3	0.27733441
0.1	1	0.27733441
0.1	3	0.27733441
0.3	0.01	0.62127293
0.3	0.03	0.55855429
0.3	0.1	0.35187484
0.3	0.3	0.27733441
0.3	1	0.27733441
0.3	3	0.27733441
1	0.01	0.65401576
1	0.03	0.64219457
1	0.1	0.50894097
1	0.3	0.34592917
1	1	0.28330978
1	3	0.28329485
3	0.01	0.67107641
3	0.03	0.64909541
3	0.1	0.51695609
3	0.3	0.35581001
3	1	0.29324047
3	3	0.28328485

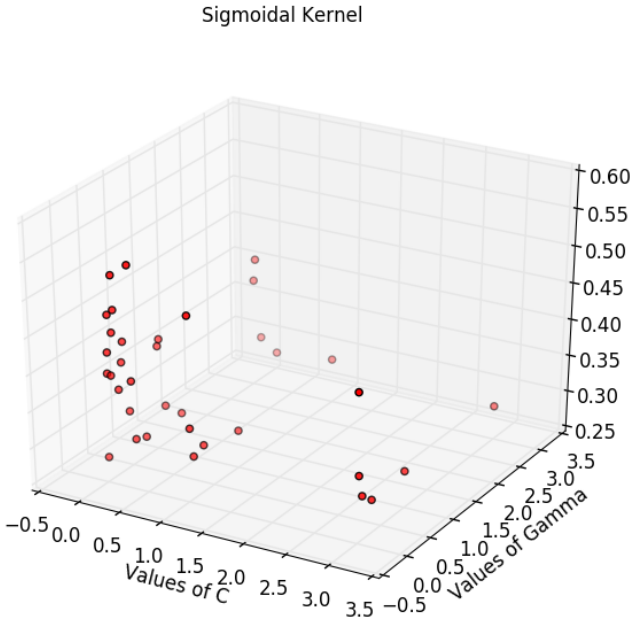
Best value of C,gamma: 3 0.01 respectively and the corresponding avg.F1-score on k-fold cross validation 0.671076405623.
The Performance of this model on test data:

Class	Precision	Recall	F1-Score	Support
0	0.60	0.60	0.60	20
1	0.89	0.85	0.87	20
2	0.67	0.90	0.77	20
3	0.57	0.40	0.47	20
avg/total	0.68	0.69	0.68	80

So we can find that guassian kernel performs slightly better than others.Number of support vectors [260 177 210 276]
(4)Sigmoidal kernel
Parameters under consideration:C,gamma,coeff0.
Ofcourse,in my code,I have taken all these criterion into account and have done a generalised validation,but for sake of visualisation we will fix coeff0 to default values and visualise the variation of F-scores with respect to C and gamma values,which are currently important for us.
let gamma $\in$ {0.01,0.03,0.1,0.3,1,3} and C $\in$ {0.01,0.03,0.1,0.3,1,3}
Now K-fold cross validation is performed for each possible combination of gamma and degree and following result is obtained.



C	Gamma	avg-F1-Score
0.01	0.01	0.27733441
0.01	0.03	0.39167239
0.01	0.1	0.38574209
0.01	0.3	0.39468717
0.01	1	0.38561179
0.01	3	0.39066751
0.03	0.01	0.42157388
0.03	0.03	0.47104405
0.03	0.1	0.44419182
0.03	0.3	0.42326063
0.03	1	0.39571682
0.03	3	0.42043989
0.1	0.01	0.52585624
0.1	0.03	0.47919829
0.1	0.1	0.36886519
0.1	0.3	0.3300631
0.1	1	0.30518585
0.1	3	0.31218162
0.3	0.01	0.54377788
0.3	0.03	0.38868239
0.3	0.1	0.30607146
0.3	0.3	0.29999191
0.3	1	0.30010669
0.3	3	0.29516545
1	0.01	0.49499958
1	0.03	0.34288483
1	0.1	0.30125545
1	0.3	0.30731083
1	1	0.29426043
1	3	0.30322162
3	0.01	0.44631621
3	0.03	0.33388887
3	0.1	0.30317624
3	0.3	0.28808191
3	1	0.29317077
3	3	0.28922067



Best value of C,gamma: 0.3 ,0.01 respectively and the corresponding avg.F1-score on k-fold cross validation 0.543777876625.
The Performance of this model on test data:

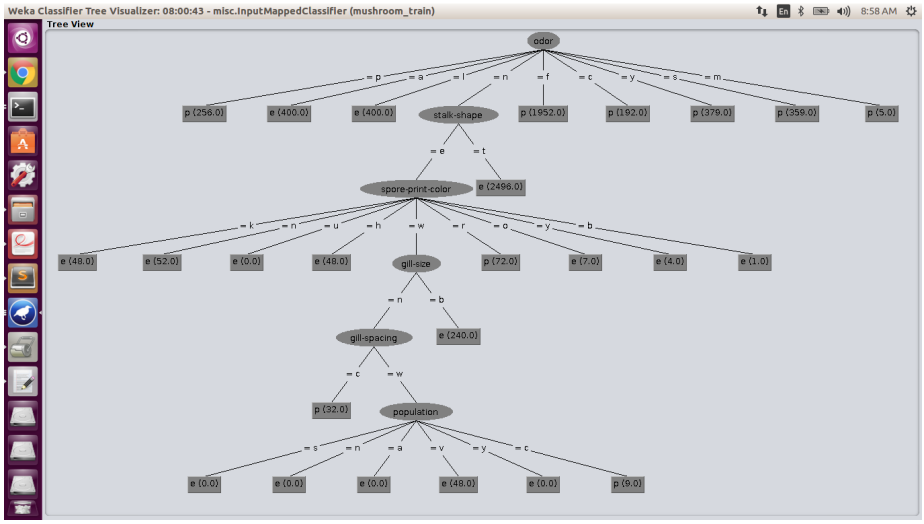
Class	Precision	Recall	F1-Score	Support
0	0.46	0.65	0.54	20
1	0.71	0.85	0.77	20
2	0.56	0.45	0.50	20
3	0.75	0.45	0.56	20
avg/total	0.62	0.60	0.59	80

The number of support vectors:[258 196 214 260]

0.5 Question 5

First,we collect data considering reduced error pruning=False and minNumObj=2(default settings) and train.Then test using test set

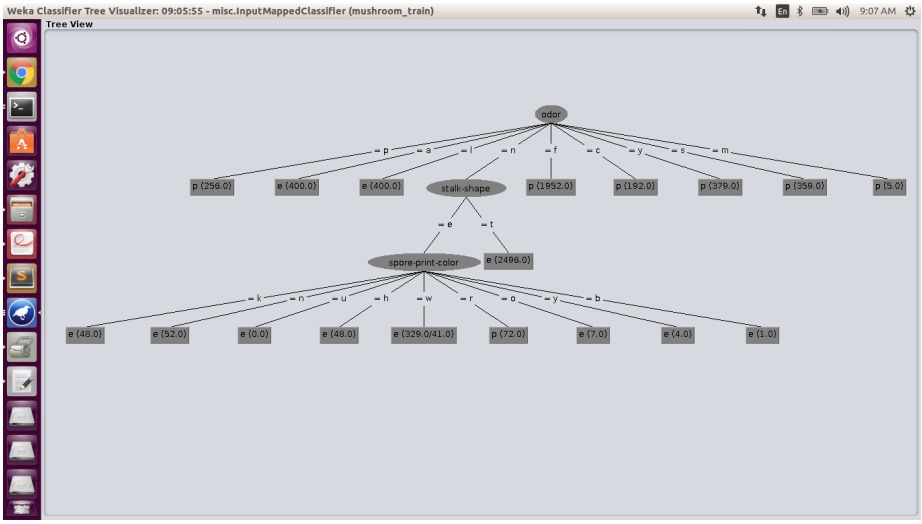
Class	Precision	Recall	F1-Score	Support
p	1.00	1.00	1.00	660
e	1.00	1.00	1.00	464
avg/total	1.00	1.00	1.00	1124



(For more details of the decision tree,look at q5M2.txt in DS)
This decision is fairly complex as it has
Number of Leaves : 25
Size of the tree : 31
Instead we can try to increase the minNumObj,reducing the size of the tree,keeping the misclassification almost the same. The following are the results on M50(ie minNumObj=50):

Class	Precision	Recall	F1-Score	Support
p	1.00	0.989	0.985	660
e	0.985	1.00	0.993	464
avg/total	0.994	0.994	0.994	1124

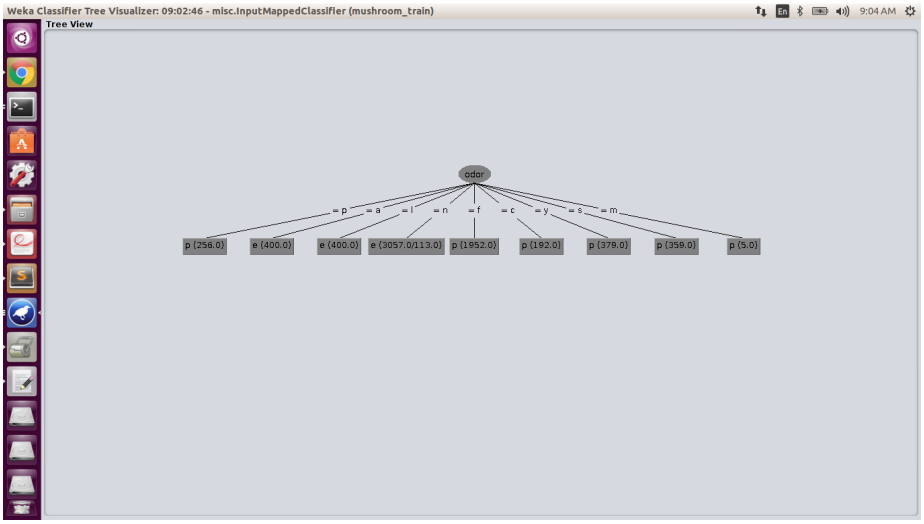
Correctly Classified Instances 1117 99.3772 %
Incorrectly Classified Instances 7 0.6228 %



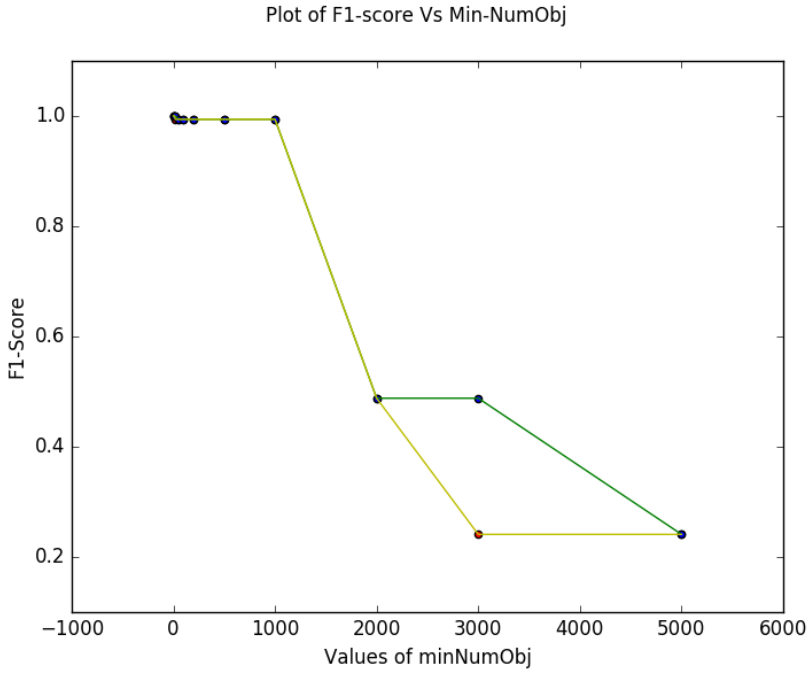
Number of Leaves : 18
Size of the tree : 21
Clearly the complexity of the tree has reduced,but the number of misclassifications is also significantly lesser.
Now lets look at M100:

Class	Precision	Recall	F1-Score	Support
p	1.00	0.989	0.985	660
e	0.985	1.00	0.993	464
avg/total	0.994	0.994	0.994	1124

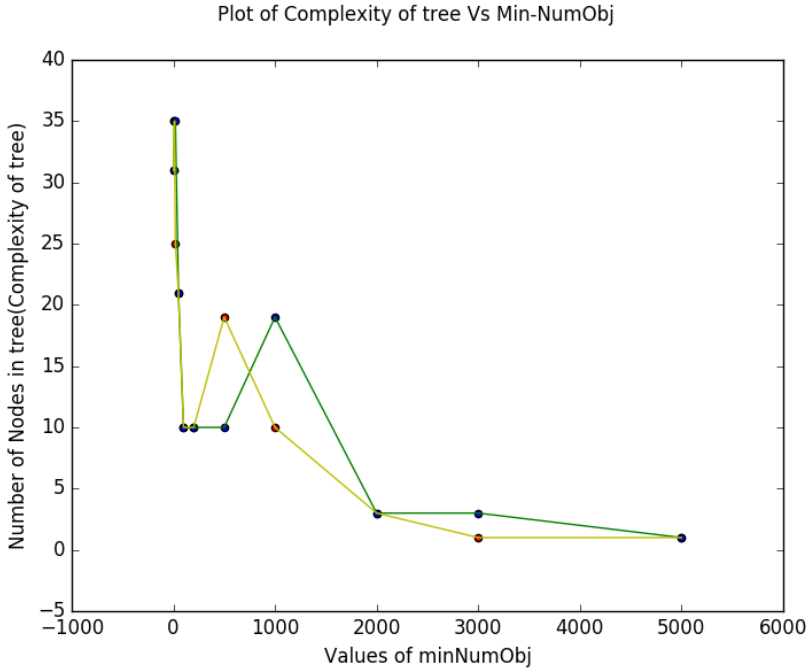
Correctly Classified Instances 1117 99.3772 %
Incorrectly Classified Instances 7 0.6228 %



Number of Leaves : 9
Size of the tree : 10
Clearly the complexity of the tree has reduced,but the number of misclassifications is also significantly lesser.
All these three trees could be considered as decision trees with best performance.
No,from M100 giving very good performance,we can say that odour is a determining feature for determining classification(edibility).
Next,from M50,we can say that spore-print-colour is also a determining feature(or maybe atleast to a good extent)
And from M2,we can say gill spacing and population are also involved in construction of decision trees.Rest of the features are not much determining.
For extensive details of the tree,we can look at M2.txt,M50.txt,M100.txt from DS3 folder.
Here is a graph plotting effect of MinNumObj in F-score:Note that green line indicates case without reduced error pruning and yellow line indicates with reduced error pruning



Here is a graph plotting effect of MinNumObj in Complexity of tree



So the answer to the question,which tree is optimal,really depends as Many

of them provide good F-score with very less misclassification errors.So in such sense,one could potentially visualise effeciency as minimum no of nodes.Also,The most complex tree we obtained sofar is 35,which is not "really" that large,to say it is complex,compared to number of features we have.Moreover we can observe that reduced error pruning doesnt cause any big difference,as tree itself is quite small.

0.6 References

1.Class Notes
2.NPTEL Video Lectures
3.sklearn library docs
4.Elements of Statistical learning(for RDA) and RDA research paper(1989)