Machine Learning Classifier

Jaimer Pastor, Mayank Kedia, Swetha Reddy

December 8, 2015

Classification

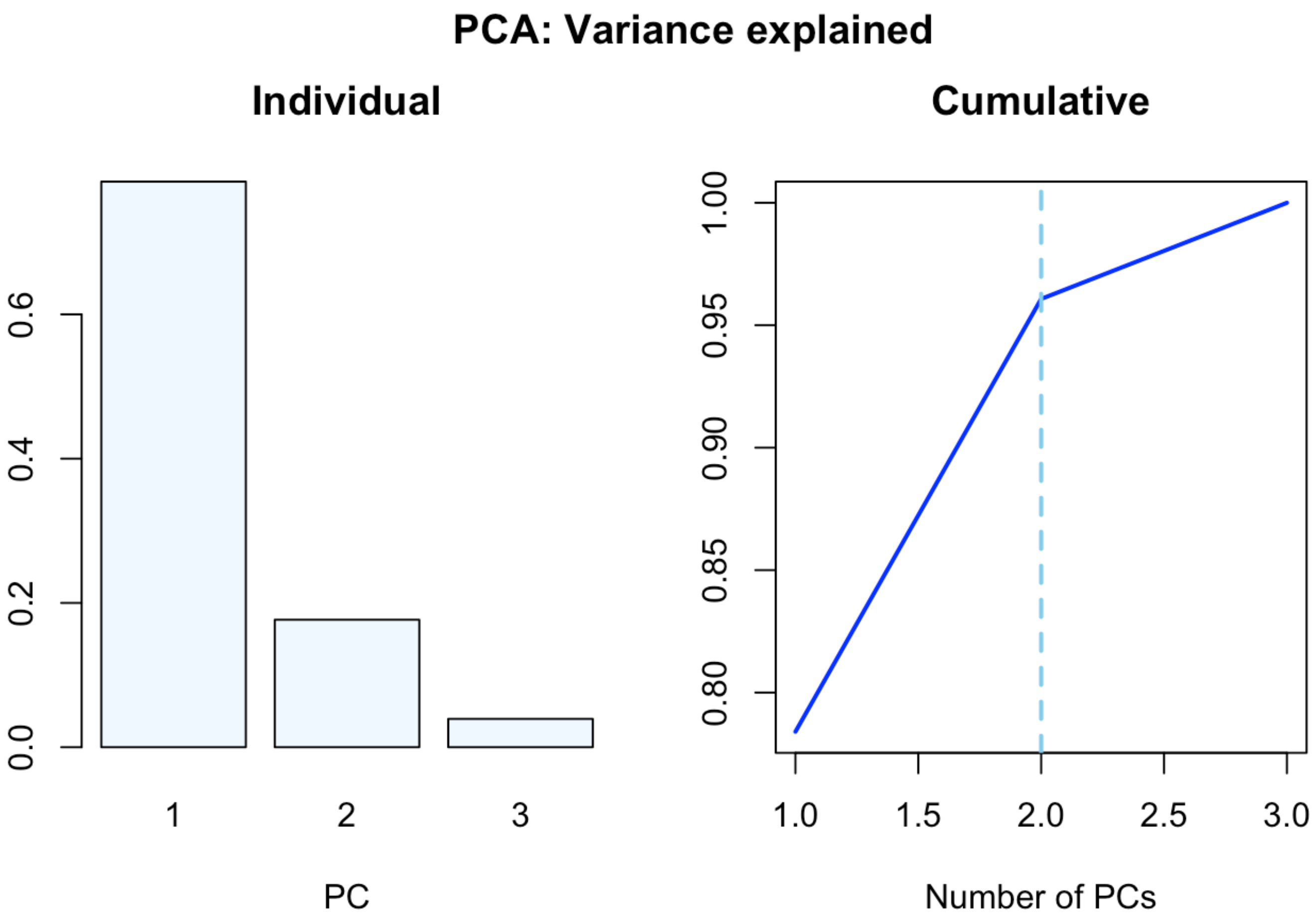
Let’s start by importing all our functionality.

Let’s import the Skin dataset from the UCI website. The summary statistics of the dataset is shown below.

```
## 'data.frame':    245057 obs. of  4 variables:
##  $ Skin: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ B   : int  74 73 72 70 70 69 70 70 76 76 ...
##  $ G   : int  85 84 83 81 81 80 81 81 87 87 ...
##  $ R   : int  123 122 121 119 119 118 119 119 125 125 ...
```

Predictor Evaluation The function classification.predictor.evaluation does the parameter evaluation from the data set. It returns number of Principal components required to explain 90% of the variance. This method also gives the summary statistics of the dataset.

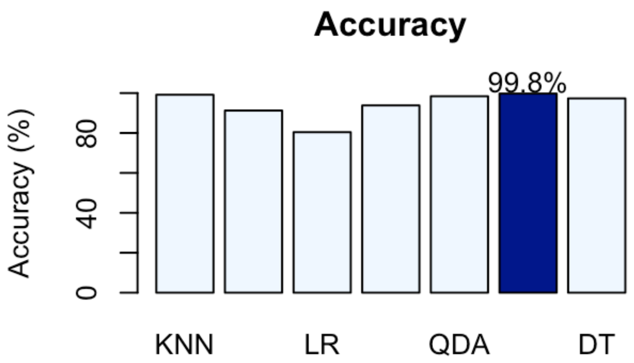
```
##
##
## ***************** PREDICTOR EVALUATION *******************************
##
## Total number of observations:       245057
## Total number of complete cases:     245057
## Total number of variables:          4
##  - Number of non-numeric variables: 0
##  - Number of numeric variables:     4
##
## Mean of each predictor:
##        B        G        R
## 125.0654 132.5073 123.1772
##
## Standard deviation of each predictor:
##        B        G        R
## 62.25565 59.94120 72.56216
##
## Significant predictors in logistic regression:
## B  ,  G  ,  R
##
## Analyzing collinearity:
## The following variables show signs of collinearity:
##   Var1 Var2 Correlation
## 1 G    B    0.855
##
##
## Principal Component Analysis (PCA): Variance explained
##    PC1    PC2    PC3
## 0.7840 0.1767 0.0393
```

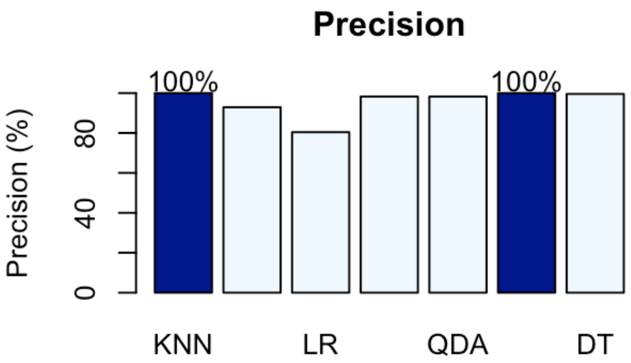


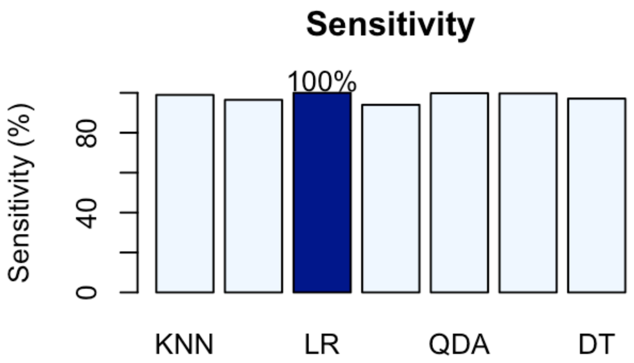
We just use random sample of 7000 rows from the original dataset of 300k records.

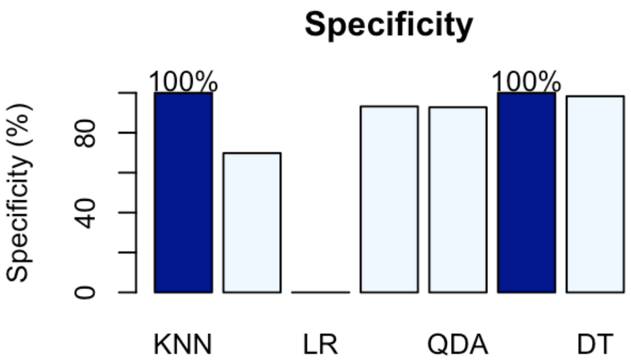
This method checks all the classifiers and returns the summary comparision of all the statistics together.

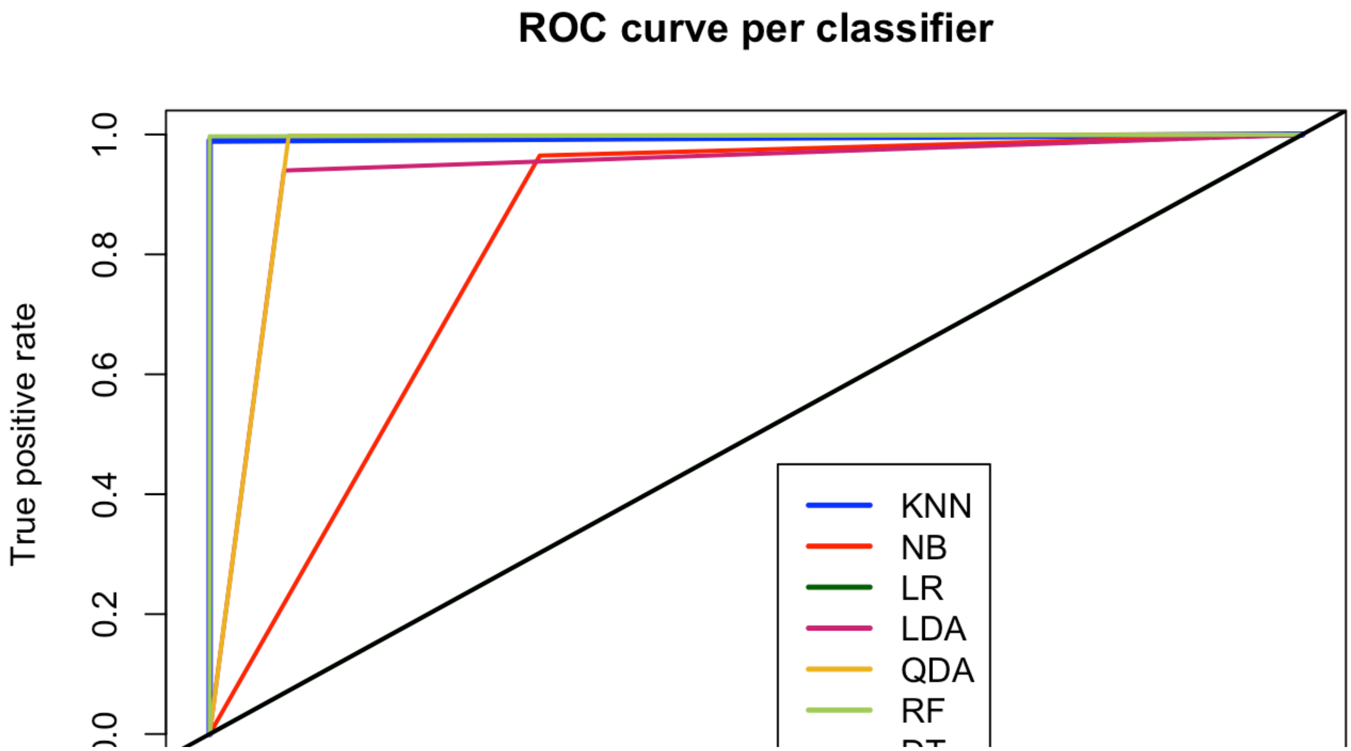
```
##
##   Classifier MSPE_test Accuracy Sensitivity Specificity Precision Ranking
## 1 KNN        0.00833   0.992    0.990       1.000       1.000
## 2 NB         0.08750   0.912    0.965       0.698       0.929
## 3 LR         0.80417   0.804    1.000       0.000       0.804
## 4 LDA        0.06167   0.938    0.940       0.932       0.983
## 5 QDA        0.01583   0.984    0.998       0.928       0.983
## 6 RF         0.00250   0.998    0.997       1.000       1.000     BEST
## 7 DT         0.02667   0.973    0.971       0.983       0.996
```

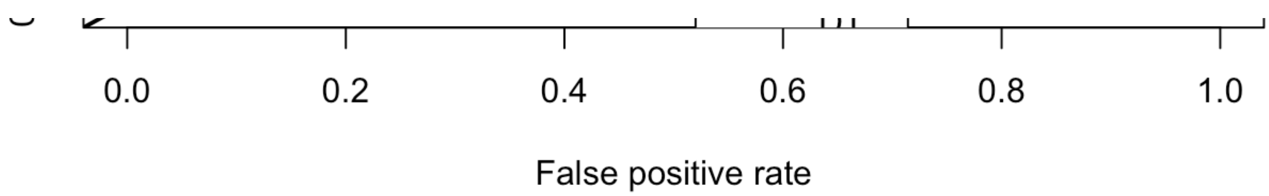

Accuracy
Accuracy (%)
0
40
80
99.8%
KNN
LR
QDA
DT


Precision
Precision (%)
0
40
80
100%
100%
KNN
LR
QDA
DT


Sensitivity
Sensitivity (%)
0
40
80
100%
KNN
LR
QDA
DT


Specificity
Specificity (%)
0
40
80
100%
100%
KNN
LR
QDA
DT

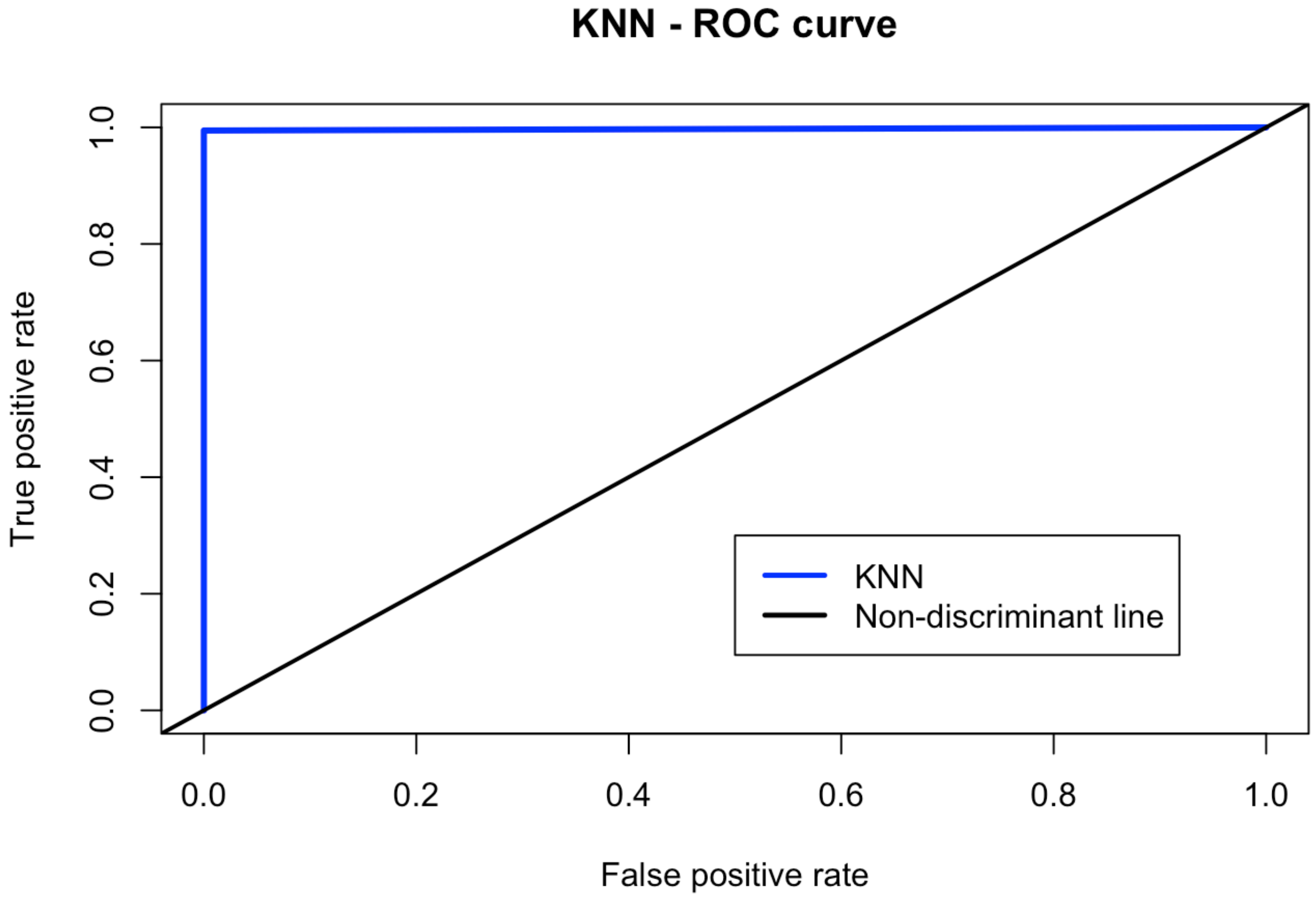

ROC curve per classifier
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
KNN
NB
LR
LDA
QDA
RF
DT



K-Nearerst Neighbor

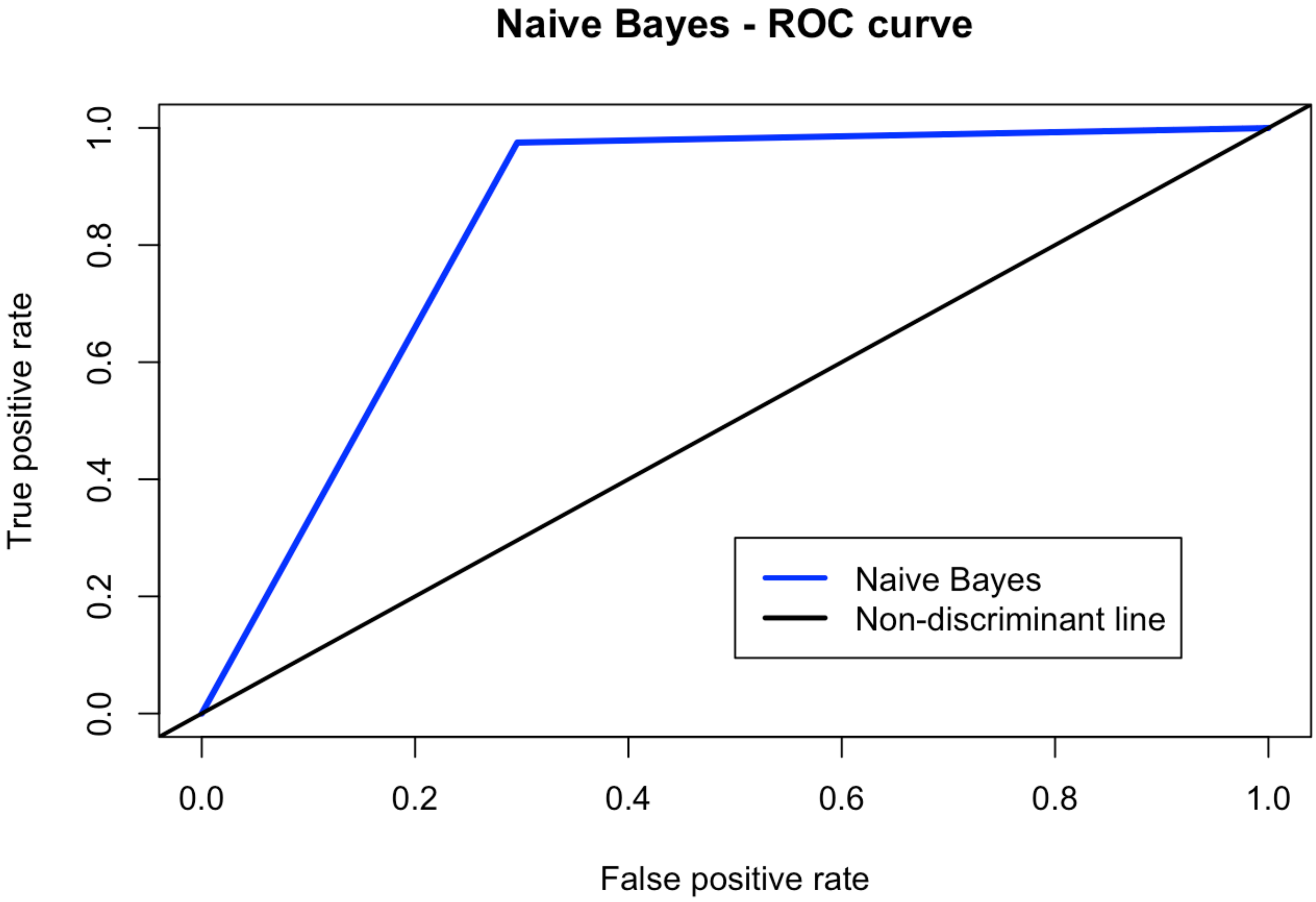
Here we test the KNN classfier alone with the classifier type we passed in as classifier = “knn” and the specify the response column as “Skin”. The output is ROC curve for the classifier.

```
##
##
## ***************** KNN Classification *******************************
##
## K-Value                         : 5
## Number of Dimensions (predictors): 3
## Training set size               : 4800
## Test set size                   : 1200
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.00417
## Accuracy   : 0.996
## Sensitivity: 0.995
## Specificity: 1
## Precision  : 1
```



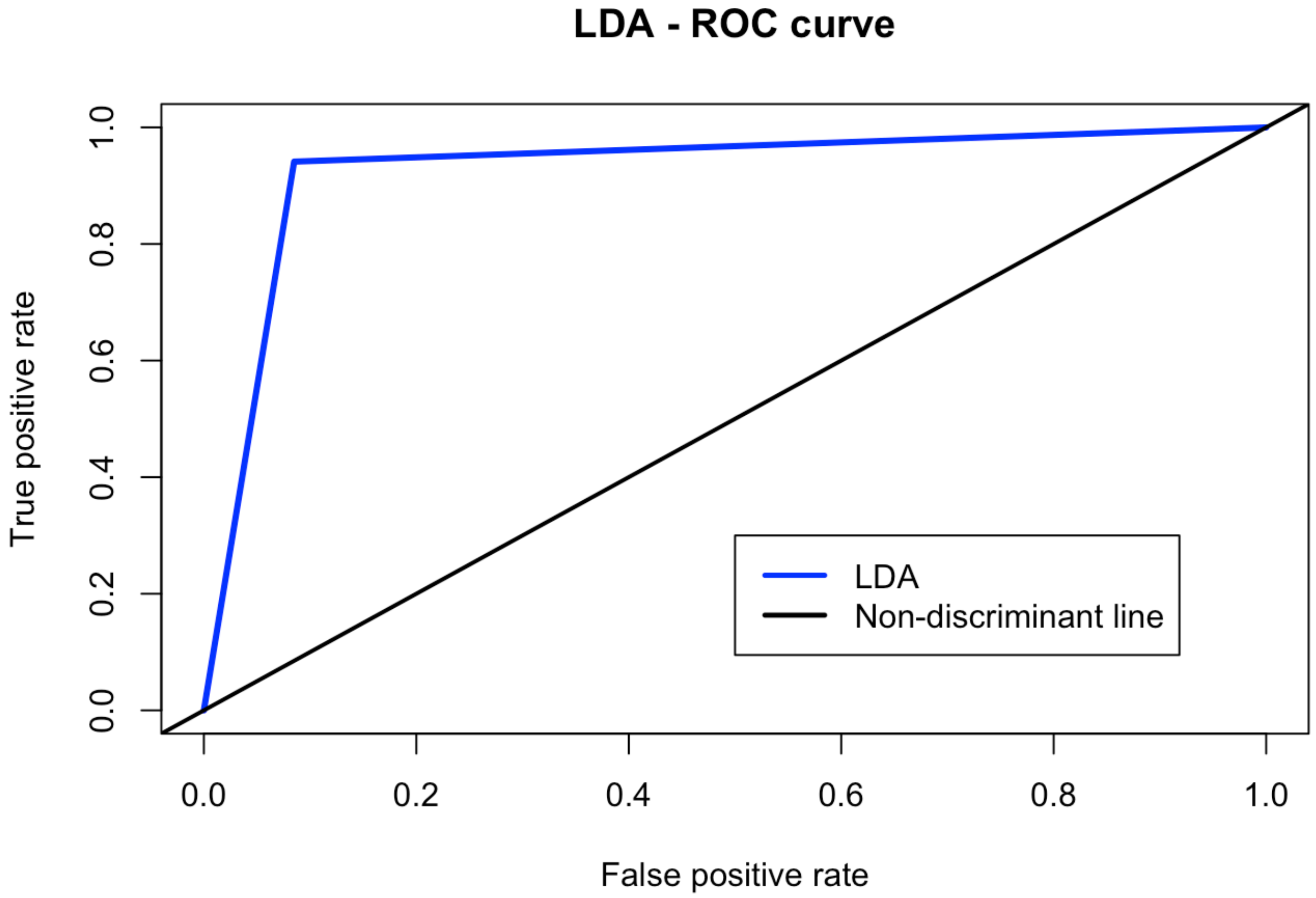
Naive Bayes Here we test the Naive Bayes classfier alone with the classifier type we passed in as classifier = “nb” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier.

```
##
##
## ***************** Naive bayes Classification *******************************
##
## Number of predictors: 3
##
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.0792
## Accuracy   : 0.921
## Sensitivity: 0.975
## Specificity: 0.704
## Precision  : 0.929
##
##
##  -------- Model Assumptions --------
##
## The conditional probability distribution of each predictor for a given output is i
ndependent of other predictors
```



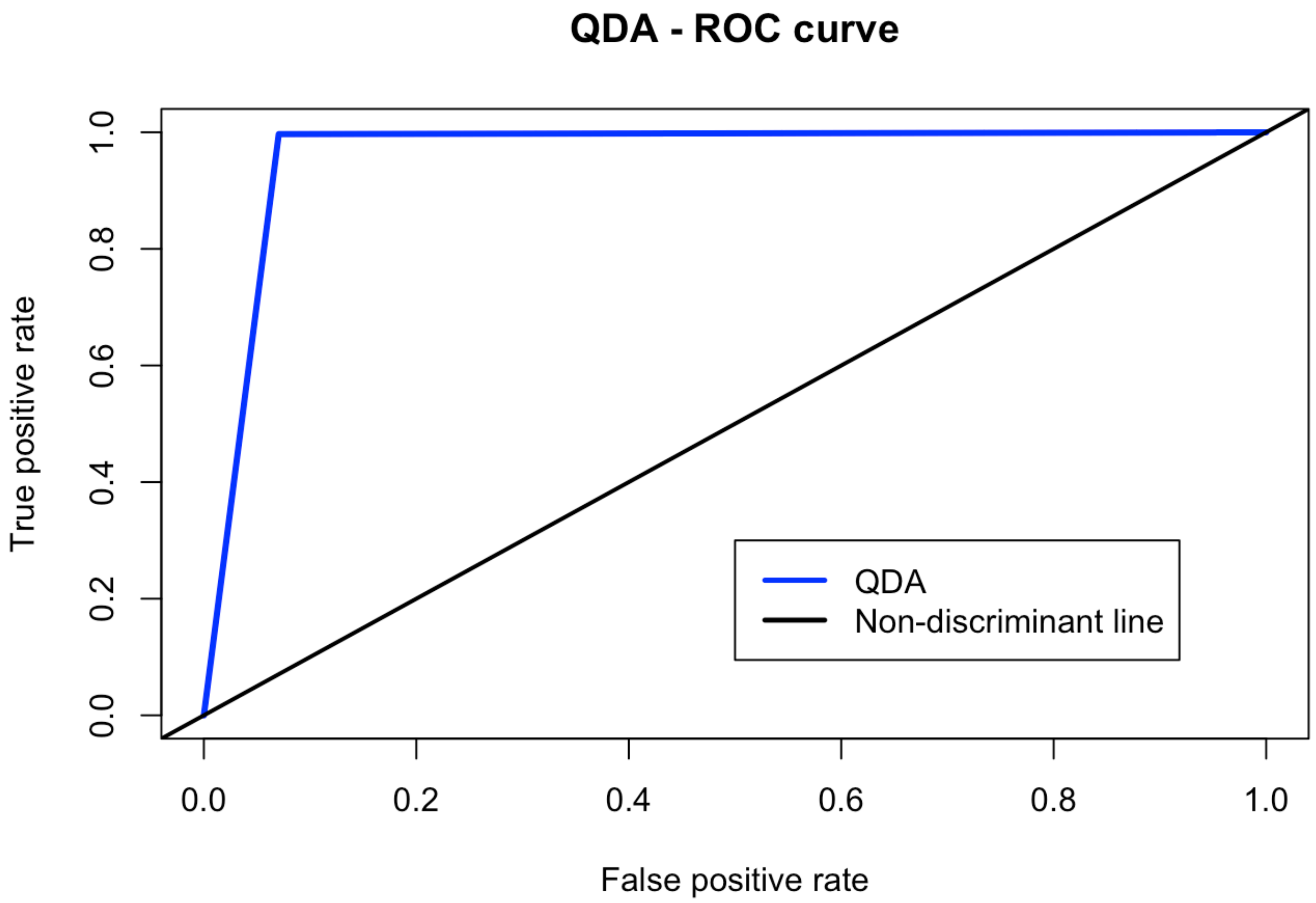
Linear discriminant analysis Here we test the Linear discriminant analysis classfier, with the classifier type we passed in as classifier = “lda” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier. The ROC curve is output.

```
##
##
## ***************** Linear Discriminant Analysis Classification ********************
***********
##
## Number of Dimensions (predictors): 3
## Training set size                : 4800
## Test set size                    : 1200
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.0642
## Accuracy   : 0.936
## Sensitivity: 0.941
## Specificity: 0.915
## Precision  : 0.977
##
##
##  -------- Model Assumptions --------
##
##  - All predictors are continuous
##  - All predictors have Normally distributed and independent conditional probabilit
ies
##  - Same covariance matrix for both sets of conditional probabilities
```



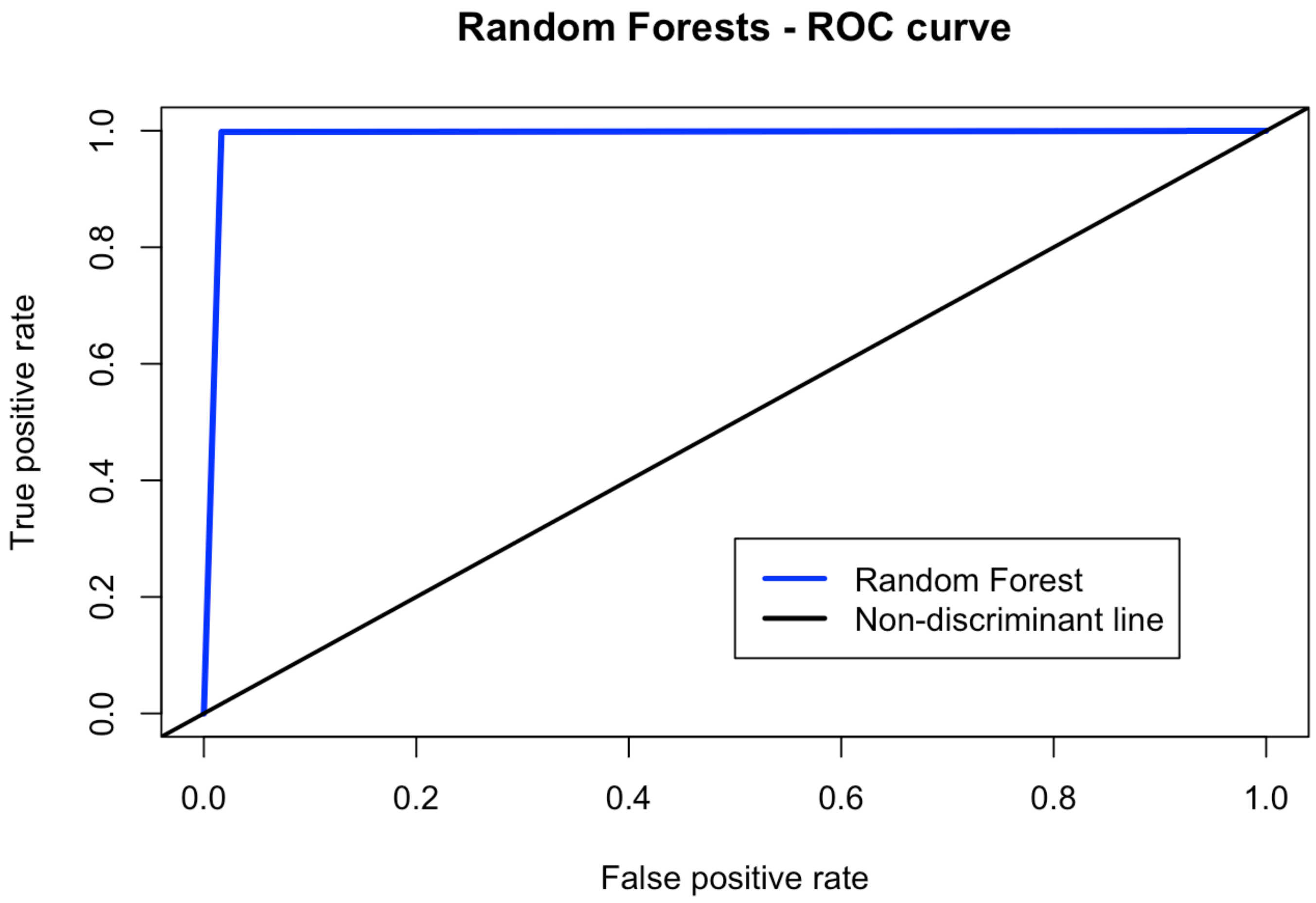
Quadratic discriminant analysis Here we test the Quadratic discriminant analysis classfier, with the classifier type we passed in as classifier = “qda” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier. The ROC curve is output.

```
##
##
## ***************** Quadratic Discriminant Analysis Classification *****************
**************
##
## Number of Dimensions (predictors): 18
## Training set size                : 4800
## Test set size                    : 1200
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.0175
## Accuracy   : 0.983
## Sensitivity: 0.997
## Specificity: 0.929
## Precision  : 0.981
##
##
##  -------- Model Assumptions --------
##
##  - All predictors are continuous
##  - All predictors have Normally distributed and independent conditional probabilit
ies
```



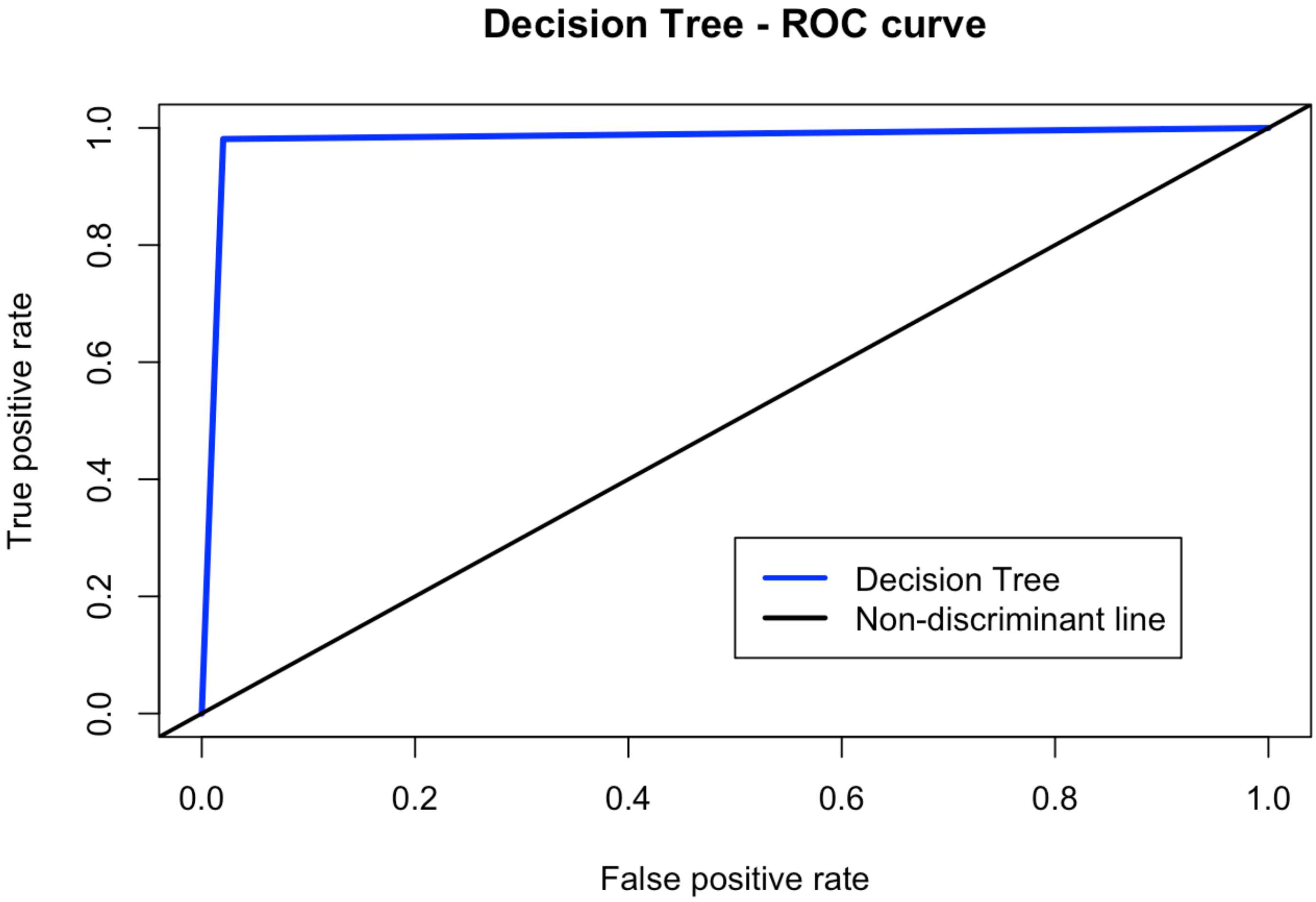
Random Forest Here we test the Random forest classfier, with the classifier type we passed in as classifier = “rf” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier. The ROC curve is output.

```
##
##
## ***************** Random Forest Classification *******************************
##
## Number of Dimensions (predictors): 3
## Number of Trees                  : 500
## Number of params in each tree    : 2
## Training set size                : 4800
## Test set size                    : 1200
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.005
## Accuracy   : 0.995
## Sensitivity: 0.998
## Specificity: 0.983
## Precision  : 0.996
```



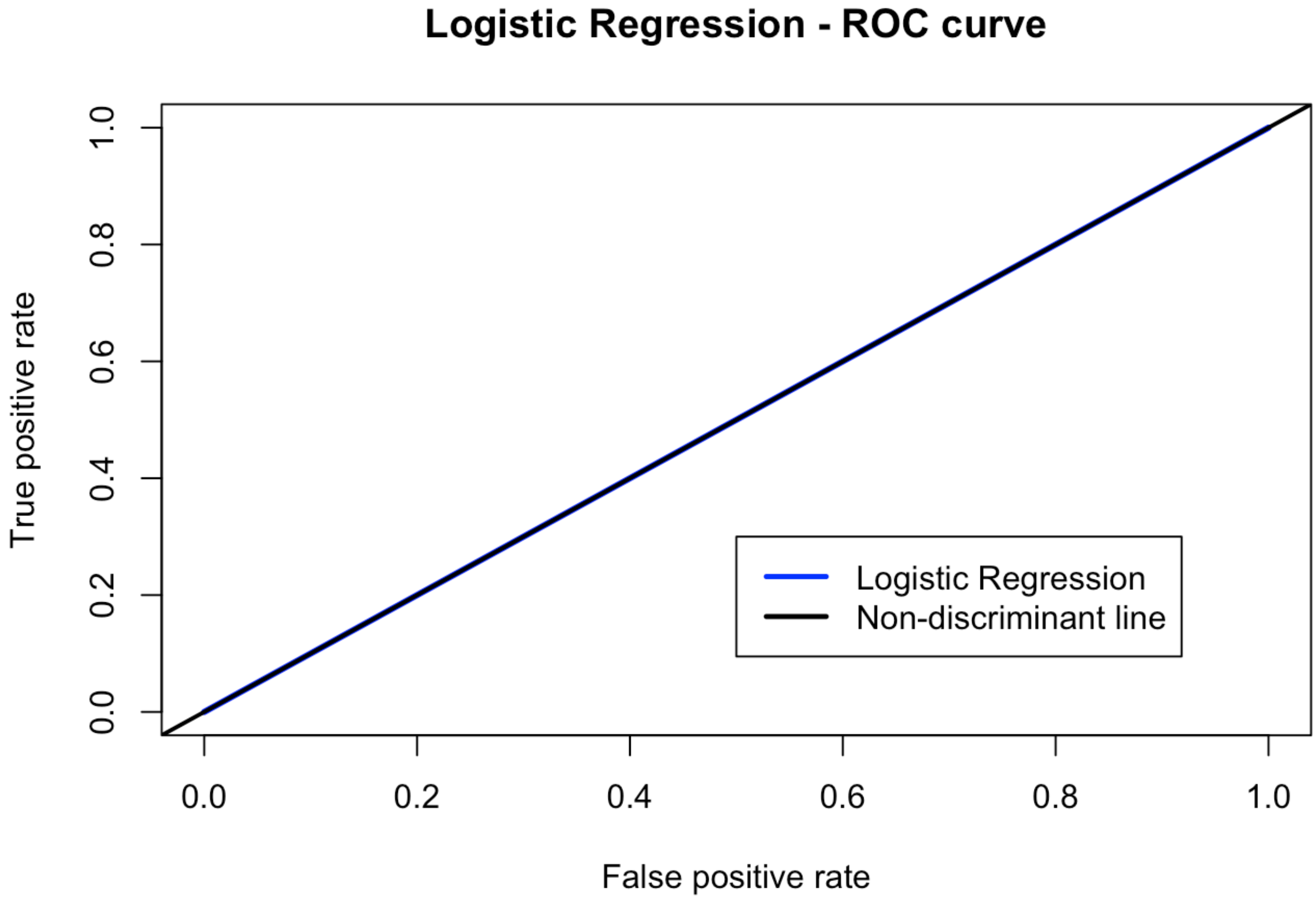
Decision Tree Here we test the decision tree classfier, with the classifier type we passed in as classifier = “dt” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier. The ROC curve is output.

```
## 
## 
## ***************** Decision Tree Classification *******************************
## 
## Training set size                : 4800
## Test set size                    : 1200
## 
##  -------- Accuracy Measures --------
## 
## MSPE       : 0.0192
## Accuracy   : 0.981
## Sensitivity: 0.981
## Specificity: 0.98
## Precision  : 0.995
## 
## 
##  -------- Tree Structure --------
## 
## 1) R <= 170; criterion = 1, statistic = 1533.658
##   2) R <= 138; criterion = 1, statistic = 229.235
##     3) R <= 122; criterion = 1, statistic = 83.336
##       4) R <= 119; criterion = 0.998, statistic = 11.948
##         5)*  weights = 2001 
##       4) R > 119
##         6) G <= 139; criterion = 1, statistic = 118.298
##           7)*  weights = 7 
##         6) G > 139
##           8)*  weights = 129 
##     3) R > 122
##       9) B <= 86; criterion = 1, statistic = 580.793
##         10) B <= 43; criterion = 0.997, statistic = 11.046
##           11)*  weights = 7 
##         10) B > 43
##           12)*  weights = 68 
##       9) B > 86
##         13)*  weights = 737 
##   2) R > 138
##     14) B <= 122; criterion = 1, statistic = 460.703
##       15) G <= 23; criterion = 1, statistic = 49.919
##         16)*  weights = 19 
##       15) G > 23
##         17) B <= 63; criterion = 1, statistic = 39.402
##           18)*  weights = 127 
##         17) B > 63
##           19)*  weights = 43
##     14) B > 122
##       20)*  weights = 508
## 1) R > 170
##   21) B <= 209; criterion = 0.988, statistic = 8.217
##     22) B <= 54; criterion = 1, statistic = 160.514
##       23) B <= 45; criterion = 0.999, statistic = 12.909
##         24)*  weights = 68
##       23) B > 45
##         25) R <= 213; criterion = 1, statistic = 15.745
##           26)*  weights = 8
##         25) R > 213
##           27)*  weights = 13
##     22) B > 54
##       28) G <= 105; criterion = 1, statistic = 76.906
##         29)*  weights = 61
##       28) G > 105
##         30) G <= 215; criterion = 1, statistic = 36.38
##           31)*  weights = 856
##         30) G > 215
##           32)*  weights = 19
##   21) B > 209
##     33) B <= 217; criterion = 0.915, statistic = 4.752
##       34)*  weights = 8
##     33) B > 217
##       35)*  weights = 121
```



Logistic Regression Here we test the Logistic regression classfier, with the classifier type we passed in as classifier = “lr” and the specify the response column as “Skin”. The output is a sumamry of the accuracy, MSPE, Sensitivity, Specificity and Precision of the classifier. The ROC curve is output.

```
##
##
## ***************** Logistic Regression Classification *****************************
**
##
## Number of Dimensions (predictors): 3
## AIC                              : 2382
## Deviance                         : 2374
## Null Deviance                    : 4905
## Training set size                : 4800
## Test set size                    : 1200
##
##  -------- Coefficients learned --------
##
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  4.26396    0.19462   21.91 2.12e-106
## B            0.02535    0.00179   14.19  9.93e-46
## G           -0.00528    0.00218   -2.42  1.54e-02
## R           -0.03481    0.00112  -31.06 7.61e-212
##
##
##  -------- Accuracy Measures --------
##
## MSPE       : 0.783
## Accuracy   : 0.783
## Sensitivity: 1
## Specificity: 0
## Precision  : 0.783
```



Output

From the individual classifier object user can access handle to the object the rss.knn@finalModel (mailto:rss.knn@finalModel), which can be used for prediction. The prediction function is used to get the ROC curves for the model. And classifier metrics is used to get all the metreics for that particular classifier. The print.flag = TRUE needs to be specified if you need the output printed to the console.

```
## 
## - Classifier metrics:
##    MSPE:  0.001
##    Accuracy:  0.999
##    Sensitivity:  0.999
##    Specificity:  1
##    Precision:  1
```

```
## [1] 0.001 0.999 0.999 1.000 1.000
```