Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Naive Bayes

Subhalaxmi Rout

Instructions

Use Penguine dataset for this assignment. Please use "Species" as your target variable. For this assignment, you may want to drop/ignore the variable "year".

Using the target variable, Species, please conduct:

(*a.*) Linear Discriminant Analysis (30 points):

- You want to evaluate all the 'features' or dependent variables and see what should be in your model. Please comment on your choices.
- Just a suggestion: You might want to consider exploring featurePlot on the caret package. Basically, you look at each of the features/dependent variables and see how they are different based on species. Simply eye-balling this might give you an idea about which would be strong 'classifiers' (aka predictors).
- Fit your LDA model using whatever predictor variables you deem appropriate. Feel free to split the data into training and test sets before fitting the model.
- Look at the fit statistics/ accuracy rates.

Linear Discriminant Analysis (LDA)

Linear discriminant analysis is an extremely popular dimensionality reduction technique. Dimensionality reduction techniques have become critical in machine learning since many multi-dimensional datasets exist these days.

Multi-dimensional data is data that has multiple features which have a correlation with one another. Dimensionality reduction simply means plotting multi-dimensional data in just 2 or 3 dimensions.

```
library(palmerpenguins)
library(stats)
library(dplyr)
library(PerformanceAnalytics)
library(DT)
library(tidyr)
library(caret)
library(e1071)
library(kableExtra)
library(caTools)
library(MASS)
```

```
library(devtools)
library(ggord)
library(klaR)
library(naivebayes)
```

Load libraries

Load data Load the data from `palmerpenguins` library and drop year, sex and island columns from dataset.

```
penguine_data <- glimpse(penguins)
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
## $ sex               <fct> male, female, female, NA, female, male, female, m...
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

```
penguine_data <- penguine_data %>% dplyr::select(-c(year, island, sex))
DT::datatable(head(penguine_data,5))
```

Show 10 entries Search:

	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
1	Adelie	39.1	18.7	181	3750
2	Adelie	39.5	17.4	186	3800
3	Adelie	40.3	18	195	3250
4	Adelie				
5	Adelie	36.7	19.3	193	3450

Showing 1 to 5 of 5 entries Previous 1 Next

EDA Have a look on summary statistics.

```
summary(penguine_data)
```

```
##       species    bill_length_mm  bill_depth_mm   flipper_length_mm
##  Adelie   :152   Min.   :32.10   Min.   :13.10   Min.   :172.0
##  Chinstrap: 68   1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0
##  Gentoo   :124   Median :44.45   Median :17.30   Median :197.0
##                  Mean   :43.92   Mean   :17.15   Mean   :200.9
##                  3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0
##                  Max.   :59.60   Max.   :21.50   Max.   :231.0
##                  NA's   :2       NA's   :2       NA's   :2
```

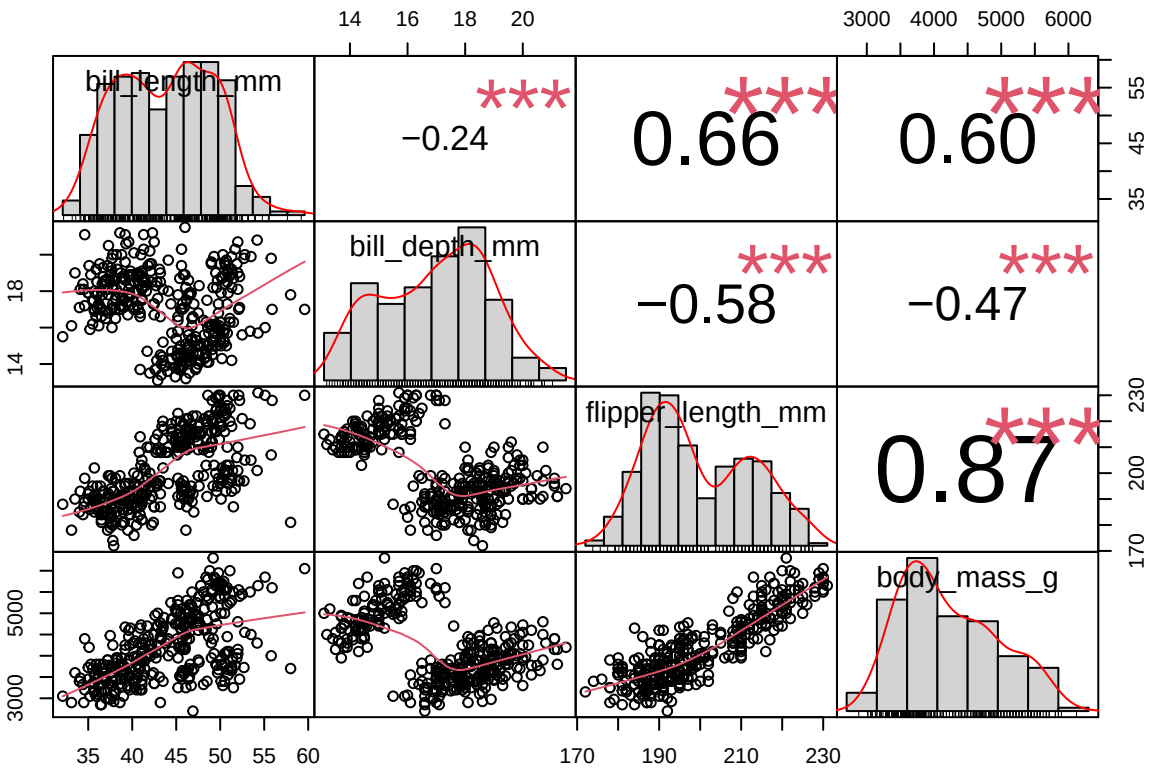
```
##   body_mass_g
##  Min.   :2700
##  1st Qu.:3550
##  Median :4050
##  Mean   :4202
##  3rd Qu.:4750
##  Max.   :6300
##  NA's   :2
```

We can see 8 NAs are present in the dataset. Remove NAs from data.

```
penguine_data <- drop_na(penguine_data)
```

To see the variables co-related to each other plot a co-relation graph.

```
my_data <- penguine_data[, c(2,3,4,5)]
chart.Correlation(my_data, histogram=TRUE, pch=19)
```



Above plot shows:

- Positive co-relation between body_mass_g and flipper_length_mm
- Negative Co-relation between bill_depth_mm and flipper_length_mm
- Positive co-relation between bill_length_mm and flipper_length_mm
- Positive co-relation between body_mass_g and bill_length_mm

Split data Split data in to 2 sets train and test. Train data and Test data ration is 70:30.

```r
set.seed(123)


# Data split
sample = sample.split(penguine_data$species, SplitRatio = 0.70)

penguine_train = subset(penguine_data, sample == TRUE)
penguine_test = subset(penguine_data, sample == FALSE)

dim(penguine_train)
```

```
## [1] 240   5
```

```r
dim(penguine_test)
```

```
## [1] 102   5
```

Train test has 240 and Test test has 102 rows.

LDA for all variables The linear Discriminant analysis estimates the probability that a new set of inputs belongs to every class. In our dataset dependant variable is `species` and all other 4 variables/fields are independent.

Load library `MASS` to perform LDA. Apply LDA on train dataset and look at the model structure.

```r
lda_all <- lda(species ~ ., data = penguine_train)
lda_all
```

```
## Call:
## lda(species ~ ., data = penguine_train)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4416667 0.2000000 0.3583333
##
## Group means:
##           bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## Adelie          38.70283      18.33585          190.0566    3704.481
## Chinstrap       49.02708      18.48958          197.1250    3766.667
## Gentoo          47.78023      15.02558          217.6860    5116.279
##
## Coefficients of linear discriminants:
##                             LD1          LD2
## bill_length_mm     0.089443430 -0.383958867
## bill_depth_mm     -0.985577100 -0.053402145
## flipper_length_mm  0.082416295  0.006036714
## body_mass_g        0.001263067  0.001683273
##
## Proportion of trace:
##    LD1    LD2
## 0.8709 0.1291
```

Above model shows, *Prior probabilities of groups:*

- 44.2% belongs to Adelie in training data
- 20% belongs to Chinstrap in training data
- 35.9% belongs to Gentoo in training data

Group means: This table shows for ecah species and each variables we have averages. For examaple, Adelie's average bill_length_mm is 38.7.

Coefficients of linear discriminants:

The first discreminant function is a linear combination of the four variables. Example: $0.089443430 * bill_length_mm - 0.985577100 * bill_depth_mm + 0.082416295 * flipper_length_mm + 0.001263067 * body_mass_g$

Proportion of trace: Percentage separations achieved by the first discreminant function is 87%. Percentage separations achieved by the second discreminant function is 13%.

LDA for body mass and bill depth From EDA, we relation between body mass and flipper length, and body mass and bill length. So to avoid co-linearty exclude body mass and flipper length variables from the model.

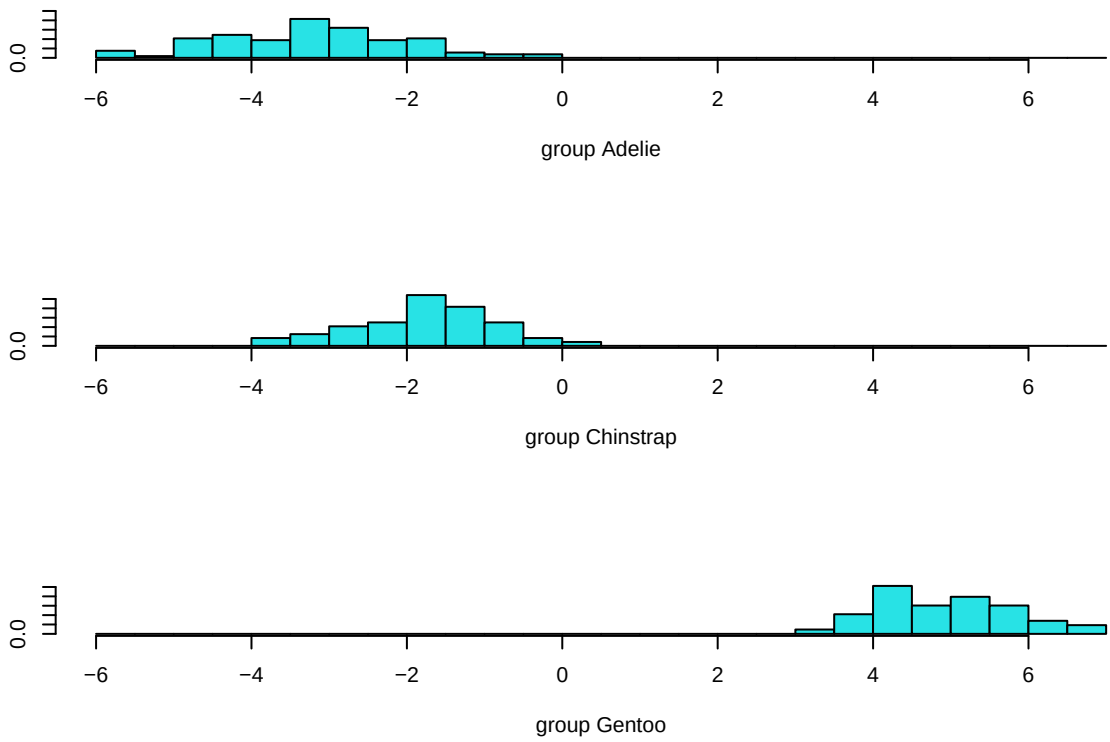
```
lda_2 <- lda(species ~ bill_length_mm + bill_depth_mm, data = penguine_train)
lda_2
```

```
## Call:
## lda(species ~ bill_length_mm + bill_depth_mm, data = penguine_train)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4416667 0.2000000 0.3583333
##
## Group means:
##           bill_length_mm bill_depth_mm
## Adelie          38.70283      18.33585
## Chinstrap       49.02708      18.48958
## Gentoo          47.78023      15.02558
##
## Coefficients of linear discriminants:
##                        LD1        LD2
## bill_length_mm  0.3367164 -0.1673935
## bill_depth_mm  -0.8441757 -0.5163900
##
## Proportion of trace:
##   LD1   LD2
## 0.922 0.078
```

Proportion of trace: Percentage separations achieved by the first discreminant function is 99.9%. Percentage separations achieved by the second discreminant function is 0.1%. Which is quite higher than our first model.

LDA Advantage Histogram and Bi-plot provides useful insights and are helpful for interpretaion of the analysis.

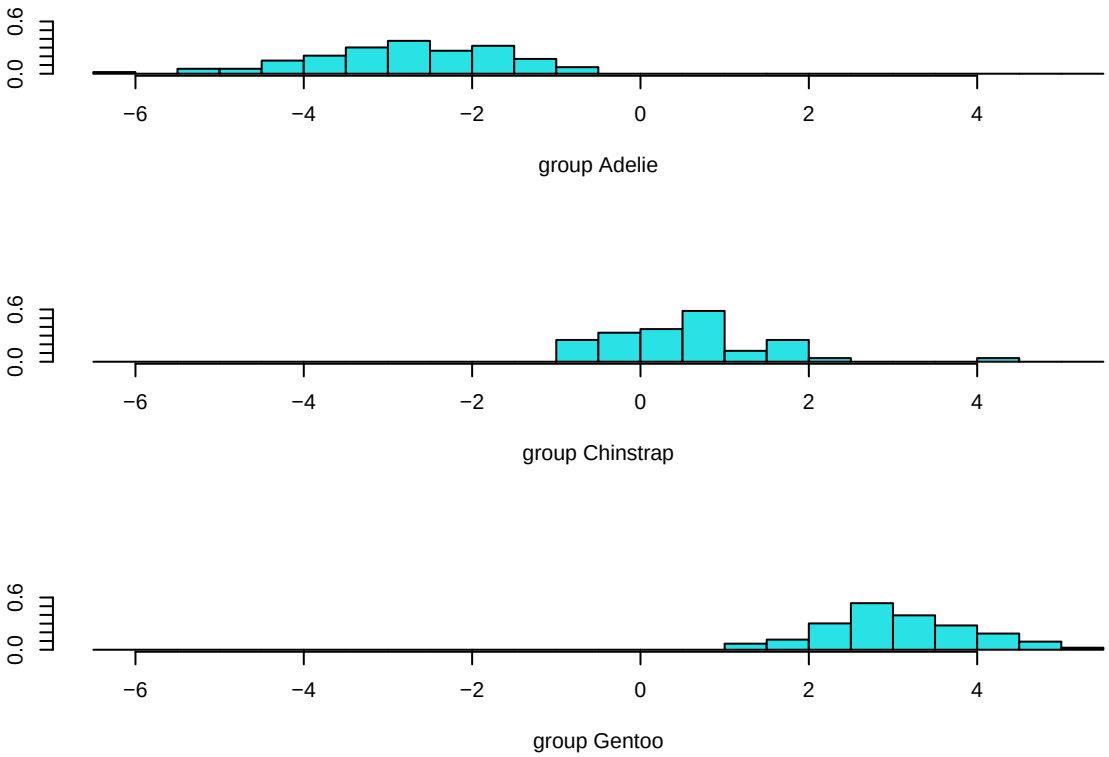
```r
# predict for train data
P_lda_all <- predict(lda_all, penguine_train)
P_lda_2 <- predict(lda_2, penguine_train)

# histigram of all variables lda models
ldahist(data = P_lda_all$x[,1], g = penguine_train$species)
```



bi-Plot

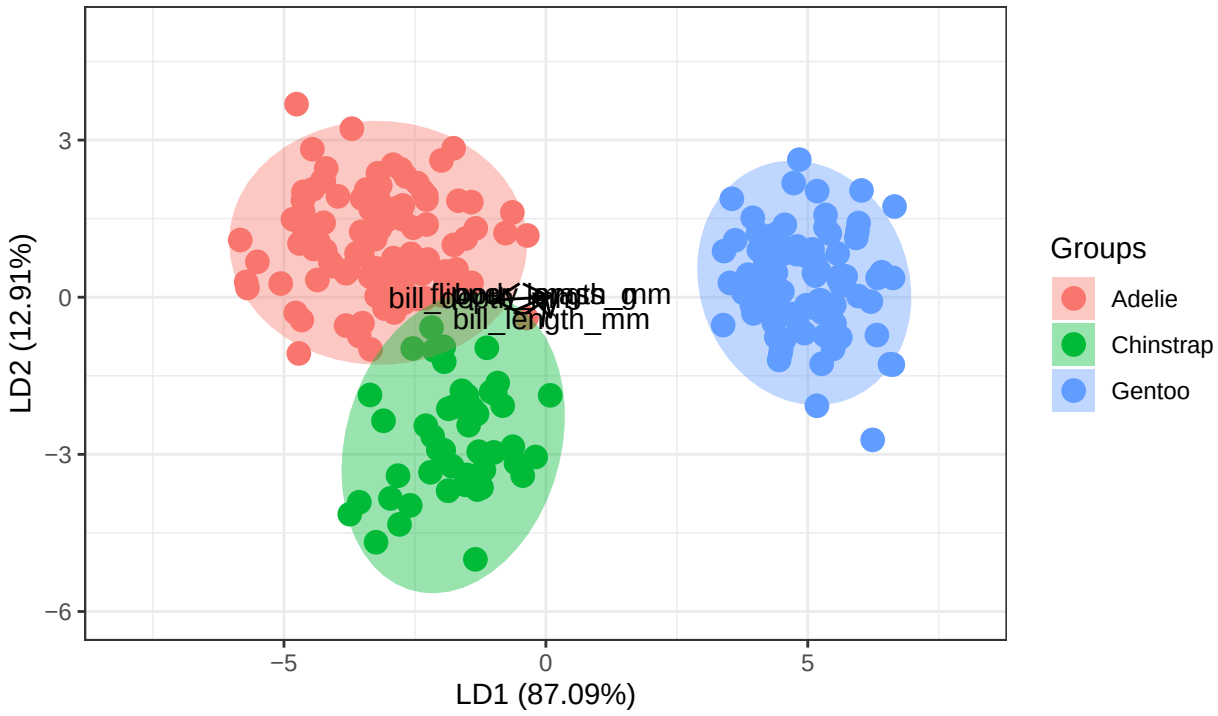
We see using our first model there is little over-lap between Adile and Chinstrap

```r
# histigram of 2nd lda models
ldahist(data = P_lda_2$x[,1], g = penguine_train$species)
```



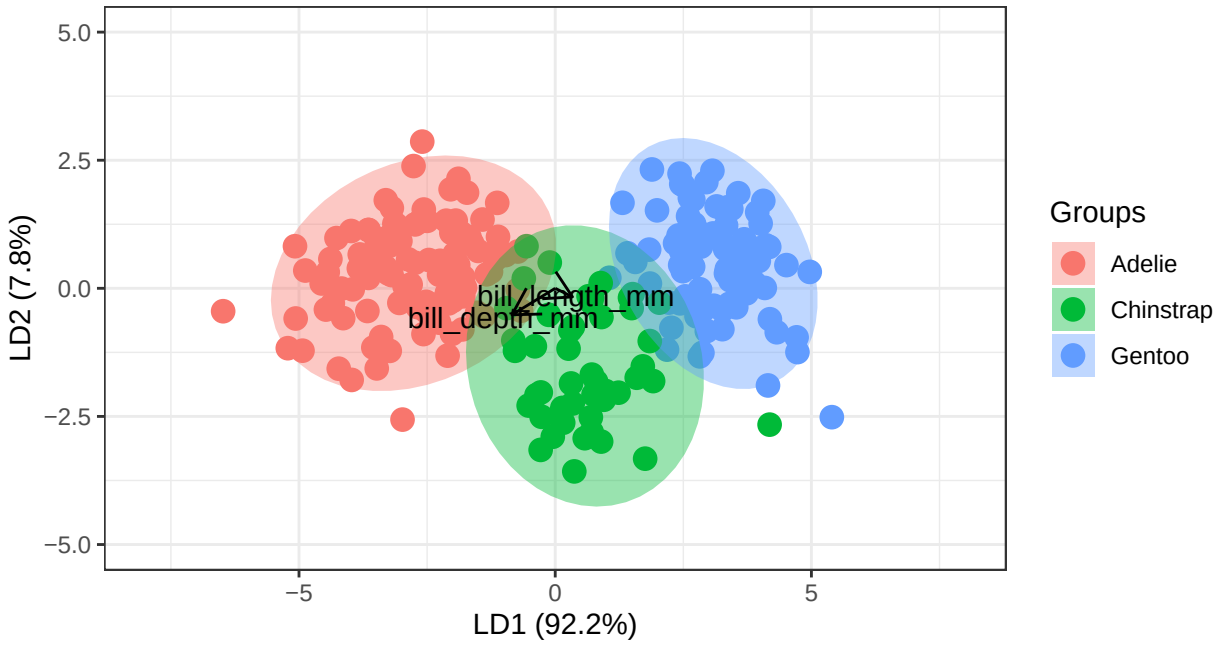
This model separates 3 species better than first model. Very few over-lap we see between species.

```
ggord(lda_all, penguine_train$species, ylim = c(-6,5), xlim = c(-8,8))
```

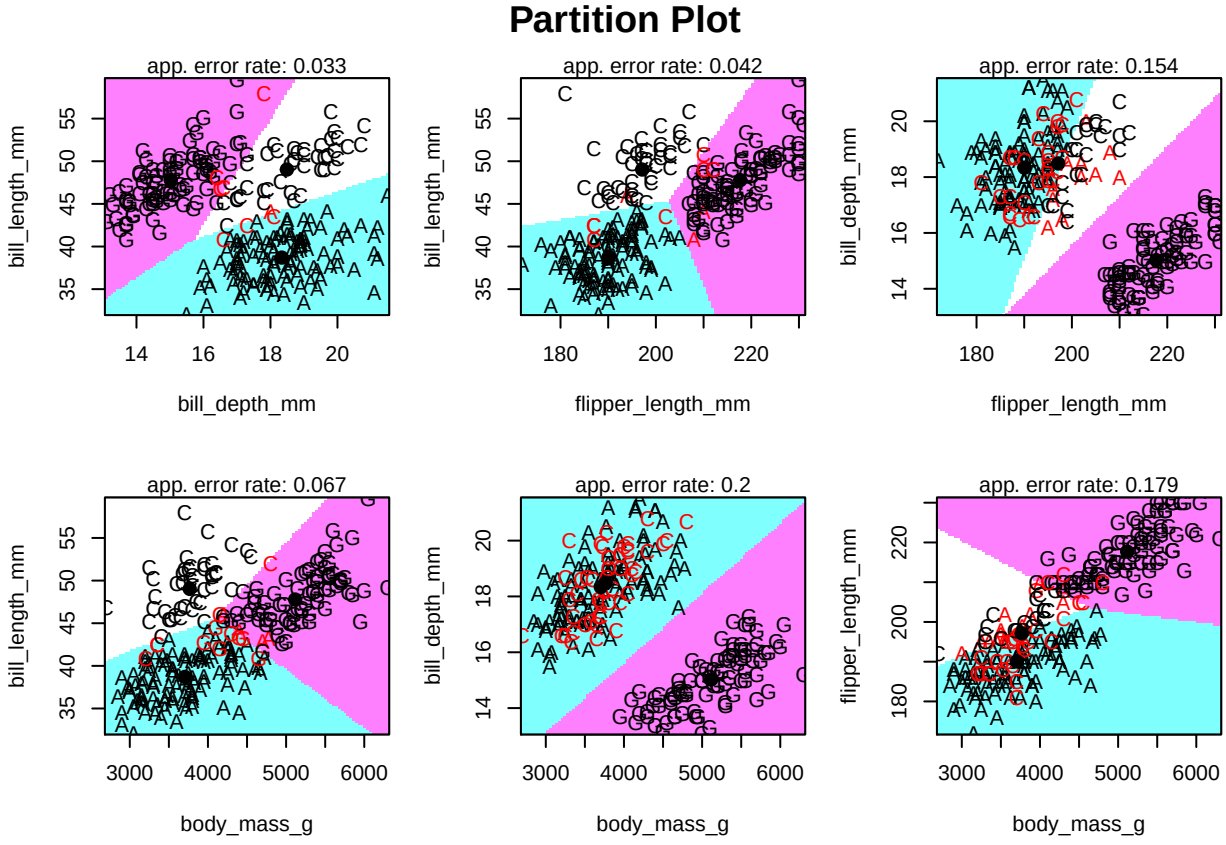


Bi-plots

```
ggord(lda_2, penguine_train$species, ylim = c(-5,5), xlim = c(-8,8))
```




```
partimat(species ~ ., data = penguine_train, method = 'lda')
```



Partition Plot

This plot gives classification of each and every observation in the training dataset based on LDA method.

Confusion matrix and accuracy training data LDA Model 1 with all variables

```
p1_train_all <- predict(lda_all, penguine_train)$class
tab_train_all <- table(Predicted = p1_train_all, Actual = penguine_train$species)
tab_train_all
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       105         2      0
##   Chinstrap      1        46      0
##   Gentoo         0         0     86
```

There are 3 mis-classification occures in Chinstrap.

```
lda_train_accuracy_all <- sum(diag(tab_train_all))/sum(tab_train_all) * 100
lda_train_accuracy_all
```

```
## [1] 98.75
```

Accuracy in training data : 98.75

Confusion matrix and accuracy for test data LDA Model 1 with all variables

```
p1_test_all <- predict(lda_all, penguine_test)$class
tab_test_all <- table(Predicted = p1_test_all, Actual = penguine_test$species)
tab_test_all
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie        44         1      0
##   Chinstrap      1        19      0
##   Gentoo         0         0     37
```

There are total 2 mis-classification occures in test data.

```
lda_test_accuracy_all <- sum(diag(tab_test_all))/sum(tab_test_all) * 100
lda_test_accuracy_all
```

```
## [1] 98.03922
```

Accuracy in test data : 98.0392157

Confusion matrix and accuracy training data (LDA Model 2) LDA Model 2 with two variables

```
p1_train_2 <- predict(lda_2, penguine_train)$class
tab_train_2 <- table(Predicted = p1_train_2, Actual = penguine_train$species)
tab_train_2
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       105         3      0
##   Chinstrap      1        41      0
##   Gentoo         0         4     86
```

There are 8 mis-classification occures in Chinstrap and Adile.

```
lda_train_accuracy_2 <- sum(diag(tab_train_2))/sum(tab_train_2) * 100
lda_train_accuracy_2
```

```
## [1] 96.66667
```

Accuracy in training data : 96.6666667

Confusion matrix and accuracy for test data (LDA model 2) LDA Model 2 with two variables

```
p1_test_2 <- predict(lda_2, penguine_test)$class
tab_test_2 <- table(Predicted = p1_test_2, Actual = penguine_test$species)
tab_test_2
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie        44         2      0
##   Chinstrap      1        18      1
##   Gentoo         0         0     36
```

There are total 4 mis-classification occures in test data.

```
lda_test_accuracy_2 <- sum(diag(tab_test_2))/sum(tab_test_2) * 100
lda_test_accuracy_2
```

```
## [1] 96.07843
```

Accuracy in test data : 96.0784314

Quadratic Discriminant Analysis

(*a*) Same steps as above to consider

For QDA use same `MASS` package to perform analysis.

First we will build the model, then calculate the prediction of train and test data and accuracy. QDA will create 2 models i.e one with 4 variable and another one with bill length and bill depth.

QDA model building With all 4 variables

```r
qda_all <- qda(species ~ ., data = penguine_train)
qda_all
```

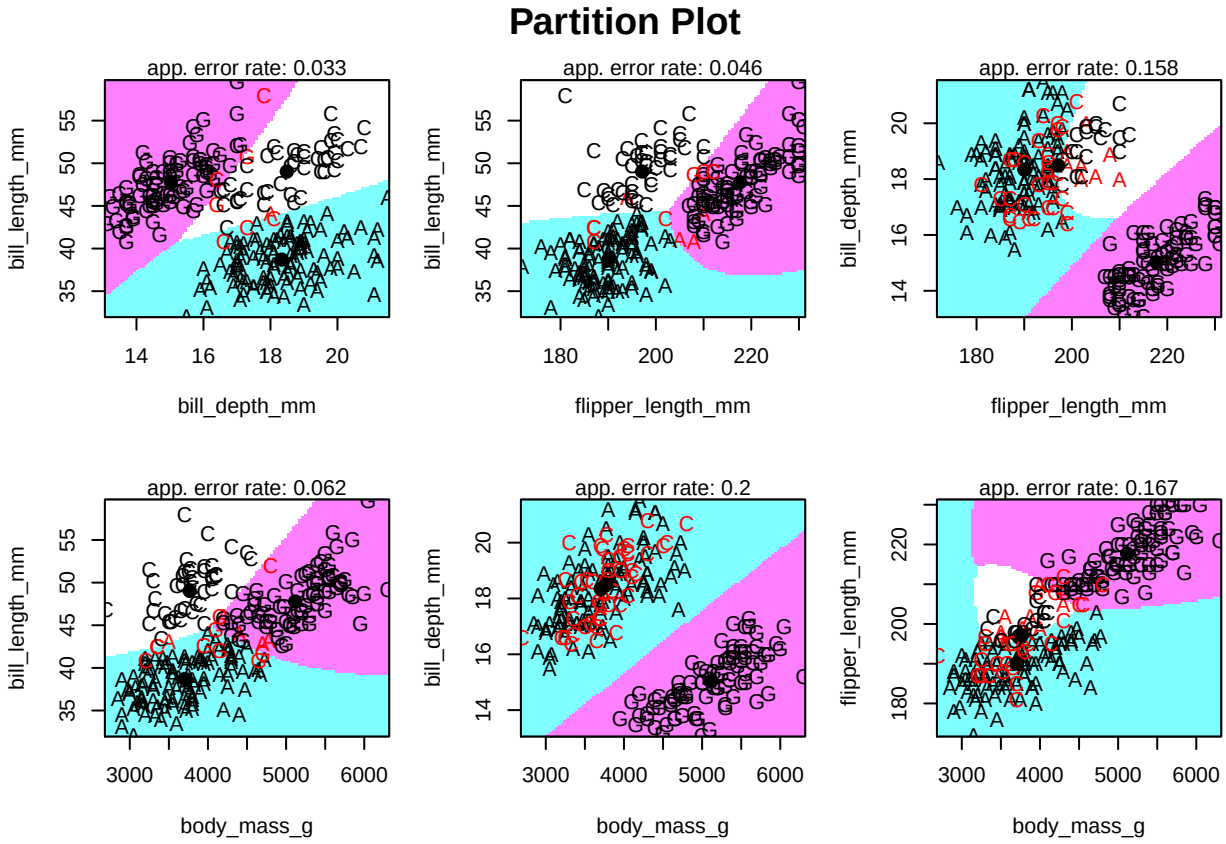
```
## Call:
## qda(species ~ ., data = penguine_train)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4416667 0.2000000 0.3583333
##
## Group means:
##           bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## Adelie          38.70283      18.33585          190.0566    3704.481
## Chinstrap       49.02708      18.48958          197.1250    3766.667
## Gentoo          47.78023      15.02558          217.6860    5116.279
```

With 2 variables

```r
qda_2 <- qda(species ~ bill_length_mm + bill_depth_mm, data = penguine_train)
qda_2
```

```
## Call:
## qda(species ~ bill_length_mm + bill_depth_mm, data = penguine_train)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4416667 0.2000000 0.3583333
##
## Group means:
##           bill_length_mm bill_depth_mm
## Adelie          38.70283      18.33585
## Chinstrap       49.02708      18.48958
## Gentoo          47.78023      15.02558
```

```r
partimat(species ~ ., data = penguine_train, method = 'qda')
```



Partition Plot

This plot gives classification of each and every observation in the training dataset based on QDA method.

Confusion matrix and accuracy training data QDA Model with all variables

```
p2_train_all <- predict(qda_all, penguine_train)$class
tab2_train_all <- table(Predicted = p2_train_all, Actual = penguine_train$species)
tab2_train_all
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       105         2      0
##   Chinstrap      1        46      0
##   Gentoo         0         0     86
```

There are 3 mis-classification occures in Chinstrap.

```
qda_train_accuracy_all <- sum(diag(tab2_train_all))/sum(tab2_train_all) * 100
qda_train_accuracy_all
```

```
## [1] 98.75
```

Accuracy in training data : 98.75

Confusion matrix and accuracy for test data QDA Model with all variables

```
p2_test_all <- predict(qda_all, penguine_test)$class
tab2_test_all <- table(Predicted = p2_test_all, Actual = penguine_test$species)
tab2_test_all
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie        44         1      0
##   Chinstrap      1        19      0
##   Gentoo         0         0     37
```

There are total 2 mis-classification occures in test data.

```
qda_test_accuracy_all <- sum(diag(tab2_test_all))/sum(tab2_test_all) * 100
qda_test_accuracy_all
```

```
## [1] 98.03922
```

Accuracy in test data : 98.0392157

Confusion matrix and accuracy training data (QDA Model 2) QDA Model 2 with two variables

```
p2_train_2 <- predict(qda_2, penguine_train)$class
tab2_train_2 <- table(Predicted = p2_train_2, Actual = penguine_train$species)
tab2_train_2
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       105         3      0
##   Chinstrap      1        43      2
##   Gentoo         0         2     84
```

There are 8 mis-classification occures in all species.

```
qda_train_accuracy_2 <- sum(diag(tab2_train_2))/sum(tab2_train_2) * 100
qda_train_accuracy_2
```

```
## [1] 96.66667
```

Accuracy in training data : 96.6666667

Confusion matrix and accuracy for test data (QDA model 2) QDA Model 2 with two variables

```
p2_test_2 <- predict(qda_2, penguine_test)$class
tab2_test_2 <- table(Predicted = p2_test_2, Actual = penguine_test$species)
tab2_test_2
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie        44         1      0
##   Chinstrap      1        19      1
##   Gentoo         0         0     36
```

There are total 3 mis-classification occures in test data.

```
qda_test_accuracy_2 <- sum(diag(tab2_test_2))/sum(tab2_test_2) * 100
qda_test_accuracy_2
```

```
## [1] 97.05882
```

Accuracy in test data : 97.0588235

Naive Bayes

(a) Same steps as above to consider

Naive Bayes algorithem is based on Bayes theorm. Mathematical expression :

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

To develop a naive bayes classigication model we need to make sure that the independant variables are not highly co-related. From EDA, we see there are co-relation exist between flipp_length and body mass. So exclude flipper length variable for NB model.

```
NB <- naive_bayes(species ~ bill_length_mm + bill_depth_mm + body_mass_g, data = penguine_train)
NB
```

```
##
## ================================== Naive Bayes ==================================
##
##  Call:
## naive_bayes.formula(formula = species ~ bill_length_mm + bill_depth_mm +
##     body_mass_g, data = penguine_train)
##
## ---------------------------------------------------------------------------------
##
## Laplace smoothing: 0
##
## ---------------------------------------------------------------------------------
##
##  A priori probabilities:
##
##    Adelie Chinstrap    Gentoo
## 0.4416667 0.2000000 0.3583333
##
## ---------------------------------------------------------------------------------
##
##  Tables:
##
## ---------------------------------------------------------------------------------
##  ::: bill_length_mm (Gaussian)
## ---------------------------------------------------------------------------------
##
## bill_length_mm    Adelie Chinstrap    Gentoo
```

```
##           mean 38.702830 49.027083 47.780233
##           sd    2.767308  3.519232  3.343292
##
## ------------------------------------------------------------------------------
##  ::: bill_depth_mm (Gaussian)
## ------------------------------------------------------------------------------
##
## bill_depth_mm    Adelie Chinstrap    Gentoo
##          mean 18.335849 18.489583 15.025581
##          sd    1.284352  1.252443  1.031297
##
## ------------------------------------------------------------------------------
##  ::: body_mass_g (Gaussian)
## ------------------------------------------------------------------------------
##
## body_mass_g    Adelie Chinstrap    Gentoo
##        mean 3704.4811 3766.6667 5116.2791
##        sd    490.3691  390.1491  516.9021
##
## ------------------------------------------------------------------------------
```

Confusion matrix and accuracy for train data Calculate Confusion Matrix and accuracy for training data using NB model

```
p3_train <- predict(NB, penguine_train)
tab3_train <- table(Predicted = p3_train, Actual = penguine_train$species)
tab3_train
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       104         4      0
##   Chinstrap      2        44      0
##   Gentoo         0         0     86
```

There are 6 mis-classification occures in train data.

```
NB_train_accuracy <- sum(diag(tab3_train))/sum(tab3_train) * 100
NB_train_accuracy
```

```
## [1] 97.5
```

Accuracy in training data : 97.5

Confusion matrix and accuracy for test data Calculate Confusion Matrix and accuracy for training data using NB model

```
p3_test <- predict(NB, penguine_test)
tab3_test <- table(Predicted = p3_test, Actual = penguine_test$species)
tab3_test
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie        43         2      0
##   Chinstrap      2        18      0
##   Gentoo         0         0     37
```

There are 4 mis-classification occured test data.

```
NB_test_accuracy <- sum(diag(tab3_test))/sum(tab3_test) * 100
NB_test_accuracy
```

```
## [1] 96.07843
```

Accuracy in test data : 96.0784314

(*d.*) **Comment on the models fits/strength/weakness/accuracy for all these three**

models that you worked with

We find out confusion matrix and accuracy of all 5 models. Compair all model based on F1, Sensitivity and Specificity.

Matrix result of all 5 models

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Pr
Class: Adelie	0.9777778	0.9824561	0.9777778	0.9824561	0.9777778	0.9777778	0.9777778	0
Class: Chinstrap	0.9500000	0.9878049	0.9500000	0.9878049	0.9500000	0.9500000	0.9500000	0
Class: Gentoo	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	0

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Pr
Class: Adelie	0.9565217	0.9821429	0.9777778	0.9649123	0.9777778	0.9565217	0.9670330	0
Class: Chinstrap	0.9000000	0.9756098	0.9000000	0.9756098	0.9000000	0.9000000	0.9000000	0
Class: Gentoo	1.0000000	0.9848485	0.9729730	1.0000000	0.9729730	1.0000000	0.9863014	0

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Pr
Class: Adelie	0.9777778	0.9824561	0.9777778	0.9824561	0.9777778	0.9777778	0.9777778	0
Class: Chinstrap	0.9500000	0.9878049	0.9500000	0.9878049	0.9500000	0.9500000	0.9500000	0
Class: Gentoo	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	0

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Pr
Class: Adelie	0.9777778	0.9824561	0.9777778	0.9824561	0.9777778	0.9777778	0.9777778	0
Class: Chinstrap	0.9047619	0.9876543	0.9500000	0.9756098	0.9500000	0.9047619	0.9268293	0
Class: Gentoo	1.0000000	0.9848485	0.9729730	1.0000000	0.9729730	1.0000000	0.9863014	0

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Pr
Class: Adelie	0.9555556	0.9649123	0.9555556	0.9649123	0.9555556	0.9555556	0.9555556	0
Class: Chinstrap	0.9000000	0.9756098	0.9000000	0.9756098	0.9000000	0.9000000	0.9000000	0
Class: Gentoo	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	0

From matrix result we see, LDA and QDA performed well than Naive Bayes. Naive Bayes performed well in normalized data, however this dataset is not normalized. I will go the QDA model 2 due to high accuract, specificity, Sensitivity, and F1.

References:

LDA: https://www.youtube.com/watch?v=WUCnHx0QDSI

Naive bayes: https://www.youtube.com/watch?v=RLjSQdcg8AM&list=RDCMUCuWECsa_za4gm7B3TLgeV_A&index=4