DATA MINING ON ABALON DATASET

# Introduction

Abalone are marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures. The economic value of abalone is positively correlated with its age and gender. Therefore, to detect the age of abalone is important for both farmers and customers to determine its price. Farmers usually cut the shells and count the rings. This complex method increases the cost and limits its popularity. Our target is to find out the best indicators and the best model to classify and the age and gender of the abalones.

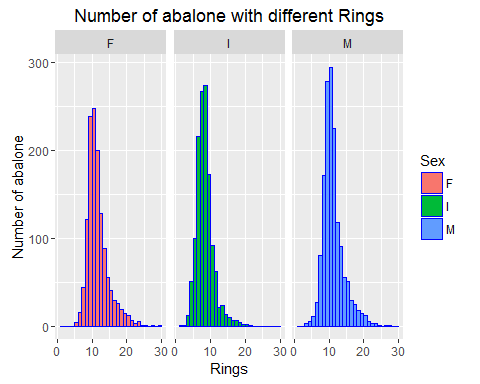
# Description of the Dataset

We are going to predict the age of the Abalone from its physical measurements. The following dataset has been downloaded from UCI repository. There are namely 9 variables in the dataset as shown below. Age of Abalone is 2+ number of rings in the Abalone.

## 'data.frame': 4176 obs. of 9 variables:  
## $ Sex : Factor w/ 3 levels "F","I","M": 3 1 3 2 2 1 1 3 1 1 ...  
## $ Length : num 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 0.525 ...  
## $ Diameter : num 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 0.38 ...  
## $ Height : num 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 0.14 ...  
## $ Whole.Weight : num 0.226 0.677 0.516 0.205 0.351 ...  
## $ Shucked.Weight: num 0.0995 0.2565 0.2155 0.0895 0.141 ...  
## $ Viscera.Weight: num 0.0485 0.1415 0.114 0.0395 0.0775 ...  
## $ Shell.Weight : num 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 0.21 ...  
## $ Rings : int 7 9 10 7 8 20 16 9 19 14 ...

# Data Cleaning

We will plot the number of Abalone with different number of rings

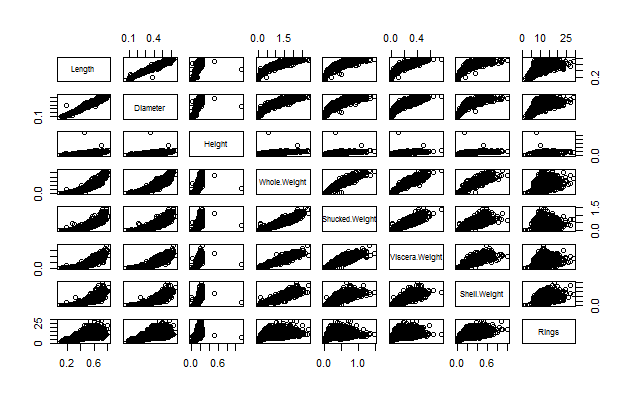


From the above plot we can see the median for class “F” lies somewhere between 10-11, for Infants it is around 8 and around 13-14 for Male. Considering the Range of Rings is between 1-30, we have decided to group Abalones with less than 6 rings, between 6 -13 rings and more than 13 rings indicating young, adult and old Abalones respectively.

We will create an Age variable in the dataset. This is a factor variable with 3 different levels where 1= Infant; 2= Adult and 3= Old respectively. The dataset is shown as below

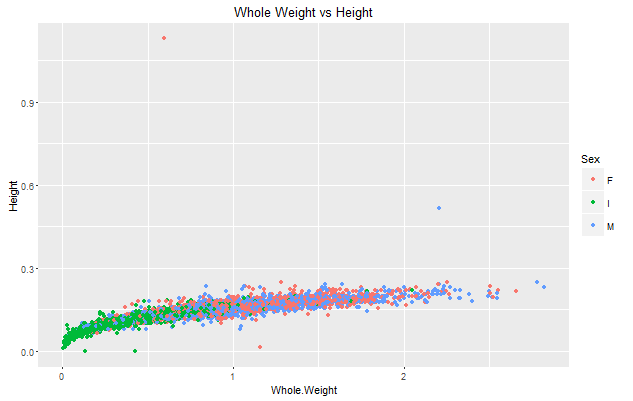
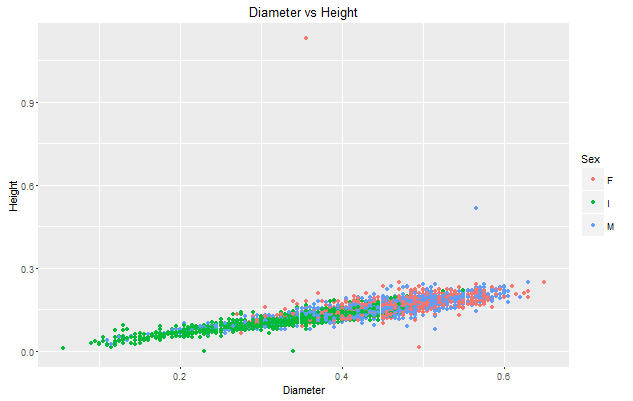
## Sex Length Diameter Height Whole.Weight Shucked.Weight Viscera.Weight  
## 1 M 0.350 0.265 0.090 0.2255 0.0995 0.0485  
## 2 F 0.530 0.420 0.135 0.6770 0.2565 0.1415  
## 3 M 0.440 0.365 0.125 0.5160 0.2155 0.1140  
## 4 I 0.330 0.255 0.080 0.2050 0.0895 0.0395  
## 5 I 0.425 0.300 0.095 0.3515 0.1410 0.0775  
## Shell.Weight Rings Age  
## 1 0.070 7 2  
## 2 0.210 9 2  
## 3 0.155 10 2  
## 4 0.055 7 2  
## 5 0.120 8 2  
## 6 0.330 20 3

Let us check the plots between the feature vectors



We can make the following observations

* We can see some outliers between Length Vs Height and Height Vs whole Weight . Let us see them closely
* We can see very high linear trends between the predictor Features. We have to check the correlations



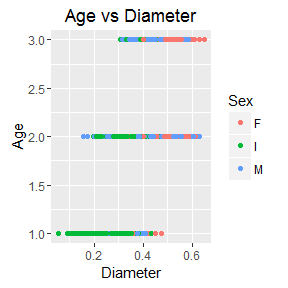
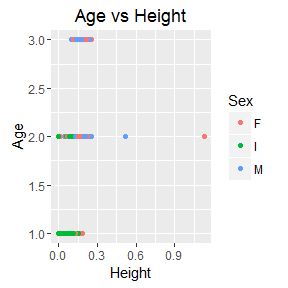
We see some outliers in the dataset. Let us remove them in the new dataset

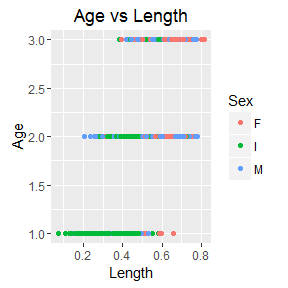
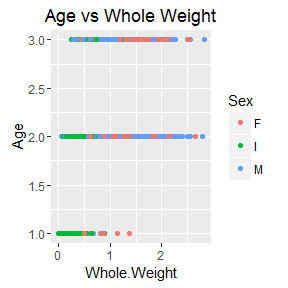
Let us check the correlation between the feature vectors as they appear to be on the high side

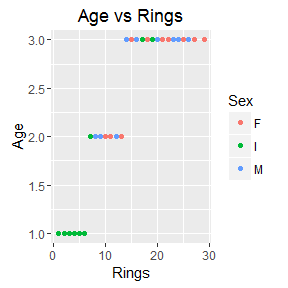
## Length Diameter Height Whole.Weight Shucked.Weight  
## Length 1.0000000 0.9868132 0.8275521 0.9252551 0.8979052  
## Diameter 0.9868132 1.0000000 0.8337053 0.9254520 0.8931591  
## Height 0.8275521 0.8337053 1.0000000 0.8192087 0.7749568  
## Whole.Weight 0.9252551 0.9254520 0.8192087 1.0000000 0.9694027  
## Shucked.Weight 0.8979052 0.8931591 0.7749568 0.9694027 1.0000000  
## Viscera.Weight 0.9030105 0.8997262 0.7982929 0.9663721 0.9319557  
## Shell.Weight 0.8976970 0.9053281 0.8173261 0.9553511 0.8826063  
## Rings 0.5571226 0.5750054 0.5581087 0.5408179 0.4212556  
## Viscera.Weight Shell.Weight Rings  
## Length 0.9030105 0.8976970 0.5571226  
## Diameter 0.8997262 0.9053281 0.5750054  
## Height 0.7982929 0.8173261 0.5581087  
## Whole.Weight 0.9663721 0.9553511 0.5408179  
## Shucked.Weight 0.9319557 0.8826063 0.4212556  
## Viscera.Weight 1.0000000 0.9076469 0.5042735  
## Shell.Weight 0.9076469 1.0000000 0.6280306  
## Rings 0.5042735 0.6280306 1.0000000

We can see very high correlations so we have to standardize the dataset. We will use the scale function and standardize the dataset. However standardization does not lower the correlations as shown above. Hence we will continue with the same dataset.

# Exploratory Data Analysis







# Finding the best regression Model

We will find the best regression model to classify age. We will regress age on all the possible predicting features and will try to find the best prediction model based on AIC and RSS values using backward deletion method.

The output is given as below

## Start: AIC=-6794.78  
## Age ~ Length + Diameter + Height + Whole.Weight + Viscera.Weight +   
## Shell.Weight + Rings  
##   
## Df Sum of Sq RSS AIC  
## - Viscera.Weight 1 0.040 189.42 -6796.2  
## - Shell.Weight 1 0.065 189.44 -6795.9  
## - Diameter 1 0.130 189.51 -6795.0  
## <none> 189.38 -6794.8  
## - Height 1 0.431 189.81 -6790.9  
## - Length 1 0.993 190.37 -6783.2  
## - Whole.Weight 1 1.218 190.59 -6780.1  
## - Rings 1 209.073 398.45 -4862.8  
##   
## Step: AIC=-6796.23  
## Age ~ Length + Diameter + Height + Whole.Weight + Shell.Weight +   
## Rings  
##   
## Df Sum of Sq RSS AIC  
## - Shell.Weight 1 0.045 189.46 -6797.6  
## - Diameter 1 0.133 189.55 -6796.4  
## <none> 189.42 -6796.2  
## - Height 1 0.411 189.83 -6792.6  
## - Length 1 0.975 190.39 -6784.9  
## - Whole.Weight 1 3.299 192.72 -6753.3  
## - Rings 1 209.239 398.66 -4863.4  
##   
## Step: AIC=-6797.61  
## Age ~ Length + Diameter + Height + Whole.Weight + Rings  
##   
## Df Sum of Sq RSS AIC  
## - Diameter 1 0.121 189.58 -6798.0  
## <none> 189.46 -6797.6  
## - Height 1 0.377 189.84 -6794.4  
## - Length 1 1.036 190.50 -6785.4  
## - Whole.Weight 1 9.039 198.50 -6678.4  
## - Rings 1 244.407 433.87 -4645.4  
##   
## Step: AIC=-6797.95  
## Age ~ Length + Height + Whole.Weight + Rings  
##   
## Df Sum of Sq RSS AIC  
## <none> 189.58 -6798.0  
## - Height 1 0.476 190.06 -6793.4  
## - Length 1 8.019 197.60 -6692.2  
## - Whole.Weight 1 8.935 198.52 -6680.2  
## - Rings 1 249.491 439.07 -4616.4

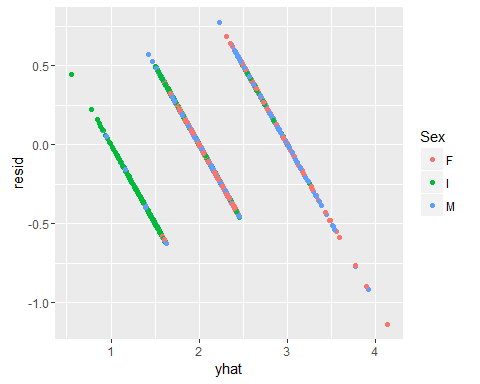
##   
## Call:  
## lm(formula = Age ~ Length + Height + Whole.Weight + Rings, data = train\_data)  
## Coefficients:  
## (Intercept) Length Height Whole.Weight Rings   
## 0.3282 1.3468 0.8707 -0.3295 0.1138

From the above output we get the best regression model on the training dataset. Let us find the accuracy which is denoted by the R^2 values in the training dataset.

Call:  
## lm(formula = Age ~ Length + Height + Whole.Weight + Rings,data=train\_data)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.328216 0.040015 8.202 3.67e-16 \*\*\*  
## Length 1.346769 0.128544 10.477 < 2e-16 \*\*\*  
## Height 0.870683 0.340968 2.554 0.0107 \*   
## Whole.Weight -0.329492 0.029794 -11.059 < 2e-16 \*\*\*  
## Rings 0.113837 0.001948 58.438 < 2e-16 \*\*\*  
## ---  
  
## Residual standard error: 0.2703 on 2595 degrees of freedom  
## Multiple R-squared: 0.7124, Adjusted R-squared: 0.7119   
## F-statistic: 1607 on 4 and 2595 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Response: Age  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Length 1 187.548 187.548 2567.2 < 2.2e-16 \*\*\*  
## Height 1 23.364 23.364 319.8 < 2.2e-16 \*\*\*  
## Whole.Weight 1 9.184 9.184 125.7 < 2.2e-16 \*\*\*  
## Rings 1 249.491 249.491 3415.0 < 2.2e-16 \*\*\*  
## Residuals 2595 189.582 0.073   
## --- 659.169  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

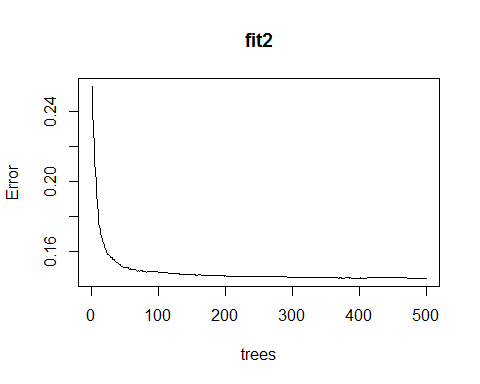
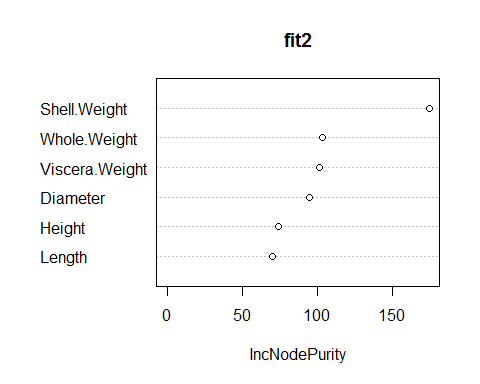
We can see that 71.24% of the variability is explained in age by the predictors and this model is 71.24% accurate having an error of 659.169. This regression is explained by the plot below where we can see a linear trend showing regression is significant.



# Finding the best Predictors to classify Age and Sex using random forest

Let us find the best predictors to classify the dataset

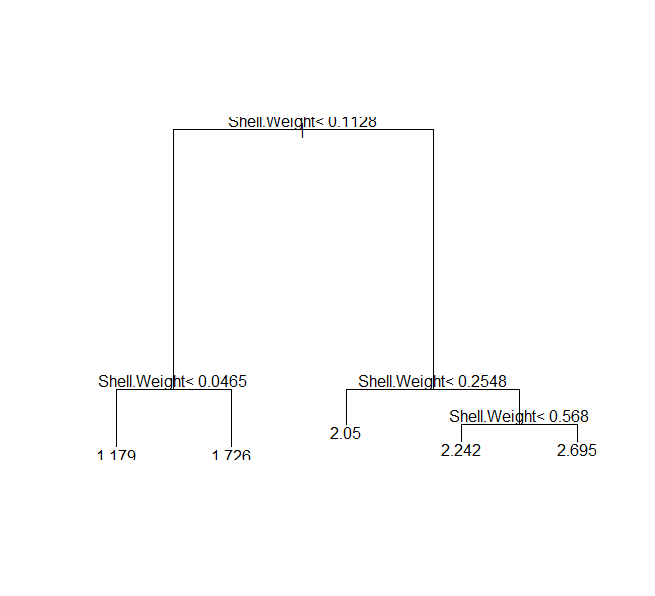
## Based on Age



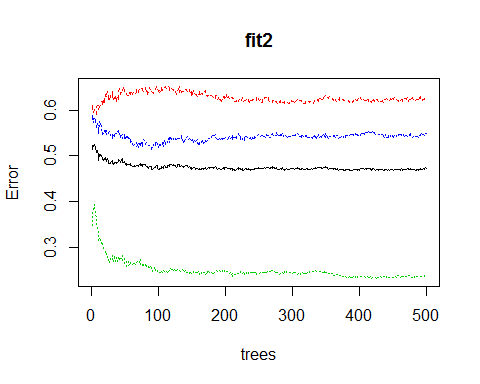
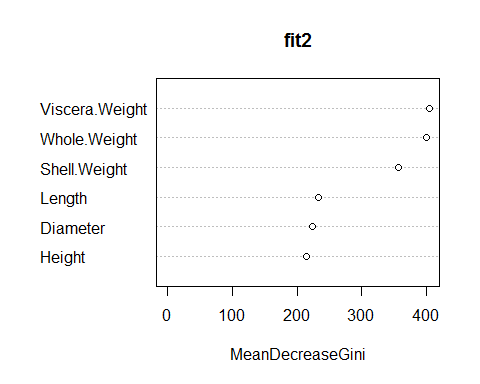
Here fit2 is the random forest of age regressed on all the predictor features from where we can see that Shell Weight is the most important variable followed by Whole Weight and so on. We can also see the error in the model reduces exponentially as the number of trees are increased, hence increasing model accuracy.

The output on the training dataset is given below

## Regression tree:  
## rpart(formula = Age ~ Length + Height + Whole.Weight + Diameter +   
## Viscera.Weight + Shell.Weight, data = train\_data, control = rpart.control(cp = 0.01))  
##   
## Variables actually used in tree construction:  
## [1] Length Shell.Weight  
##   
## Root node error: 677.02/2757 = 0.24556  
##   
## n= 2757   
##   
## CP nsplit rel error xerror xstd  
## 1 0.265843 0 1.00000 1.00065 0.033355  
## 2 0.061450 1 0.73416 0.76283 0.022810  
## 3 0.033257 2 0.67271 0.70605 0.022212  
## 4 0.016797 3 0.63945 0.66419 0.020638  
## 5 0.010088 4 0.62265 0.65434 0.020937  
## 6 0.010000 5 0.61257 0.64755 0.020523



## Based on gender



Here fit2 is the random forest of Sex regressed on all the predictor features from where we can see that Viscera Weight is the most important variable followed by Whole Weight and so on. We can also see the error in the model reduces exponentially for Infants as the number of trees are increased, however it decreases till 50 trees then remain constant for all other 2 hence model accuracy is constant.

## RANDOM FOREST FROM R

# E:\semester 2\IE-5318\project 2\Rplot18.png

# Random forest on Age using Weka

Let us find the accuracy and confusion matrix

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation:     New Microsoft Excel Worksheet123456-weka.filters.unsupervised.attribute.Remove-R8

Instances:    4177

Attributes:   8

              Length

              Diameter

              Height

              Whole weight

              Shucked weight

              Viscera weight

              Shell weight

              Age

Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 2.43 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0.17 seconds

=== Summary ===

Correctly Classified Instances        1173               82.6056 %

Incorrectly Classified Instances       247               17.3944 %

Kappa statistic                          0.4479

Mean absolute error                      0.1576

Root mean squared error                  0.2851

Relative absolute error                 63.7949 %

Root relative squared error             82.3379 %

Total Number of Instances             1420

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class

                 0.545    0.030    0.672      0.545    0.602      0.567    0.943     0.723     A

                 0.931    0.559    0.859      0.931    0.894      0.429    0.825     0.933     B

                 0.348    0.031    0.589      0.348    0.437      0.402    0.861     0.503     C

Weighted Avg.    0.826    0.446    0.810      0.826    0.813      0.440    0.841     0.863

=== Confusion Matrix ===

    a    b    c   <-- classified as

   78   65    0 |    a = A

   38 1039   39 |    b = B

    0  105   56 |    c = C

# Random forest on Gender using Weka

Let us find the accuracy and the confusion matrix

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation:     avalone

Instances:    4177

Attributes:   9

              Sex

              Length

              Diameter

              Height

              Whole weight

              Shucked weight

              Viscera weight

              Shell weight

              Rings

Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 3.06 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0.18 seconds

=== Summary ===

Correctly Classified Instances         803               56.5493 %

Incorrectly Classified Instances       617               43.4507 %

Kappa statistic                          0.3462

Mean absolute error                      0.336

Root mean squared error                  0.4162

Relative absolute error                 75.7653 %

Root relative squared error             88.3868 %

Total Number of Instances             1420

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class

                 0.505    0.304    0.489      0.505    0.497      0.199    0.669     0.496     M

                 0.419    0.225    0.466      0.419    0.441      0.200    0.715     0.480     F

                 0.783    0.129    0.737      0.783    0.760      0.645    0.902     0.802     I

Weighted Avg.    0.565    0.224    0.560      0.565    0.562      0.340    0.757     0.587

=== Confusion Matrix ===

   a   b   c   <-- classified as

 262 180  77 |   a = M

 215 190  48 |   b = F

  59  38 351 |   c = I

# Classification Using LDA

Linear Discriminant analysis

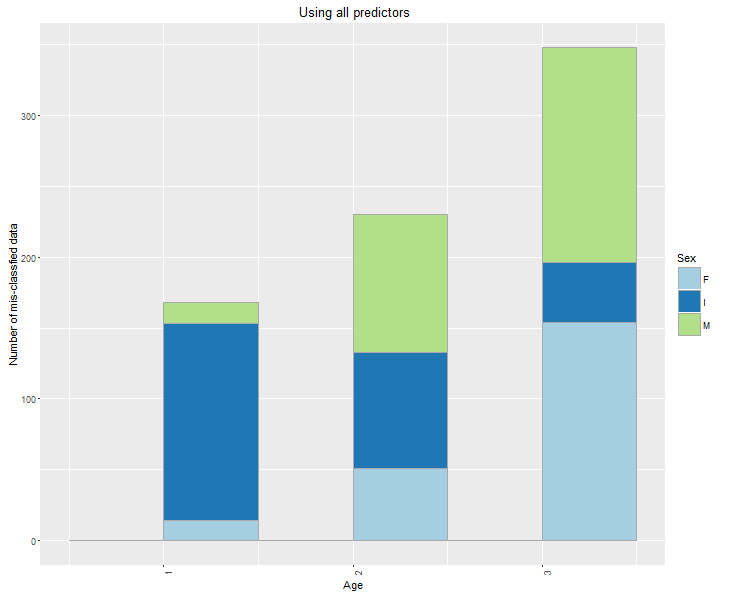
Linear Discriminant Analysis (LDA) was formulated by Ronald A. Fisher in 1936. It is a classification methodology. It is used to find a linear combination of features and classify objects.

It is simple and mathematically robust. It produces models whose accuracy is equally good as the more complex methods. Its concept is such that it searches for a linear combination of variables or predictors that separate the classes.

We first calculate the mean vectors, covariance matrices and class probabilities and then calculate pooled covariance matrix and finally the coefficients of the linear model.

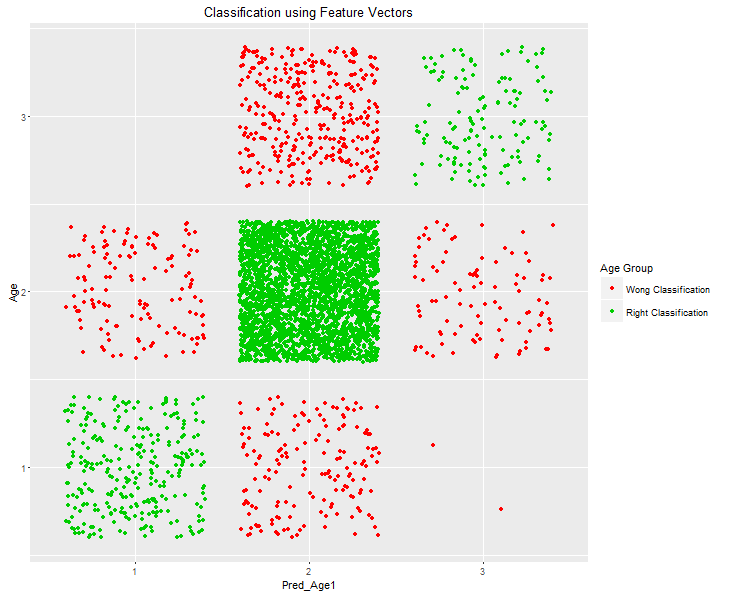
## Based on Age

Let us regress age on training dataset and find the classification plot on the testing dataset, we get



The above plot shows the classification of Age using the model on the testing dataset where Male , Infants and Female are classified according to the model. The interesting fact is that the column 1 should be entirely dark blue in color which shows there is misclassification of data.

# Plot for the accuracy of classification



From the above plot we can see that the accuracy of the LDA classification which gives a decent percentage of misclassification. Let us find the percentage of misclassification.

Call:  
## lda(Age ~ Length + Diameter + Height + Whole.Weight + Viscera.Weight +   
## Shell.Weight, data = gData)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.1073311 0.7755151 0.1171538   
##   
## Group means:  
## Length Diameter Height Whole.Weight Viscera.Weight Shell.Weight  
## 1 0.3214621 0.2405134 0.07934152 0.1987165 0.04262946 0.05736049  
## 2 0.5421872 0.4221501 0.14305684 0.8688054 0.19066559 0.24512573  
## 3 0.5890082 0.4666360 0.16850716 1.1390389 0.23998671 0.36325460  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Length 7.969149 -15.0665489  
## Diameter 10.311185 -0.7794949  
## Height 16.030201 10.9924667  
## Whole.Weight -3.284777 -2.1906461  
## Viscera.Weight -1.726759 -1.7482654  
## Shell.Weight 2.745354 21.3135980  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.7939 0.2061

The above value of LD1 shows that around 79.39% of the model is accurate and around 20.61% of the model is inaccurate which can be verified by the plots above

## Sex

Call:  
## lda(factor(Sex) ~ Length + Diameter + Height + Whole.Weight +   
## Viscera.Weight + Shell.Weight, data = gData)  
##   
## Prior probabilities of groups:  
## F I M   
## 0.3128893 0.3215141 0.3655966   
##   
## Group means:  
## Length Diameter Height Whole.Weight Viscera.Weight Shell.Weight  
## F 0.5791884 0.4548086 0.1572665 1.0468786 0.23077642 0.3021390  
## I 0.4277459 0.3264940 0.1079955 0.4313625 0.09201006 0.1281822  
## M 0.5613663 0.4392529 0.1511796 0.9909738 0.21544201 0.2819050  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Length 6.6303085 1.666405  
## Diameter -11.7480211 -11.385199  
## Height -11.1533706 -3.055235  
## Whole.Weight 0.4695207 11.617462  
## Viscera.Weight -5.3058845 -22.681957  
## Shell.Weight -0.9104921 -15.654075  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9895 0.0105

## [1] I M I I I M  
## Levels: F I M

We can see that around 98.95 % of the testing data is correctly classified by this algorithm.

**Conclusion**

|  |  |  |
| --- | --- | --- |
| Classification Method | Gender Accuracy | Age Accuracy |
| Linear Regression | 23.66 | 69.86 |
| Linear Discriminant Analysis | 99.34 | 85.34 |
| Decision Tree | 56.54 | 82.60 |

on the analysis above, we can conclude that quantity features of the dataset namely length, Height and Whole.weight (weight) can classify abalone to different age group, especially for Sex. Hence Farmers who weight the abalone and decide the price of the abalone according to size are right. Their technique is correct,