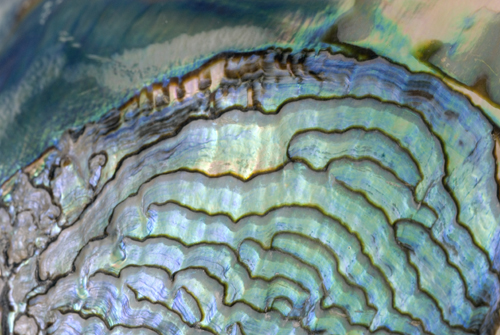
**Data Mining Project Report**

**Gender prediction of Australia Abalone**

****

**Professor: Dr. Chris Rump**

**Students: Prakash Thummisetti, Jing-Yi Wu, Connor Delong**

**Executive Summary:**

Abalone data set is from school, UCI. We wish to predict the sex of abalone. In order to employ some methods, we deleted Infants and kept female and male. In order to randomly select test data, we randomize all data and assigned first 60% of data as training data, 20% of data as validation data and 20% as test data.

For Naivebayes method, we need to check independence assumption. We plotted the correlation matrix to see how variables are correlated. We tried 4 techniques, which are Naive Bayes, Logistic Regression, CaRT, and K- nearest neighbor. In general, all accuracy rates are pretty close.

In the process of model selection, all techniques gave us very close accuracy rate. They are around 51% to 55%. Based on the outputs of 4 techniques, Logistic Regression gave us the best accuracy. We tested the test data set by using Logistic regression. The accuracy rate is 56%.

1. Introduction:

Source: <http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names>

In this data set, there are 9 variables:

Sex nominal -- M, F, and I (infant)

Length continuous mm Longest shell measurement

Diameter continuous mm perpendicular to length

Height continuous mm with meat in shell

Whole weight continuous grams whole abalone

Shucked weight continuous grams weight of meat

Viscera weight continuous grams gut weight (after bleeding)

Shell weight continuous grams after being dried

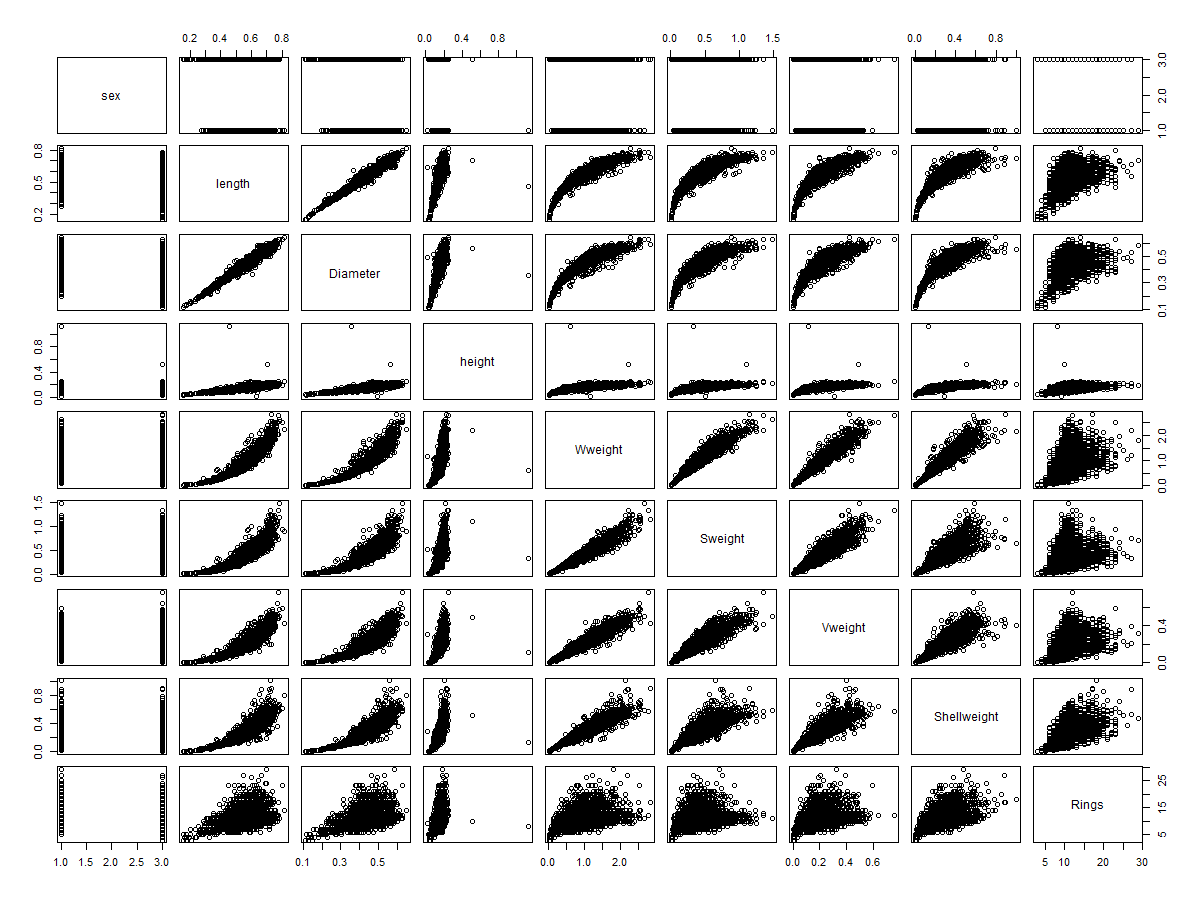
Rings integer -- +1.5 gives the age in years

We want to use Length, Diameter, Height, Whole weight , Shucked weight, viscera weight, and shell weight, and Rings to predict Sex, which is the response variable.

2. Data Modification:

We decided to delete We want to predict the sex of Abalone from other 8 variables. the data of Infant, since we want to utilize the methodology of Logistic Regression. We still can predict the female, male, and Infant by using other methods. To have consistency and comparisons among these methods, we deleted the data of infant. We want to predict the sex of Abalone from other 8 variables. There were 4177 observations before we deleted the data of infant. In total, there are 2835 observations now. We don’t need to transformer any part of the data because we don’t employ any model or technique depending on Normality assumption.

3. Data Exploration:



Since all methods don't require normality assumption, we don't need to transform variables or normalize data. However, the NaiveBayes method requires independence assumption. We plotted correlation matrix for original data to see if some variables are highly correlated. We can see that Length, Diameter, Whole weight, Viscera weight and Shell weight are highly correlated. We decided to delete those variables and used remaining variables.

4. Model:

5. Results:

6. Conclusions:

7. Appendices: