Classification and Pricing of Farmed Abalone

A. Gill

December 12, 2022

Introduction

An enquiry has been made as to the potential use of lightweight software, installed on diving equipment, to help abalone farmers to quickly assess whether an abalone should be harvested or left in the water. There are two main parameters in making such an assessment:

- **Sex.** Female abalone are more valuable, because their eggs are used to reduce water toxicity (Atlantic, 2010). On the other hand, mature males may be sought, to ensure that enough females are left to allow reproduction to occur.
- **Age.** Infant abalone (that have not yet reached sexual maturity) should be avoided, because their small size yields less meat. In addition, they should be allowed to reach maturity and contribute to reproduction.

The aim is to use non-intrusive measurements - length, diameter and height - to predict an abalone's sex and infancy status, as well as estimate its weight (and thus market value), allowing farmers to return rejected abalone to the water without having harmed them. Modelling techniques shall be applied to data from previously harvested abalone, to develop predictive tools for this purpose.

1 Exploratory Analysis

Data on more than 4000 harvested abalone have been supplied. The dataset includes the following columns (variables):

- Sex
- Length, diameter and height
- Whole, shucked, viscera and shell weights
- Number of rings on the shell (an indicator of age)

The data have been checked for missing or impossible values. No entries are missing, but two abalone have a recorded height of 0, a physical impossibility. These two rows have been removed.

The modelling techniques to be used depend on the supplied data meeting certain conditions:

- **Normality.** When the probability of each value occurring is graphed against the value itself, the curve (known as a density plot) should form the classical bell shape expected of "normal" data.
- **Linearity.** The relationship between each pair of variables should be linear (or close to a straight line).
- **Homoscedasticity.** The variation (spread) in one variable should remain the same, regardless of the value of another variable.

1.1 Normality

The density plots for all the numerical variables are shown in Figure 1.

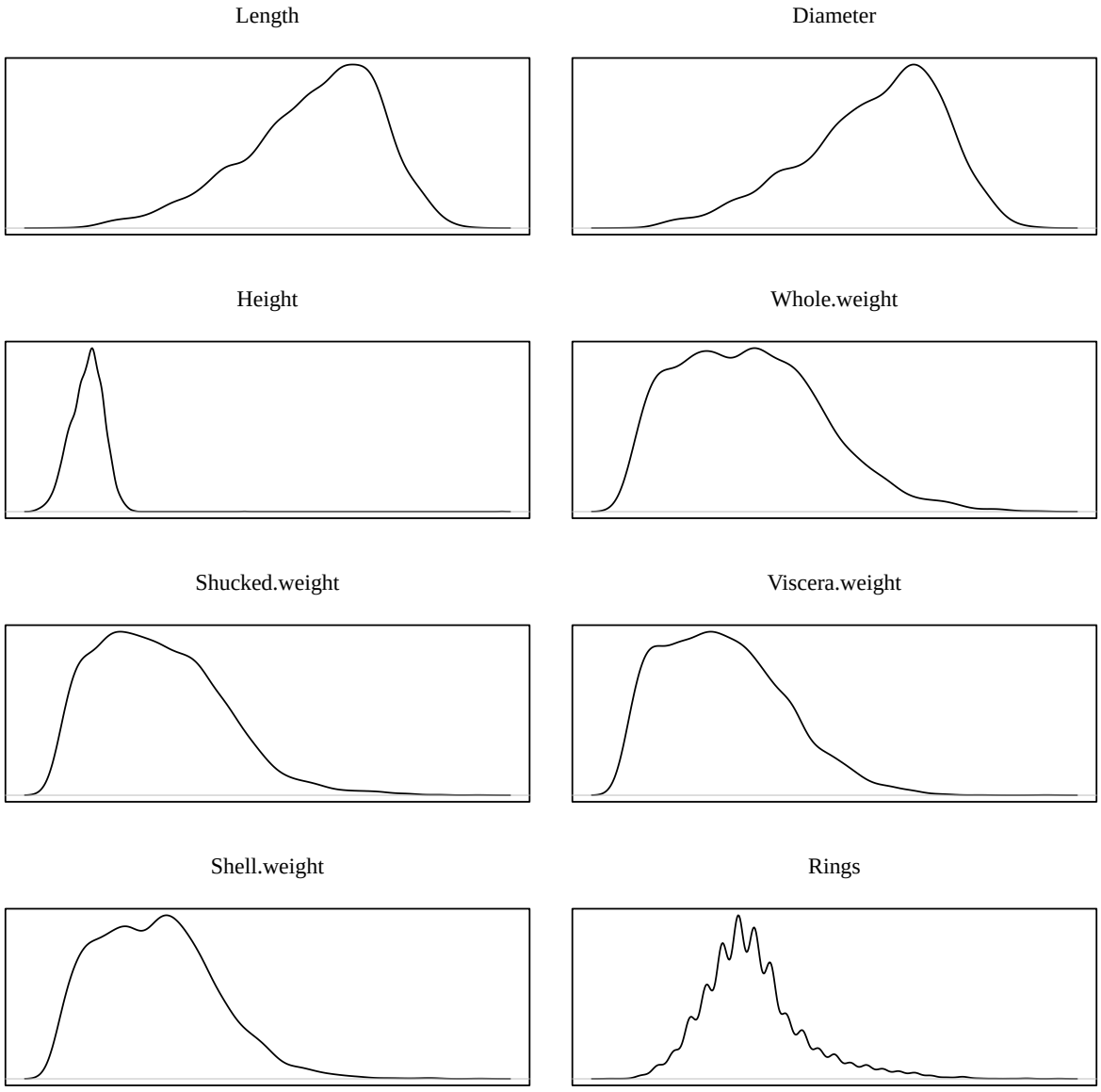

Figure 1: Density Plots

All of the variables are skewed; that is, the bell curve is stretched either left or right, meaning the data may not be sufficiently "normal". This can be repaired, to some extent, by transforming the data i.e. mathematically altering the values to force them into a more symmetrical shape.

1.2 Linearity and Homoscedasticity

These two requirements can be checked by plotting pairs of the variables against one another, as shown in Figure 2. The following observations can be made:

- The relationships between all the weight variables, and the length and diameter, are non-linear.
- As the number of rings increases, the variation (spread) in all the other variables (except height) also increases. However, because the counting of rings is an intrusive process, this

variable is to be excluded from the modelling, and this failure of homoscedasticity can be ignored.

- The slightly oval shape of the plots pairing the four weights suggests that these variables are normally distributed. Examining Figure 1 confirms that the four weight variables are the most "normal", with low kurtosis (pointiness) but some skew to the right.

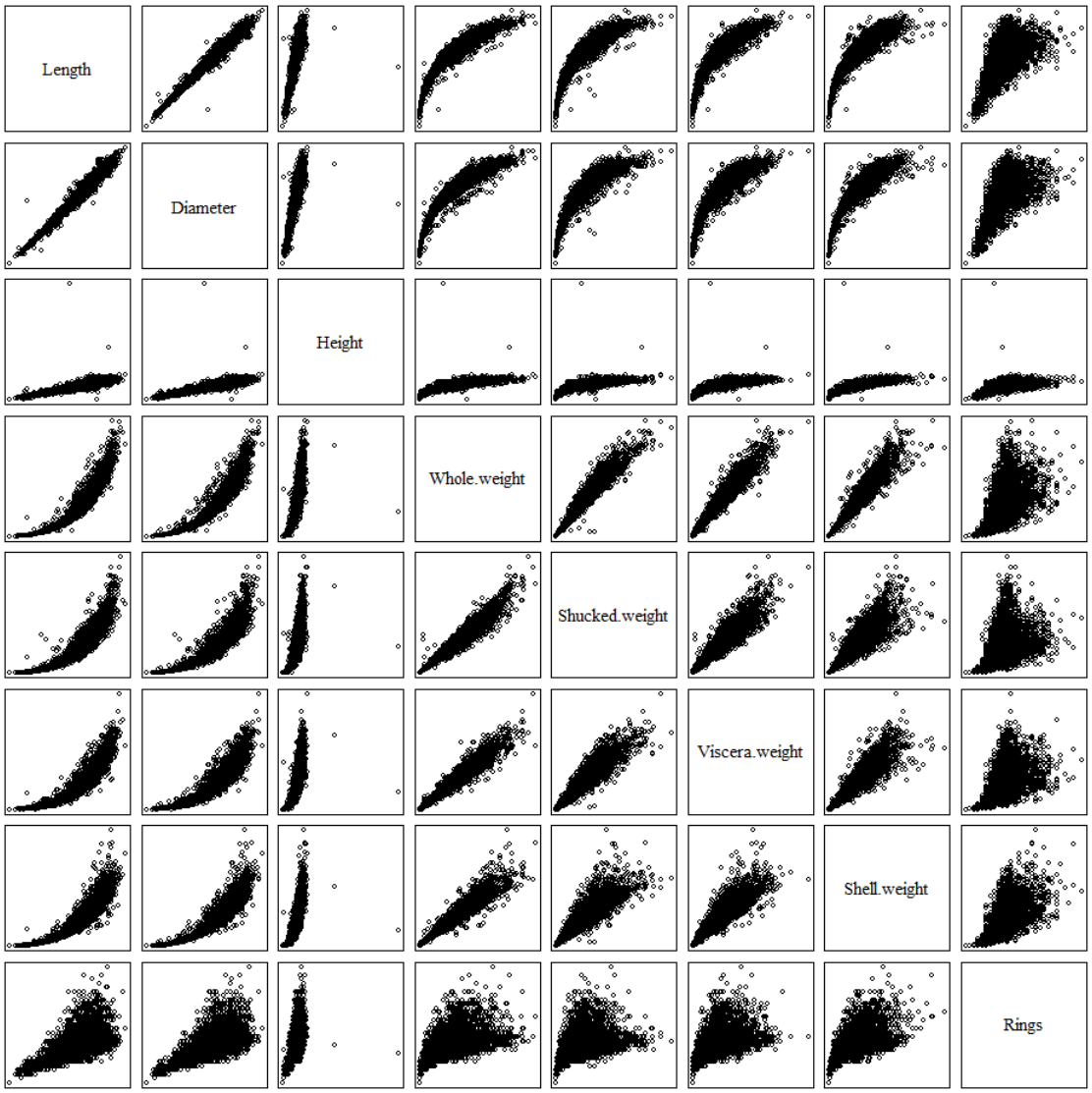

Figure 2: Scatter Plots of Variable Pairs

1.3 Data Correction

As mentioned earlier, the variables can be transformed to normalise them. A transformation can be as simple as taking the square root or logarithm of every value in the data column.

Transformations have the additional benefit of pulling outliers (extreme values) in towards the bell, so that they are no longer considered extreme. Then, any remaining outliers may be removed from the dataset. They can also help to improve the linearity of the relationships between the variables.

Table 1 shows that most outliers have been successfully addressed by the transformations, leaving only a small number to be removed from the dataset (with the exception of height). Figure 3 illustrates a substantial improvement in every variable, after applying a transformation and removing the remaining outliers.

Variable	Initial Outliers	Remaining Outliers
Length	49	8
Diameter	59	9
Height	27	156
Whole weight	30	1
Shucked weight	48	3
Viscera weight	26	2
Shell weight	35	8
Rings	278	268

Table 1: Outliers Before and After Transformations

Having confirmed, through measurements of skewness and kurtosis, that the variables are now much closer to "normal" than they were before, data modelling to predict sex and infancy status can now proceed.

2 Predicting Sex and Infancy

The supplied data have been fed into two machine learning algorithms in order to develop a "classifier" that labels an abalone an infant, female or male, based on non-intrusive inputs (length, diameter and height).

The first algorithm involves **discriminant analysis**; the second involves a **support vector machine**. Detailed explanations for these algorithms can be found in XXXX and XXXX.

2.1 Three-Way Classifier

Two variants of discriminant analysis have been attempted; LDA (linear discriminant analysis) and QDA (quadratic discriminant analysis). LDA assumes that the variance (spread) of values for all variables is the same. However, testing has confirmed that this is not the case for abalone. Nevertheless, LDA may still produce good results and is thus worth considering.

In addition to LDA and QDA, two support vector machines have been attempted; linear and radial. Table 2 reports the accuracy for each of the attempted algorithms.

Model	Accuracy %
QDA	0.5050
LDA	0.5153
SVM linear	0.5147
SVM radial	0.5168

Table 2: Accuracy of Three-Way Classifiers

The overall accuracy for all models is poor, but is slightly better for SVM. Table 3 shows the accuracy of the two SVM models by class (infant, female, male).

Nearly all of the female abalone are classified incorrectly by these models. By contrast, their accuracy is acceptable (above 70%) for infants, and approaching good (nearly 80%) for males. In order to understand this phenomenan, the three input variables can be summarised as boxplots, as shown in Figure 4.

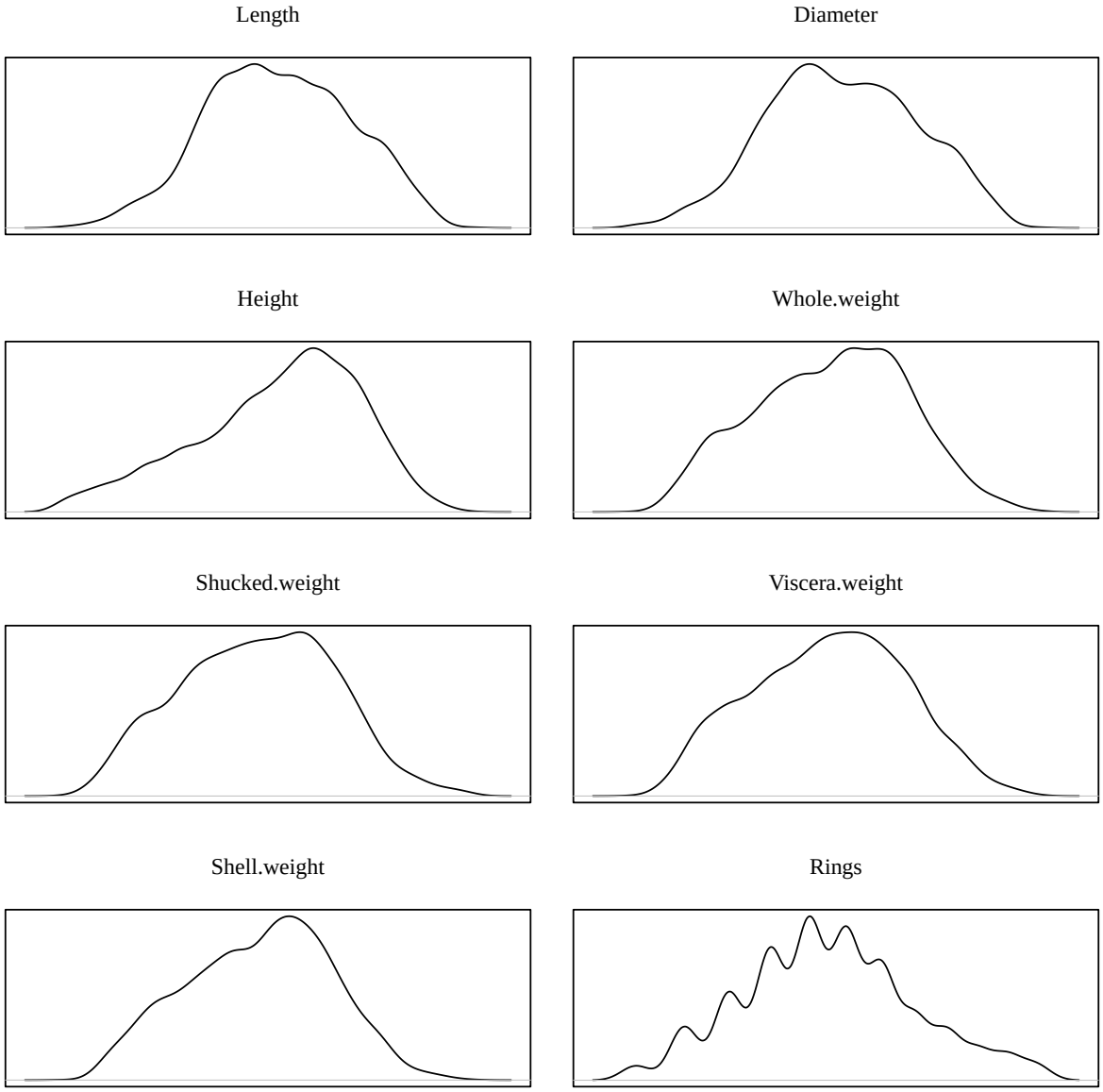

Figure 3: Density Plots After Transformations and Removal of Outliers

The boxplots show that the values of length, diameter and height for males and females are considerably overlapped, while the values for infants are substantially separated from the other classes. This explains why the models are unable to differentiate between males and females.

Because of this, **none** of the models is suitable as a three-way classifier, although the best-performing model, SVM with a radial kernel, may be used to identify infants and males.

2.2 Binary Classifiers

Having failed to specify a classifier suitable for labelling infants, females and males, binary classifiers have been attempted, to differentiate between:

- Infants and non-infants (to avoid harvesting infant abalone)
- Females and non-females (to selectively harvest females for increased profits)
- Males and non-males (to selectively harvest males to preserve female populations)

Class	Accuracy (Linear) %	Accuracy (Radial) %
Infant	0.7206	0.7121
Female	0	0.0025
Male	0.7820	0.7926

Table 3: Accuracy of Three-Way Classifiers

A performance summary for the best models in these scenarios is given in Table 4. While it appears that all three models perform reasonably well, closer inspection confirms otherwise. The following are the observations from the **confusion matrix** shown in Table 5:

- **I vs Not I**: Very good at identifying non-infants, but poor at identifying infants.
- **F vs Not F**: Classifies nearly all abalone as non-female, regardless of true sex.
- **M vs Not M**: Classifies nearly all abalone as non-male, regardless of true sex.

Classifier	Best Model	Accuracy %
I vs Not I	SVM radial	0.7950
F vs Not F	SVM radial	0.6926
M vs Not M	SVM radial	0.6824

Table 4: Accuracy of Three-Way Classifiers

Therefore, it is recommended that a binary classifier for infants and non-infants **only** be used to positively identify non-infants, not for identifying infants. In addition, it is recommended that a binary classifier **not** be used at all for females and non-females, and males and non-males.

Unfortunately, this means that the objective to use non-intrusive measurements to predict the sex of abalone cannot be met, because males and females are not differentiable by their physical proportions; rather, the best method of determining sex is to examine the colour of the abalone's underside (XXXX).

3 Estimating Price

	Predicted	Predicted Not
Actual I	675	506
Actual Not I	273	2346
Actual F	84	1122
Actual Not F	46	2548
Actual M	1	1412
Actual Not M	0	2387

Table 5: Confusion Matrix for Binary Classifiers

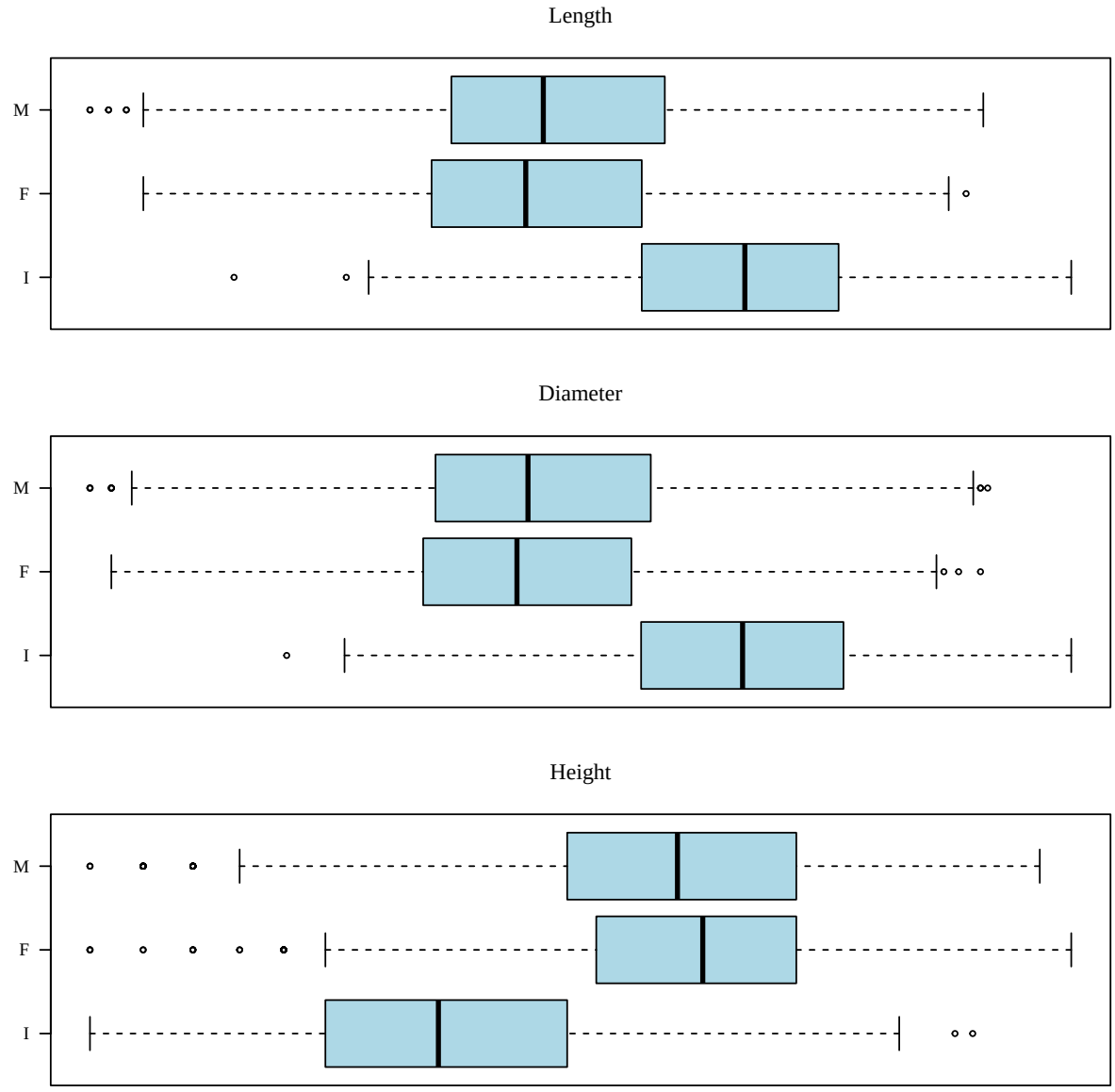

Figure 4: Input Variables by Sex Class