ZZSC5855 Project

Abalone Harvesting

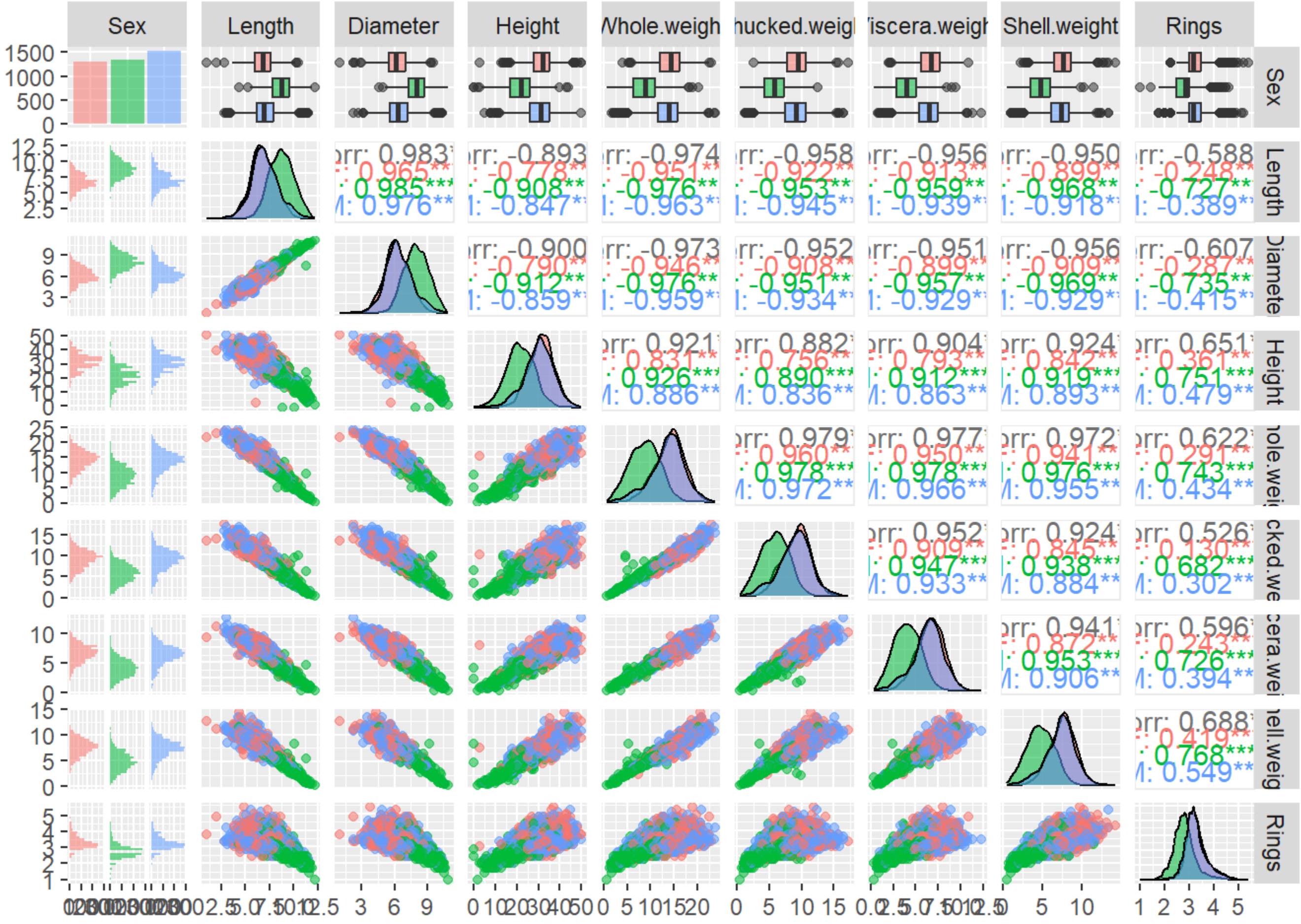
By Waikei Lau (z5349878)

Abalone harvesting

Abalones are marine snails that are considered a delicacy in many countries. This report investigates data to identify models and methods to better categorise and predict physical characteristics of these animals to promote sustainability and profitability in the abalone harvesting.

Data

The data used for this report comes from the UCI Machine learning repository.
https://archive.ics.uci.edu/ml/datasets/abalone



The data was transformed to approximate multivariate normality.

Sustainability

One approach to gender prediction is the use of statistical classification. This class of methods include Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machines (SVM). A method is chosen based on the set of assumptions the dataset satisfies. This LDA and QDA models rely heavily on assumptions for normality and equality of variance.

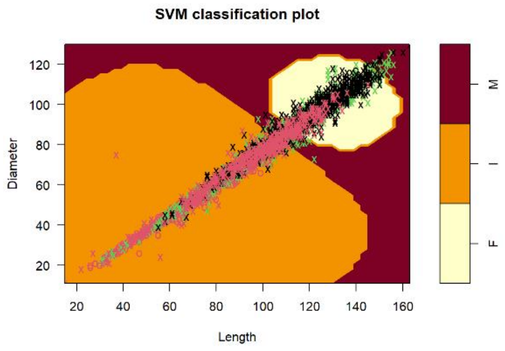
Since our data did not exhibit equal variance and the assumption of normality could not be rigorously proven, the most applicable method was the use of Support Vector Machines. A One-against-rest approach (*) was chosen, where a unique dataset was generated based on the following scenarios:

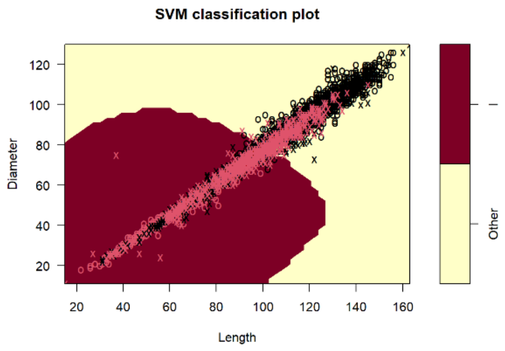
1. Unchanged
2. Non-female records were classified as “other” *
3. Non-male records were classified as “other” *
4. Non-infant records were classified as “other” *

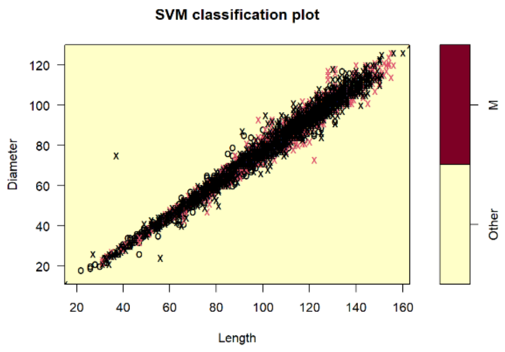
Tuning the SVM to use radial kernels and “low” cost parameters optimized accuracy while minimizing the use of computing power (i.e. support vectors). The One-against-rest approach demonstrated improved accuracy with the following accuracy rates corresponding to the above scenarios:

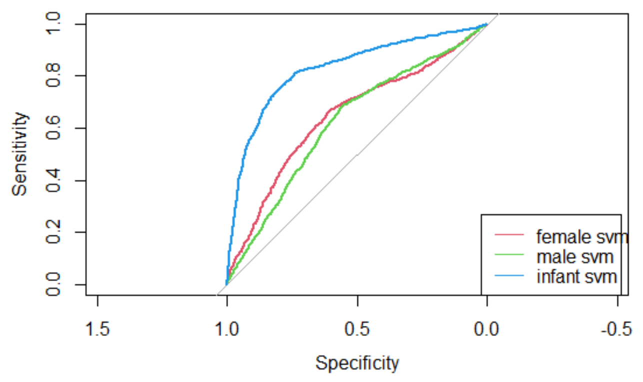
Results: {All|51%} - {female|68%} - {male|63%} - {infant|79%}

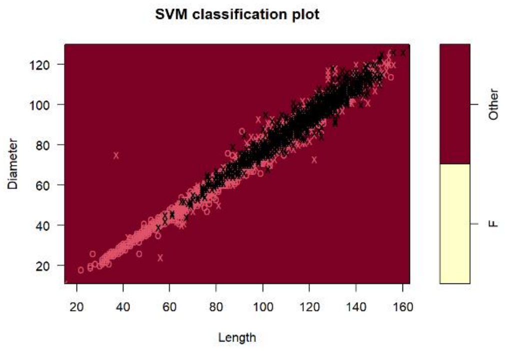
SVM performed significantly better when classifying infant abalone compared to male and female abalone. We see this was a result of the significant overlap in sizes for the mature adults and the type of measurements used. While weight and size measurements can easily distinguish infant from adult abalone, classification between male and female abalone is less accurate as they are very similar in size and weight.



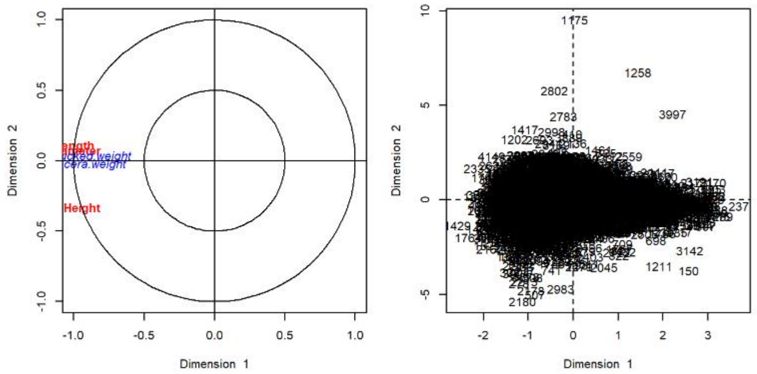








Profitability

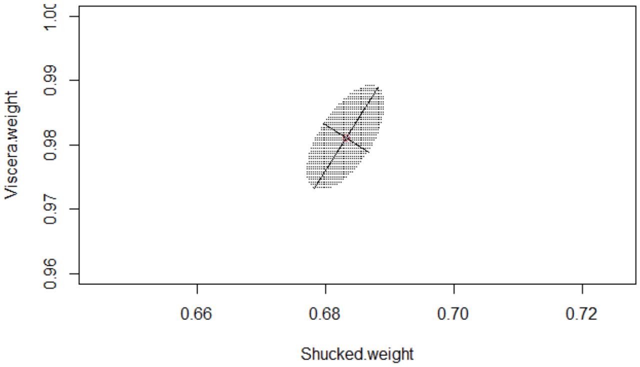


Canonical correlations measure the largest possible correlation between a linear combination of the variables in the first set and a linear combination of the variables in the second set.

To predict the value of abalone, and its shucked and viscera weight from size measurements, the use of canonical correlations (CC) was applied.

Satisfying the assumptions of CC:

- Variables are multivariate normal
- Large sample size
- No multicollinearity (correlation between variables less than 1)



Abalone data was transformed to approach approximate normality in addition to satisfying these assumptions, the weights (Shucked + Viscera) and size (Length + Diameter + Height) measurements were extracted as sets and analysed.

A linear relationship was tested and confirmed between the weight and size variables. All variables showed a significant effect on the correlation between weight and size.

Utilising properties of multivariate Normal, a correlation could be built such that given the size measurements, abalone weight could be predicted, leading to an estimate of the value of each abalone.

Additionally, a 90% prediction interval containing the true value of the abalone was conducted based on the shucked and viscera weight. That is, 90% of abalone would be in this weight interval. Hence, given price per gram of shucked and viscera weight, a 90% value interval can be determined using CC and properties of Multivariate normal.

Conclusion

The sustainability and profitability of abalone harvesting can be optimized using statistical techniques such as the ones mentioned in this report. Improving the accuracy of predictions and reducing the complexity of models allows greater operational efficiency and longer term business outcomes.