# Predict the age of abalone from physical measurements

Rohit Joshi
Computer Science and Electronic Engineering Department,
University of Essex,
Colchester, United Kingdom

## ABSTRACT

Abalones are sea snails or molluscs otherwise commonly called as ear shells or sea ears. Because of the economic importance of the age of the abalone and the cumbersome process that is involved in calculating it, much research has been done to solve the problem of abalone age prediction using its physical measurements available in the dataset. this paper approaches it as a classification and regression to predict the age. Furthermore, in contrast to previous research that saw this as a classification problem.

## Keywords

Abalone, Regression, classification.

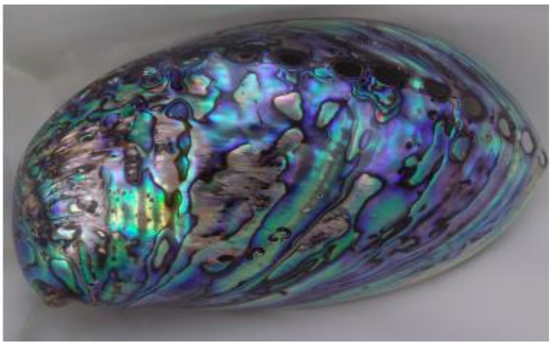


**Fig. A: Abalone**

## 1. INTRODUCTION

Abalones is a shellfish considered a delicacy in many parts of the world. An excellent source of iron and pantothenic acid, and a nutritious food resource and farming in Australia, America and East Asia. 100 grams of abalone yields more than 20% recommended daily intake of these nutrients. The economic value of abalone is positively correlated with its age. Therefore, to detect the age of abalone accurately is important for both farmers and customers to determine its price. However, the current technology to decide the age is quite costly and inefficient. Farmers usually cut the shells and count the rings through microscopes to estimate the abalones age. Telling the age of abalone is therefore difficult mainly because their size depends not only on their age, but on the availability of food as well. Moreover, abalone sometimes form the so-called 'stunted' populations which have their growth characteristics very different from other abalone populations This complex method increases the cost and limits its popularity. Our goal in this report is to find out the best indicators to forecast the rings, then the age of abalones.

Because some rings are hard to make out using this method, 1.5 is traditionally added to the ring count as a reasonable approximation of the age of the abalone. Knowing the correct price of the abalone is important to both the farmers and consumers while knowing the correct age is important to environmentalists who seek to protect this endangered species. Due to the inherent inaccuracy in the manual method of counting the rings and thus calculating the age, researchers have tried to employ physical characteristics of the abalone such as sex, weight, height and length to determine its age. The corresponding dataset is found at UCI’s repository.

Most of the research on the dataset has seen the abalone age prediction problem being categorized as a classification problem, that is, assigning a label to each example in the dataset. The label in this case is the number of rings of the abalone, which is a real number. This leads the classifier to distinguish among many classes and is thus bound to do poorly as can be seen in Zhengjie Wang’s results. To improve upon this approach, the number of classes is reduced. However, doing so beats the purpose of easing the process of calculating age (and thereafter price), especially in the absence of concrete data about the degree of correlation between age and price. For instance, two ages belonging to one of the reduced class but nonetheless causing a large variation in price would render the reduced class model useless. To overcome the problems associated with the classification model, this paper experiments with regression models and analyses the performance. Mean Absolute Error (MAE) is used as the evaluation metric to downplay the significance of outliers (too young or too old abalones, which are rare in nature) and because it allows us to make a straightforward conclusion: a MAE below 0.5 would guarantee that the regressor has made a correct and useful prediction.

## 2. DATASET ANALYSIS

The abalone dataset is a dataset that contains measurements of physical characteristics of different abalones. It has 4177 instances. The physical characteristics along with the unit of its measurement in brackets are (Table 1)

**Table 1. Description of variables in the abalone dataset**

| Index | Attribute | Measuring unit | Description |
|---|---|---|---|
| - | Sex | - | It can be either one of Male, Female or Indeterminate (Infant). Abalone gender is not determined at birth but rather when they mature a little [5] |
| 1 | Length | mm | Longest shell measurement |
| 2 | Diameter | mm | Perpendicular to length |
| 3 | Height | mm | Height of abalone with meat in shell |
| 4 | Whole weight | grams | Weight of the whole abalone |

| Index | Attribute | Measuring unit | Description |
|---|---|---|---|
| 5 | Shucked weight | grams | Weight of just the meat |
| 6 | Viscera weight | grams | Gut weight (after bleeding) |
| 7 | Shell weight | grams | Weight of shell after being dried |
| 8 | Rings | - | This is the dependent variable (label). Number of rings + 1.5 gives age |

Figure 1 shows the distribution of rings in the abalone dataset. It can be seen that the dataset is skewed with majority examples having rings in the range of 7-14 with very few examples having rings above 20. The exact number of examples in ascending order of number of rings in the examples is: (1, 1, 15, 57, 115, 259, 391, 568, 689, 634, 487, 267, 203, 126, 103, 67, 58, 42, 32, 26, 14, 6, 9, 2, 1, 1, 2, 0, 1).

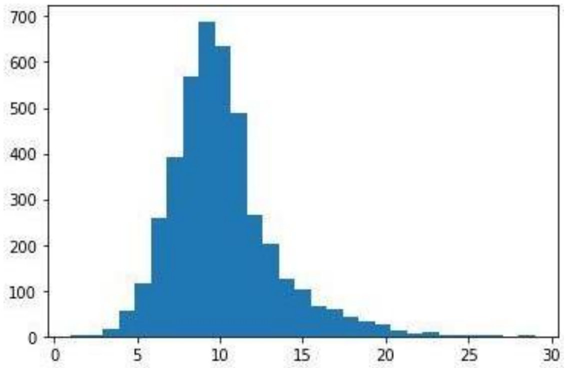


**Fig. 1: Distribution of Rings Variable**

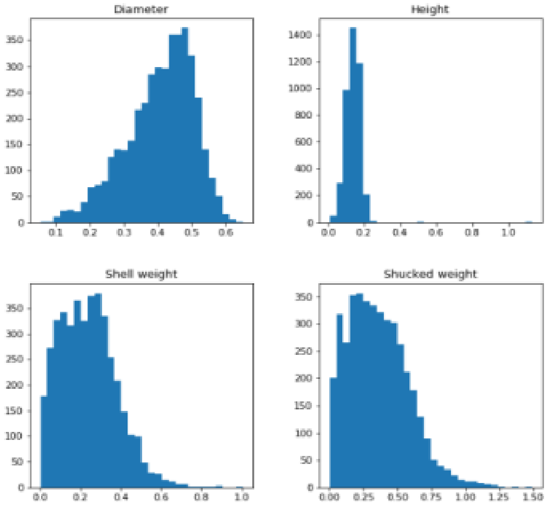


**Fig. 2: Histogram for Diameter, Height, Shell and Shucked weights.**

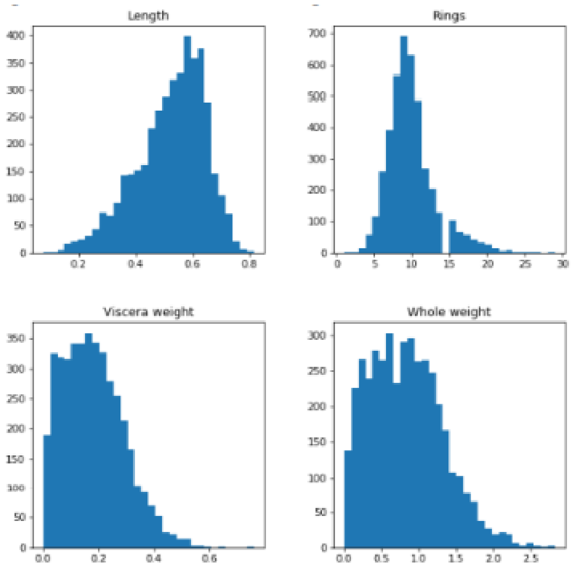


**Fig 3: Histogram for Length, Rings, Viscera and Whole weights.**

The minimum, maximum, mean, median, standard deviation and interquartile range of all the numeric attributes along with dependent variable of the dataset is calculated and plotted using a boxplot for easy visualization of outliers. Due to the larger range of "Rings" variable, an unnormalized boxplot renders the other variables' boxplots incomprehensible by squeezing their ranges. To bring all the variables on the same scale, they are normalized such that they all have zero mean and standard deviation 1. Figure 2 shows this boxplot. The attributes Length and Diameter have almost the same normalized range while there are a few outlying values for the Height attribute which might make the task of regression difficult. All the Weight attributes also have almost the same normalized range. The Rings label is not analyzed since it will be used in an unnormalized form for the regression to obtain a proper value of MAE.

Next, the relation among the variables and between the variables and label is analyzed. The normalized attribute values for each example are first passed through a logit function to fit them in the range (0,1) which removes any negative values that may be there after normalization. A parallel plot (Figure 3) is constructed which plots these values of all attributes and it is coloured based on the value of the label (Rings). Dark brown represents less rings while dark blue represents higher number of rings. The parallel plot reveals significant correlation between each of the attributes and the label for each of the examples; similar colour shades are grouped together at several attributes for similar values.

This suggests that the prediction model will be fairly accurate. However, there are a few examples which do not follow the above trend. Dark blue lines mixed with lighter blue lines on the right and some blue lines in between the brown ones suggest these examples will be difficult to predict correctly.

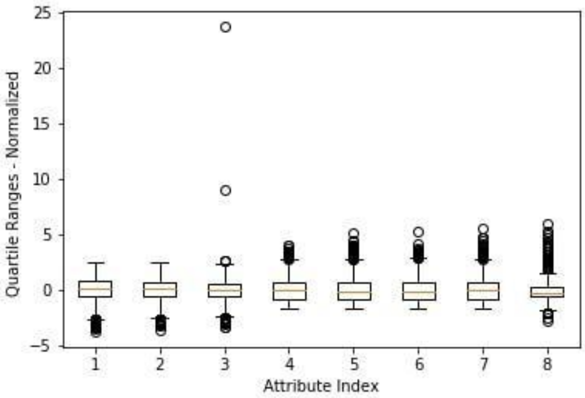


**Figure 4: Normalized Boxplot**

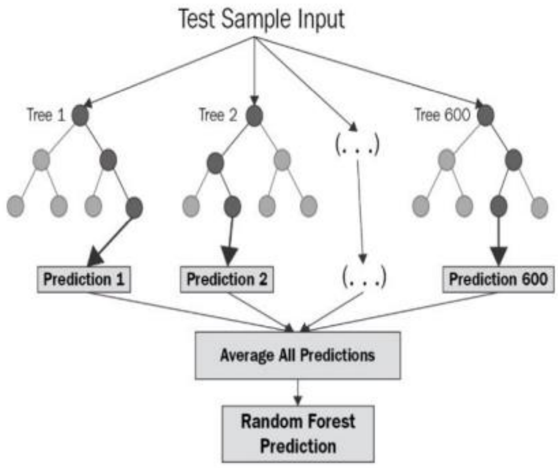
Figure 4 specifically looks at the values of correlation among the numeric attributes and between the numeric attributes and label. Apart from the index starting at zero, the order of attributes is as shown in Table 1 (the Rings label forms the last column and last row). It can be clearly seen that the different attributes have a strong correlation with each other which confirms the analysis of the parallel plot. However, the correlation of the attributes with the label is markedly less contradicting the findings of the parallel plot. Based on the correlation of the attributes with the label, it can be concluded that Shell Weight is the most important attribute for prediction.

## 3. LITERATURE REVIEW

Investigators are building new notions to conclude the age of abalone by altered methods. Let's say, marine natural scientist are spending the laboratory investigation to define the age of abalone, machine learning scientists are using classification procedure expending physical faces of abalone to define the age, econometricians and statisticians are also expending physical faces of abalone to define the age using different kinds of regression as well as clustering, and many other people are expending different techniques to detect the age of abalone. Naval natural scientist Takami, H. et al. [10] advanced an age determination way for larval and newly changed post-larval abalone Haliotis discuss hannai in a test site testing and resolute the age of field caught individuals. Day, R. W. et al. [6] developed a method where they assessed the potential of five fluorochromes in marking shells of the abalone 3 Haliotis rubra, using an immersion technique. Such marks are required to 'time stamp' the shells and thus determine whether shell layers are deposited regularly enough to be used to age abalone. They also reference that juvenile growth does not right the commonly used von Bertalanffy model and they present a modified deterministic Gompertz model for tagging data and three stochastic versions in which asymptotic length is a random parameter. They also reference that juvenile growth does not right the commonly used von Bertalanffy model and they present a modified deterministic Gompertz model for tagging data and three stochastic versions in which asymptotic length is a random parameter. They use Kullback's informative mean to discriminate between models with respect to the fit to data. Siddeek, M. S. M., and Johnson, D. W. [9] define that length frequency data for Omani abalone (Haliotis mariae) from two zones (Sadh and Hadbin) of the Dhofar coast of the Sultanate of Oman were used to right von Bertalanffy development curves by ELEFAN, MULTIFAN and Non-Linear Least Square Fitting methods. The first two methods were directly applied to length-frequencies whereas the last method was used on the length modes determined by the MIX method. The growth stricture values by sex and area were not meaningfully different. Al-Daoud, E. [3] uses neural network technique to classify the number of rings using physical characteristics. Using the von Bertalanffy growth equation Bretos, M. [4] proposes a method to determine the age of abalone. Gurney, L. J., et al. [7] describe the stable oxygen isotopes procedure to determine the blacklip abalone Haliotis rubra in south-east Tasmania. However, Naylor, et al. [8] find that the method, variations in the ratios of carbon isotopes, showed no consistent patterns and unlike some mosllusc, do not appear to be useful predictors of reproductive status at length.

## 4. METHODOLOGY

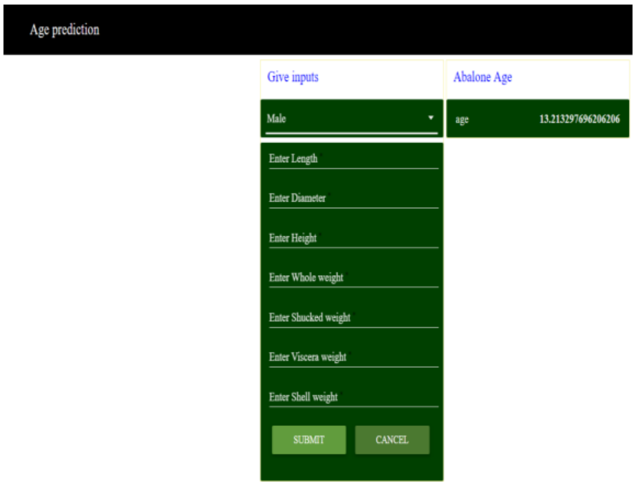
KNN: K-Nearest Neighbor Regression (KNN) everything in much the identical way as KNN for classification. The difference lies in the features of the dependent variable. With classification KNN the needy variable is categorical. With regression KNN the dependent variable is incessant. Both involve the use adjacent samples to forecast the class or value of other examples. The k-nearest neighbors are a simple, easy-to-appliance supervised machine learning algorithm that can be used to control the strength and character of the relationship between the dependent variable and the independent variables. It undertakes that related things exist in close proximity. In other words, things are near to each other. The following are the steps Of KNN algorithm that is to be applied to the prediction of regression problem. 1. Load the data. 2. Adjust K to your selected number of neighbors. 3. For each example in the data, Calculate the distance between the query example and the current example from the data and add the distance and the index of the example to an ordered collection. Euclidean distance: It is designed as the square root of the sum of the squared differences between a new point (x) and an existing point (y). Manhattan Distance: This is the space between real vectors using the totality of their complete difference. Hamming Distance: It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0. Otherwise, D=1. Once the distance of a new reflection from the points in our training set has been measured, the next step is to pick the nearby points. The number of points to be considered is defined by the value of k. 4. Type the orderly gathering of distances and keys from smallest to largest (in ascending order) by the distances. 5. Best the first K items from the sorted crowd. 6. Get the markers of the selected K entries. 7. If regression, arrival the mean of the K labels. 8. If classification, arrival the mode of the K labels. To select the K that’s right for your data, we run the KNN algorithm several times with different values of K and choose the K that moderates the number of errors we meet while keeping the algorithm’s skill to exactly make guesses when it’s given data it hasn’t seen before. Random Forest Regressor: Random Forest is a Supervised Learning algorithm which uses collective learning method for classification and regression. Random forest is a trapping technique and not a boost up. The trees in random forests are run in parallel. There is no contact between these trees while building the trees.



It initiates by construction of an assembly of decision trees at training time and resulting the class that is the manner of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta-estimator (i.e., it chains the outcome of many forecasts) which combinations many decision trees, with some supportive adjustments: 1. The quantity of features that can be divided on at each node remains limited to some percentage of the total (which is known as the hyperparameter). This guarantees that the group model does not depend on too heavily on any discrete feature, and makes fair use of all possibly predictive landscapes. 2. Each tree attractions a random sample from the inventive data set when generating its splits, adding a further division of randomness that prevents overfitting.

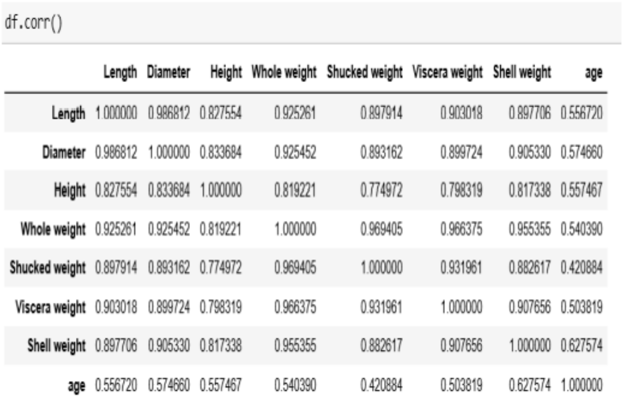
## 5. RESULTS AND DISCUSSION

The output is seen through user interface which is a web UI developed by using Node-RED in IBM Watson Studio. This interface consists of different fields that user has to give the physical measurements of abalone. If the user enters all the physical values of abalone and click on submit button then it predicts the age of abalone.



age 13.213297696206206

Variable within the dataset can be related for lots of reasons. For example, one variable could cause or depend on the values of another variable or one variable could be lightly associated with another variable or two variables could depend on a third unknown variable. The correlation for the variables presents in the abalone dataset is



```
df.corr()
```

| | Length | Diameter | Height | Whole weight | Shucked weight | Viscera weight | Shell weight | age |
|---|---|---|---|---|---|---|---|---|
| **Length** | 1.000000 | 0.986812 | 0.827554 | 0.925261 | 0.897914 | 0.903018 | 0.897706 | 0.556720 |
| **Diameter** | 0.986812 | 1.000000 | 0.833684 | 0.925452 | 0.893162 | 0.899724 | 0.905330 | 0.574660 |
| **Height** | 0.827554 | 0.833684 | 1.000000 | 0.819221 | 0.774972 | 0.798319 | 0.817338 | 0.557467 |
| **Whole weight** | 0.925261 | 0.925452 | 0.819221 | 1.000000 | 0.969405 | 0.966375 | 0.955355 | 0.540390 |
| **Shucked weight** | 0.897914 | 0.893162 | 0.774972 | 0.969405 | 1.000000 | 0.931961 | 0.882617 | 0.420884 |
| **Viscera weight** | 0.903018 | 0.899724 | 0.798319 | 0.966375 | 0.931961 | 1.000000 | 0.907656 | 0.503819 |
| **Shell weight** | 0.897706 | 0.905330 | 0.817338 | 0.955355 | 0.882617 | 0.907656 | 1.000000 | 0.627574 |
| **age** | 0.556720 | 0.574660 | 0.557467 | 0.540390 | 0.420884 | 0.503819 | 0.627574 | 1.000000 |

## 6. CONCLUSION AND FUTURE WORK

On the source of this study, it appears the future regression systems effort well to forecast the age of abalone. The study directs that we do not prerequisite to count the quantity of rings consuming microscopic test. In other disputes, we do not need any laboratory experiment to predict the age of abalones. We can predict the age and price of abalone using the very simple physical individualities like weight, height, diameter, and length.

## 7. REFERENCES

[1] Abalone: https://en.wikipedia.org/wiki/Abalone

[2] Hossain, M, & Chowdhury, N (2019) Econometric Ways to Estimate the Age and Price of Abalone. Department of Economics, University of Nevada.

[3] UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone

[4] Wang, Z (2018) Abalone Age Prediction Employing A Cascade Network Algorithm and Conditional Generative Adversarial Networks. Research School of Computer Science, Australian National University

Bowles, M (2015). Machine Learning in Python: Essential Techniques for Predictive Analysis, John Wiley & Sons, Inc.

[5] Alsabti, K., Ranka, S., & Singh, V (1999) CLOUDS: A decision tree classifier for large datasets.

[6] Mayukh, H. (2010) Age of Abalones using Physical Characteristics:A Classification Problem. Department of Electrical and Computer Engineering, University of Wisconsin-Madison.

[7] Fahlman, S & Lebiere, C (1990) The Cascade- Correlation Learning Architecture. Neural Information Processing Systems Conference, 1990.

[8] Goodfellow, I et al (2014) Generative Adversarial Nets. Department of Computer Science and Operations Research, University of Montreal, Canada.

[9] Mirza, M (2014) Conditional Generative Adversarial Nets. Department of Computer Science and Operations

Research, University of Montreal, Canada.

[10] Chawla, N. (2002) C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure.

[11] Saina, H & Purnamia, S (2015). Combine Sampling Support Vector Machine for Imbalanced Data Classification, The Third Information Systems International Conference.

[12] Pedregosa et al (2011). Scikit-learn: Machine Learning in Python,JMLR12.