Abalone Age Prediction Using IBM Watson

**1.INTRODUCTION**

**2.LITERATURE SURVEY**

**3.DATASET**

**4.METHODOLOGY USED**

4.1. Collect/Observe Data

4.2. Explore and Prepare Data

4.3. Train Model Based On Data

4.4. Evaluate the Performance of the Model

4.5. Improve the Performance of Model

**5.THEORITICAL ANALYSIS**

5.1 Block diagram Diagrammatic overview of the project.

5.2 Hardware / Software designing Hardware and software requirements of the project

**6.FLOWCHART**

**7.WEBSITE INTERFACE**

**8.RESULT**

**9.ADVANTAGES & DISADVANTAGES**

**10.APPLICATIONS**

**11. CONCLUSION**

**12.FUTURE SCOPE**

**13.BIBILOGRAPHY**

**1.INTRODUCTION:**

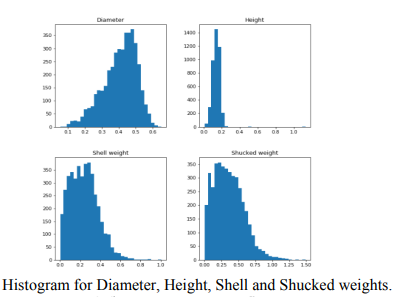
Abalones are endangered marine snails that are found in the cold coastal waters worldwide, majorly being distributed off the coasts of New Zealand, South Africa, Australia, Western North America, and Japan. They are considered a delicacy and a highly nutritious food and extensively consumed in certain parts of Latin America, France, New Zealand, Southeast Asia, China, Vietnam, Japan, and Korea. They are also commercially farmed as a source of mother-of-pearl. The shells of abalone are used for decorative purposes owing to their iridescence. This makes abalone a highly sought after commodity and economically significant. The price of an abalone is positively correlated to its age.However, determining the age of an abalone is a highly involved process. Rings are formed in the inner shell of the abalone as it grows, usually at the rate of one ring per year. Getting access to the rings of an abalone involves cutting the shell. After polishing and staining, a lab technician examines a shell sample under a microscope and counts the rings.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.

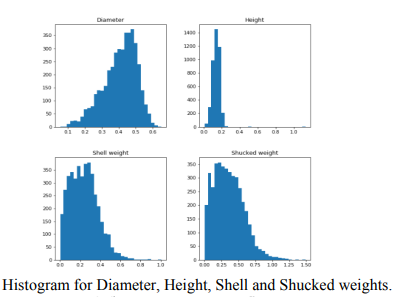
**2.LITERATURE SURVEY:**

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.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.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 use Kullback's informative mean to discriminate between models with respect to the fit to data. Siddeek, M. S. M., and Johnson, D. W.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.uses neural network technique to classify the number of rings using physical characteristics. Using the von Bertalanffy growth equation Bretos, M. proposes a method to determine the age of abalone. Gurney, L. J., et al. describe the stable oxygen isotopes procedure to determine the blacklip abalone Haliotis rubra in south-east Tasmania. However, Naylor, et al.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.

**3.DATASET:**

The University of California, Irvine Machine Learning Repository [2] affords a data set consisting of 4177 samples of physical faces of abalones and their age. Abalones are sea-snails that are fished for their missiles and meat. Technical trainings on abalones need expressive the age of an abalone, but the route of defining age is thorny. It comprises calculating the number of layers of shell (“rings”) that make up the abalone’s shell. This is done by pleasing a sample of shell, staining it and including the number of rings beneath the microscope. To avoid the cumbersome process, this data set providing has been used to build education algorithms to predict age using easily and quickly computable physical appearances. This project uses this data set recast as a regression problem, slightly a prediction problem. The data set involves of 8 features and the number of rings which is directly related to the age. The 8 features are sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight. All variables are quantifiable but Sex. The variable rings are someway linked to the age of an abalone, as age equals to number of rings plus 1.5. Histograms display the data may be slanted, so it will be sensible to measure it. It also shows that there are probable outliers in Height and that there might be a strong connection between the Diameter and Length and between Shell weight, Shucked weight Viscera weight and Whole weight.





**4.METHODOLOGY USED:**

**4.1. Collect/Observe Data**

For analysis purposes, this article uses a data set from a web page. The data set contains the real attribute values of multiple abalone, including Sex, Length, Diameter, Height, Whole. Weight, Shucked. Weight, Viscera. Weight, Shell. Weight, Rings and other properties. Therefore, it can roughly reflect the situation of abalone in the real world.

**4.2. Explore and Prepare Data**

The first step is to prepare data. We need to deal with missing values. Generally, there will be some problems with the dataset we got for the first time. Thus, we use the ‘sum’ function to determine if there are missing values and the completeness of the data. After the previous step, the appropriate data set has been collected in this article, which needs to be initially processed in this article. In this section, in general, one may need to check whether values are available, convert values, detect/eliminate outliers, or some other section. However, in this data set, after testing, the data set basically meets the conditions for further processing.Therefore, the relationship between features (correlation coefficient matrix) and the relationship between visual features (scatter diagram matrix) are mainly explored in this part. In doing so, this article calls the functions of cor (), pairs (), and pairs. Panels () to complete the determination. The first two functions can be called directly from Rstudio, while the last requires the psych package to be pre-installed in Rstudion.In view of the this,all variables except Height are positively correlated with abalone age. Therefore, in the initial variable selection, this article will take all variables except the Height variable as the dependent variables.

**4.3. Train Model Based On Data**

This paper adopts multiple linear regression model for modeling. The function Im () that needs to call is included by default in Rstudio software, so people can call it directly. After modeling the six variables selected above, the results are as follows (sex variable will not be considered because it is a factor value and make the process more complex):

**4.4. Evaluate the Performance of the Model**

After getting the model of training, this paper needs to evaluate the performance of the model. In doing so, the functions of summary () and plot are called to perform significance and residual analysis of the function.Significance analysis of the initiatory model and Residuals analysis of the initiatory model It is not difficult to see from the result graph that the significance of the Length variable is not high enough. In addition, there is a faint contour line in the last figure (bottom right) of the residual analysis diagram, which means that there may be outliers in the model.

**4.5. Improve the Performance of Model**

After the performance evaluation, it is not difficult to find some problems in the current model. Since Length is not significant for dependent variables, the Length variable is removed later in this article for modeling. This article will take a step further after the performance assessment of the new model is completed. However, in the actual operation process, the results obtained by repeating steps 3 and 4 in this paper are found to be completely in line with the expectation.

**5.THEORITICAL ANALYSIS:**

**5.1 Block diagram**

Diagrammatic overview of the project

Input

(dataset)

prediction

Output

(age)

**5.2 Hardware / Software designing**

Hardware and software requirements of the project

1.Anaconda Navigator

2.Jupyter Notebook & Spyder

**6.FLOWCHART :**

Diagram showing the control flow of the solution

Start

Data set

Dataset pre-processing

Model setup

Model training

Metrics score and model loss

Model Evaluation

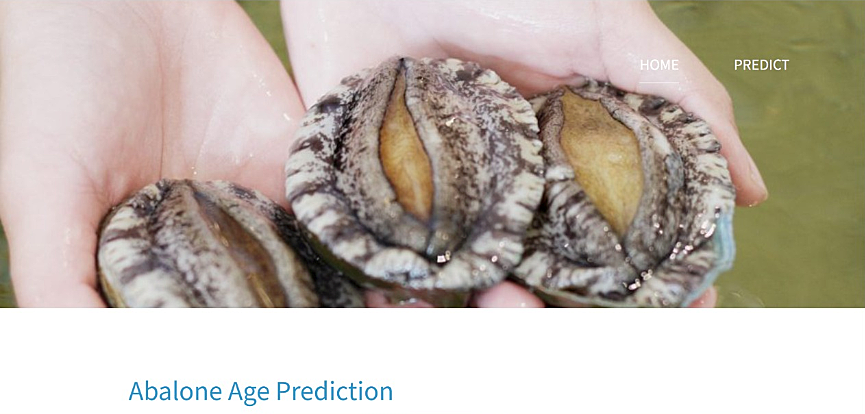
Performance Analysis

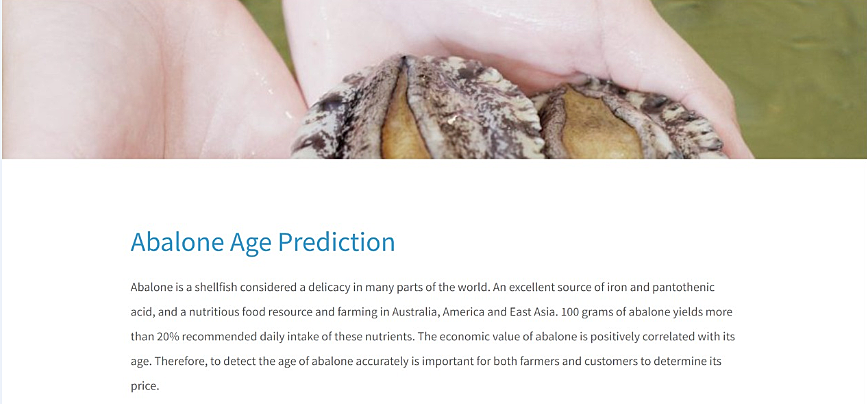
Parameters Tuning

End

**7.Website interface:**

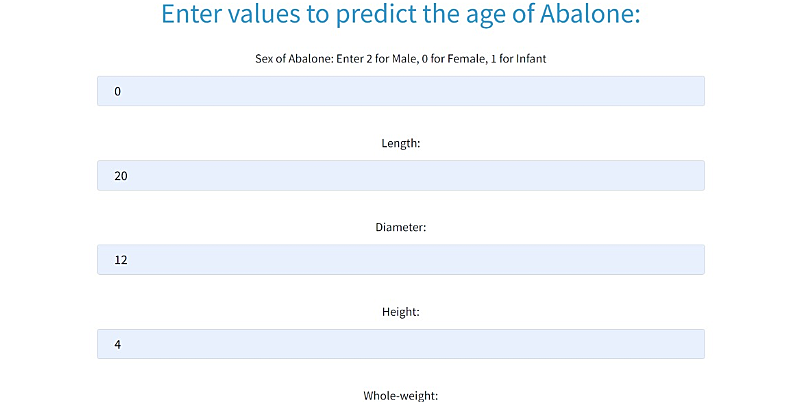
1.Home page:

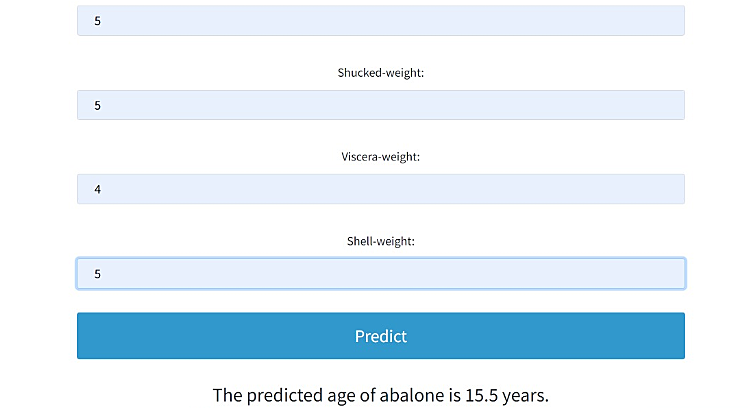




2.Predict page:







**8.RESULT:**

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.

**9.ADVANTAGES & DISADVANTAGES:**

It is salient to predict abalone age as it **helps farmers and sellers to determine the market price of abalones**. The economic value of abalone is positively correlated with their respective ages.

**10.APPLICATIONS:**

The highly iridescent inner nacre layer of the shell of abalone has traditionally been used as

A decorative item,

In jewelry, buttons, and as inlay in furniture and musical instruments, such as on fret boards and binding of guitars.

**11. CONCLUSION :**

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.

**12.FUTURE SCOPE :**

In the task of predicting age of an abalone (by predicting number of rings) through its physical characteristics, the RANSAC regression model works best with a MAE of 1.332. Huber regressor does a pretty good job in comparison (MAE= 1.399) while penalized regression models cannot outperform OLS (MAE=1.5). All over, robustness regression models do a good job in dealing with outliers present in the abalone dataset. Techniques such as SMOTE and Cross Validation do not improve the performance of the models with RANSAC performing the best here too achieving a MAE of 1.830. This cements its position as the best model. Normalization of attributes seems to result in the same performance as unnormalized data. The scatter plots for individual folds of cross validations show that Mean Absolute Error can be brought down below 1 for certain arrangements of data, the best being 0.936. However,it also shows that the error is still above the acceptable limits in some regions. It is the author’s belief that given an adequate and balanced dataset, RANSAC along with SMOTE and Cross Validation can achieve the goal of Mean Absolute Error less than 0.5 across all the labels.

**13.BIBILOGRAPHY :**

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

2.Al-Daoud,E.(2009).A comparison between three neural network models for classification problems. Journal of artificial intelligence, 2(2), 56-64.

3.Bretos, M. (1980). Age determination in the keyhole limpet Fissurella crassa Lamarck (Archaeogastropoda:Fissurellidae), based on shell growth rings. The Biological Bulletin, 159(3), 606-612.