ABALONE

AGE PREDICTION

- s15089 - Sanjani Wickramasinghe
- s14982 - Poornima Dissanayake
- s14953 - Buddhima Senarathna
- s15006 - Pasindu Sachintha

GROUP 10

ST 3082

Abstract

Abalone is a valuable seafood that is sold based on its age, which is determined by counting the number of rings in its shell. However, this method is time-consuming and invasive, leading to inefficiencies in the abalone market. To address this issue, we propose the development of an ANN model that can predict the number of rings in abalone based on physical measurements, such as the length, diameter, and weight of the shell. We will use a dataset of abalone with known ages and physical attributes to train and validate the ANN model, optimizing its hyperparameters for maximum accuracy. Our aim is to create a model that can predict the number of rings in abalone with a high level of accuracy, providing an efficient and non-invasive method for determining their age.

We will start by conducting descriptive analysis on the abalone dataset, which includes information on its physical characteristics, age, and price. This will help us to identify any patterns or trends in the data and to select the most relevant features for our model.

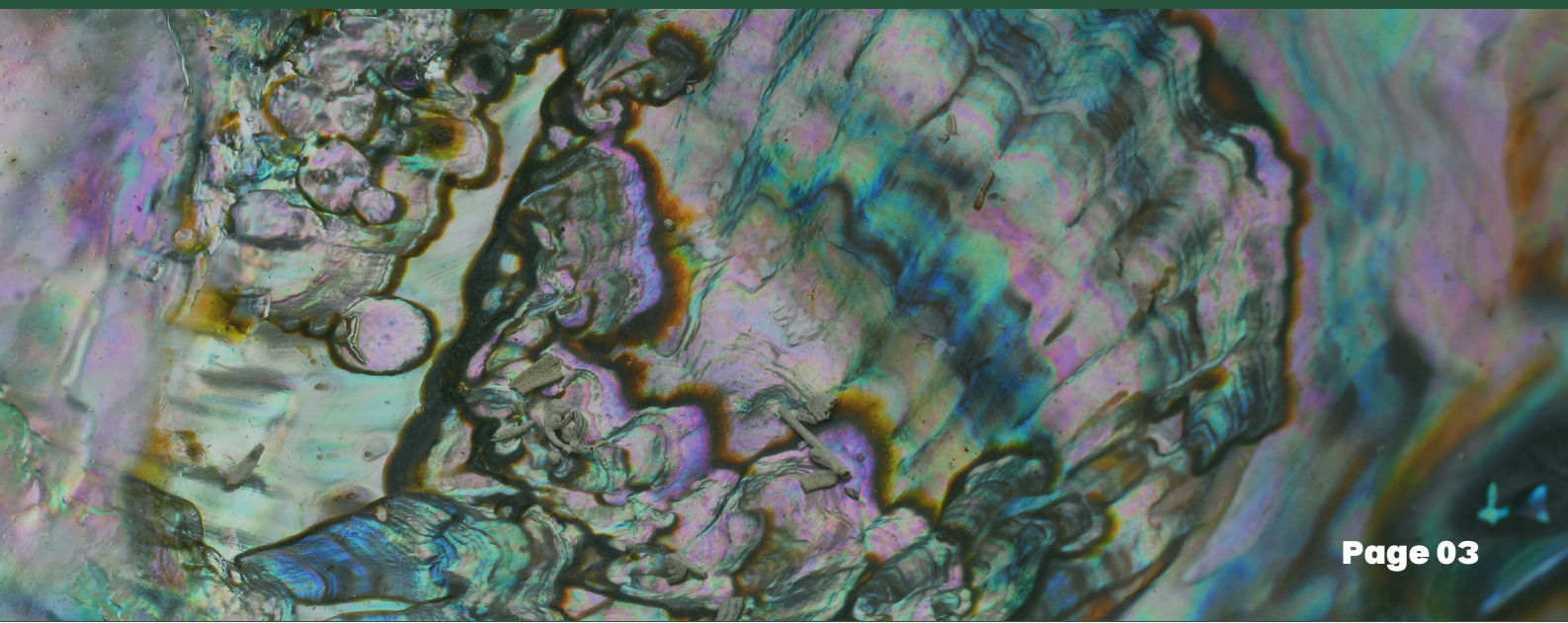
We will then use multiple linear regression to create a baseline model, which will help us to understand the relationships between the physical measurements and the number of rings in abalone. We will also use regularization techniques, such as Lasso and Ridge regression, to improve the model's performance and to prevent overfitting.

Finally, we will explore the use of random forest, a powerful machine learning technique that can handle non-linear relationships and interactions between variables. This will allow us to create a highly accurate model that can predict the number of rings in abalone with minimal error.
Overall, the successful development of such a model could have significant implications for the abalone industry, providing an efficient and non-invasive method for determining the age of these valuable seafood products.

Table of content

List of Figures

List of Tables



Introduction

Abalones are a type of marine snail.These sea creatures are found along the coasts of California, Mexico, Japan, and other parts of the world. Abalones have a unique appearance, with a flattened, oval-shaped shell that is often adorned with a variety of colors and patterns. They also have a muscular foot that allows them to cling tightly to rocks and other surfaces in the ocean.

Abalone shells have a distinctive low, open spiral structure and are notable for a series of respiratory pores that are arranged in a row along the outer edge. The inner layer of the shell is made up of nacre, which is commonly known as mother-of-pearl. In several abalone species, this nacre layer has a striking iridescence that creates a wide range of bright and shifting colors. This feature has made abalone shells highly desirable to people for their ornamental value, such as in jewelry or as decorative objects. Additionally, the shells are also used as a source of mother-of-pearl, which is valued for its vibrant hues and durability.

The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked by a variety of cultures and they have long been a valuable food source for humans in every area of the world where a species is abundant. The meat of this mollusk is considered a delicacy in certain parts of Latin America, France, New Zealand, East Asia and Southeast Asia. In the Greater China region and among Overseas Chinese communities.

However,our goal is to create a model that can accurately and efficiently estimate the age of abalone, which will have important applications for both commercial and scientific purposes.

Problem Statement

Abalone are a valuable food source that are sold based on their age, as determined by the number of rings in their shells. However, counting the rings is a time-consuming and invasive process, making it difficult to accurately determine the age of abalone in a timely manner. This creates challenges for the fishing industry, as well as for researchers studying the biology and ecology of these animals. To address this problem, we aim to build a machine learning model that can predict the number of rings in abalone shells based on a set of physical measurements. Our goal is to create a model that can accurately and efficiently estimate the age of abalone, which will have important applications for both commercial and scientific purposes.

Description of the dataset

The abalone dataset is a collection of measurements and other attributes for a total of 4,177 abalone specimens. Abalones are a type of sea snail, and this dataset includes specimens from the Western Pacific Ocean. The data was collected in 1995 by researchers at the Marine Resources Division of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO).
Table below lists down the variables along with a brief description about the variables.

#	Name	Data Type	Measurement Unit	Description	Attribute Type
1	Sex	Categorical		M, F, and I (infant)	Input
2	Length	Continuous	mm	Longest shell measurement	Input
3	Diameter	Continuous	mm	perpendicular to length	Input
4	Height	Continuous	mm	with meat in shell	Input
5	Whole weight	Continuous	mm	whole abalone	Input
6	Shucked weight	Continuous	mm	weight of meat	Input
7	Viscera weight	Continuous	mm	gut weight (after bleeding)	Input
8	Shell weight	Continuous	mm	after being dried	Input
9	Rings	integer		+1.5 gives the age in years	Output

Table 1: Description of the dataset

Data Preprocessing

For the analysis of the abalone dataset, some data cleaning and preprocessing steps were carried out. One of these steps was to remove any records where the height of the abalone was zero, which may have been due to measurement errors or other issues. Additionally, one outlier was identified where the height was greater than 1 and was also removed from the dataset.

In order to prepare the data for different machine learning techniques, different normalization and standardization techniques were applied. For multiple linear regression and artificial neural network models, data normalization was used to ensure that all the features were on a similar scale. On the other hand, partial least square regression and random forest models used data standardization, which scales the features to have zero mean and unit variance, in order to improve model performance.

Important Results of Descriptive Analysis

Our main objective of the analysis is to predict the number of rings present in an abalone shell, given its characteristics. Hence our target variable is the number of rings.

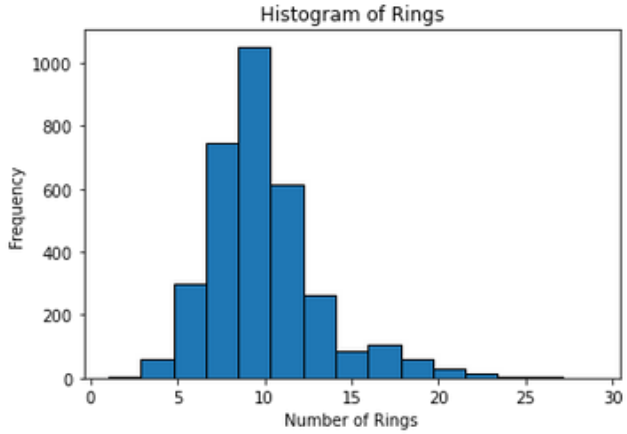


Figure 1: Histogram of rings

The rings variable ranges from 1 to 29. As we can see in the figure, the distribution of rings is positively skewed. This indicates that most abalones in the dataset have a relatively small number of rings, with a smaller proportion of abalones having a larger number of rings.

As abalones grow older, they tend to accumulate more rings. Rings in abalones are formed as a result of the growth of the shell, and each ring represents one year of growth. Abalones face various threats in their environment, such as predation, disease, and other ecological factors. As they grow older, the chances of mortality increase, leading to a smaller population of abalones with higher rings count. This could be a possible reason to the skewness of the distribution.

According to the figure of boxplots, it's clear that the variables such as length and diameter are left skewed, whole weight, shucked weight, viscera weight and shell weight are right skewed while height variable is approximately normally distributed.

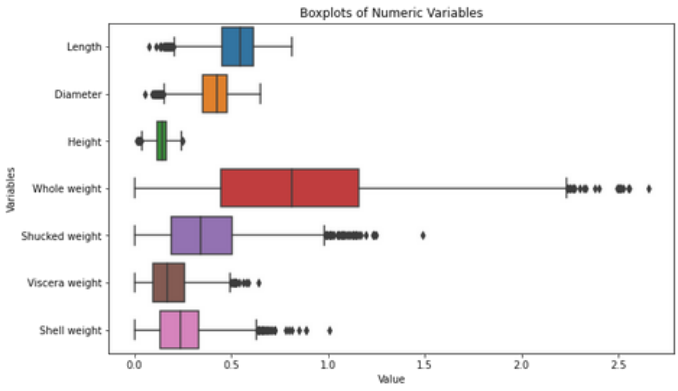


Figure 2: Boxplot of numeric variables

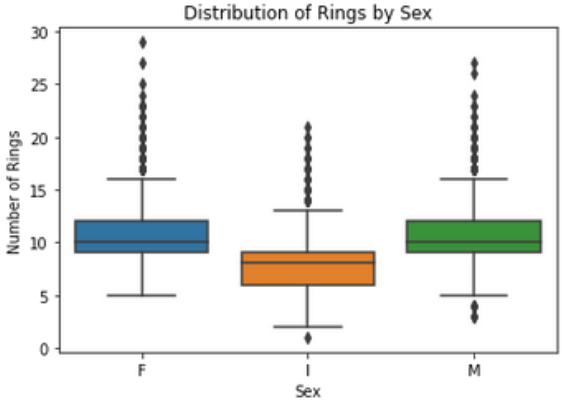


Figure 3: Boxplot of distribution of rings by sex

When considering the distribution of rings according to the gender, male and female abalones have approximately the same distribution implying sexual dimorphism may not have a significant influence on the growth rate or ring formation in abalones. Infant abalones tend to have lesser rings compared to other categories.

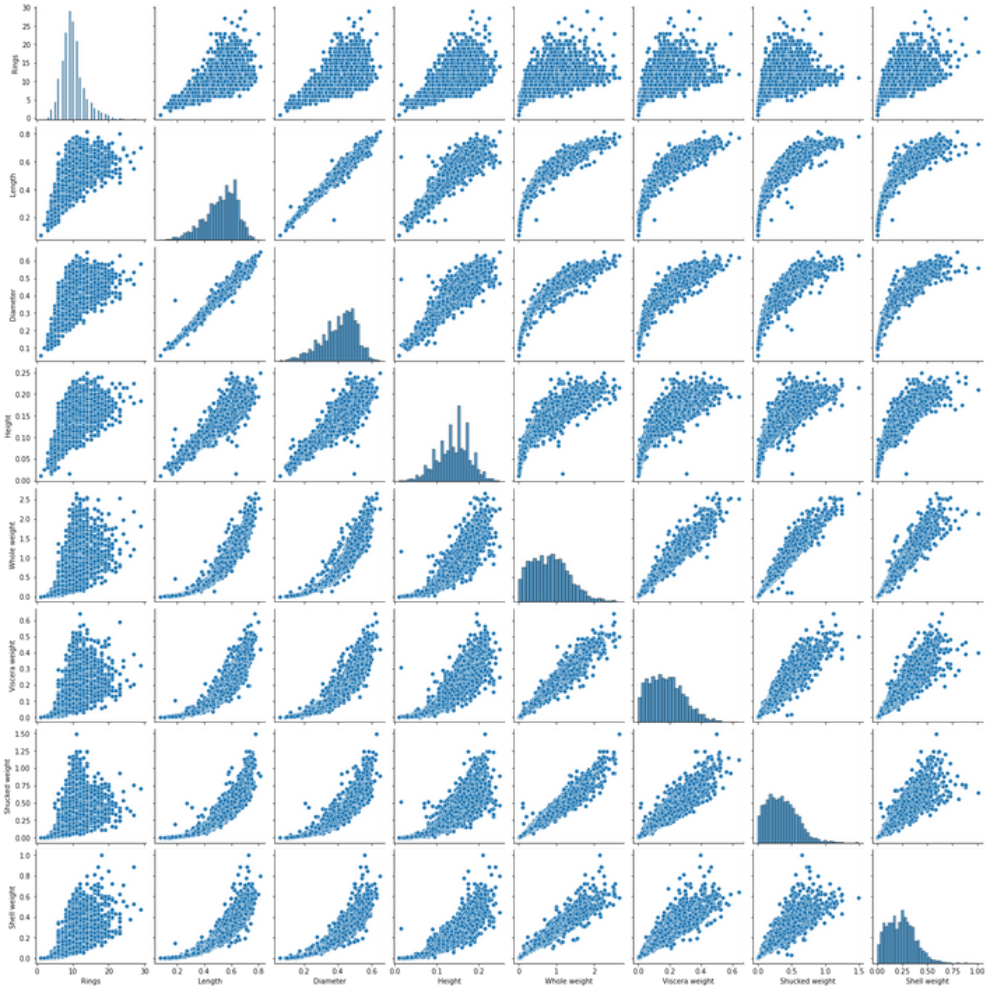


Figure 4: Correlation scatterplot matrix

As anticipated, there is a strong correlation between the height, diameter, and length of the abalone. Specifically, the length and diameter exhibit a very high degree of correlation, as the diameter is essentially determined by the length. This suggests that it may be possible to discard one of these features during the modeling process. Additionally, the length, diameter, and height also show some correlation with the number of rings, although to a lesser extent than the previously mentioned variables. The correlation matrix presented below provides further evidence supporting the aforementioned observations.

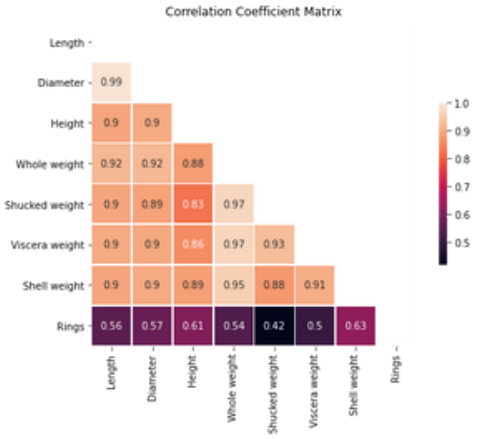


Figure 5: Correlation coefficient matrix

This correlation matrix helps identify potential multicollinearity issues, where highly correlated variables may introduce redundancy in the model. In such cases, it may be necessary to remove or consolidate correlated variables to improve the model's performance and interpretability.

The Kruskal-Wallis test was performed to determine if there is a statistically significant difference in the distribution of rings based on sex. The test yielded significant results, indicating that there is a notable difference in the rings distribution among different sexes.

Partial Least Square Analysis

PLS analysis can be used as an initial step for more advanced analysis. It can be used to identify whether there is a possibility to have observational clusters, variable clusters, outliers etc.

A partial least squares regression model will be fitted with the response variable being the number of rings. The first component accounts for 73.68% of the total variation while the second component explains 2.89% of the variation.

The score plot does not show distinct groupings of observations. However, a cluster analysis will be performed for further reference.

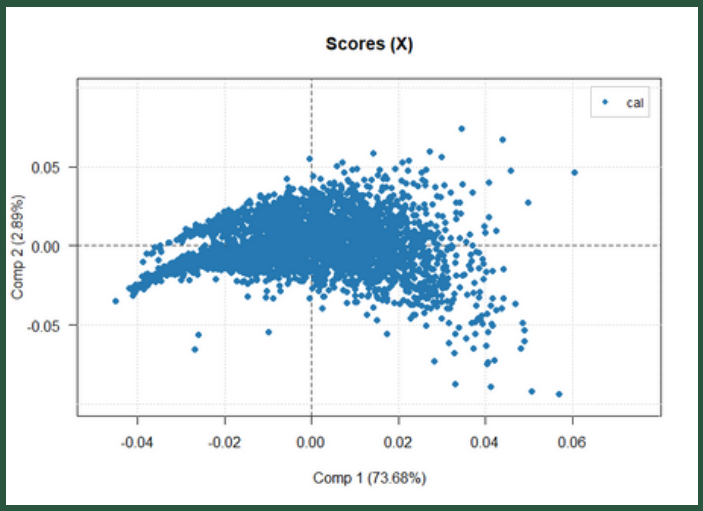


Figure 6: Score plot of PLS

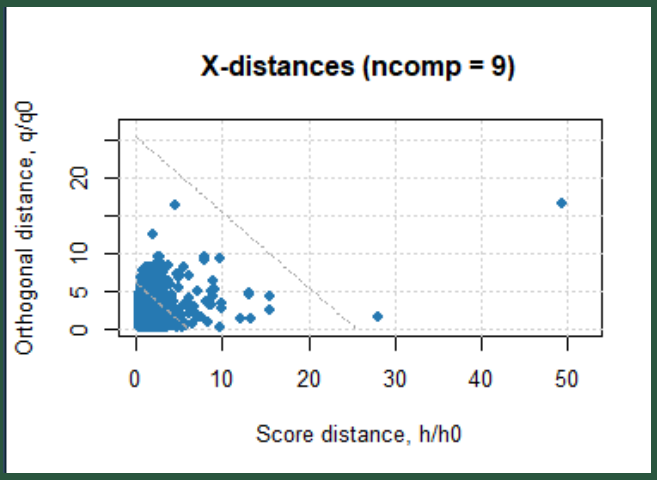


Figure 7: Scatterplot between score distances and orthogonal distances

Figure 8: Loadings plot

The X-distances plot indicates that out training dataset may not contain significant number of outliers. The loadings plot suggest that there may be some variable clusters present. As some predictor points are very close to each other in the loadings plot, it indicates that these predictors are highly correlated with each other. Having highly correlated predictors can lead to multicollinearity issues in regression models. In such cases, it may be necessary to consider removing one or more of the highly correlated predictors or performing dimensionality reduction techniques to address multicollinearity and improve the model's performance.

As the response variable is not closer to the predictor variables, it suggests that there may not be a strong linear relationship between the response variable and the predictor variables. This means that the predictors may not be effective in explaining the variation in the response variable. It could indicate that additional or different predictors may be needed to better capture the relationship with the response variable.

Cluster Analysis

Clustering techniques allow us to group data objects into similar classes in such a way that items within a group share similar characteristics, while items in different groups are not similar at all.

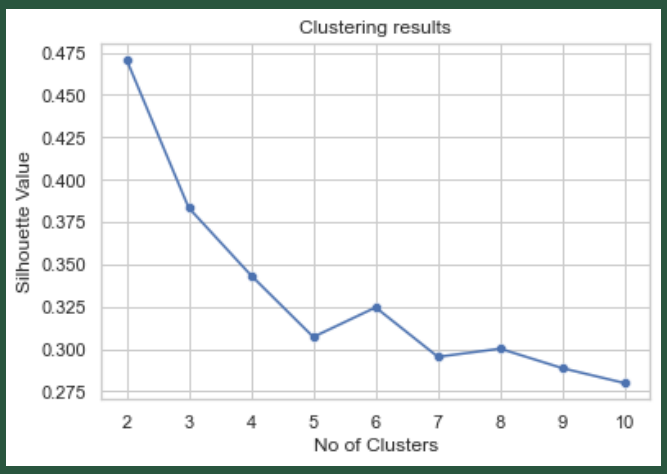


Figure 9: Clustering results

The training dataset contains both categorical and continuous data. Hence the kproto algorithm will be used to identify the clusters. The following figure illustrates how the mean silhouette values will change according to the number of clusters. We can identify the optimal number of clusters as two.

Although we could identify clusters, the mean silhouette value (0.47) indicates weaker clustering or potential misclassifications as it's not closer to 1.

Mean of variables according to each cluster:

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
Cluster 1	0.611	0.480	0.166	1.189	0.515	0.259	0.339	11.498
Cluster 2	0.419	0.321	0.107	0.397	0.173	0.086	0.119	8.103

Table 2: Clusters

Male and female abalones dominate cluster 1 while infant abalones dominate cluster 2. This would be the possible reason for identifying higher mean values for cluster 1 and lower values for cluster 2.

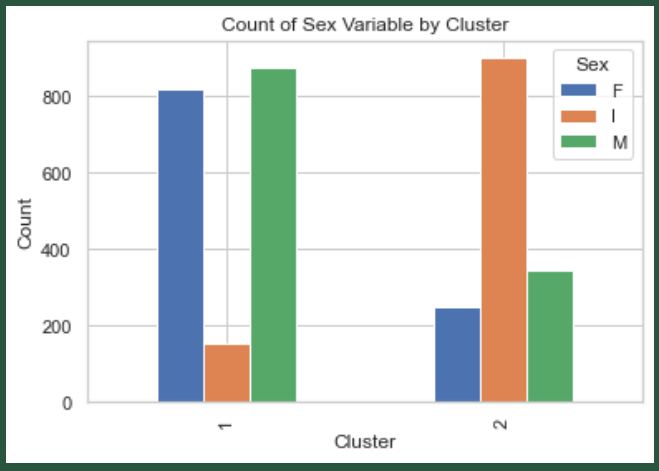


Figure 10: Bar graph of count of sex variables by cluster

Development and Assessment of the Predictive Models

Multiple Linear Regression

Initially, multiple linear regression model will be fitted to predict the number of rings. The summary of the model is as follows.

OLS Regression Results

Dep. Variable:	Rings	R-squared:	0.544
Model:	OLS	Adj. R-squared:	0.543
Method:	Least Squares	F-statistic:	497.2
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	0.00
Time:	23:46:52	Log-Likelihood:	-7373.2
No. Observations:	3339	AIC:	1.476e+04
Df Residuals:	3330	BIC:	1.482e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.7047	0.281	9.618	0.000	2.153	3.256
Diameter	7.9876	1.177	6.789	0.000	5.681	10.295
Height	23.3689	2.527	9.246	0.000	18.414	28.324
Whole weight	9.2846	0.804	11.541	0.000	7.707	10.862
Shucked weight	-19.7949	0.905	-21.883	0.000	-21.569	-18.021
Viscera weight	-11.4549	1.441	-7.947	0.000	-14.281	-8.629
Shell weight	6.9948	1.265	5.528	0.000	4.514	9.475
Sex_F	0.7542	0.116	6.519	0.000	0.527	0.981
Sex_M	0.8740	0.109	8.049	0.000	0.661	1.087

Omnibus:	788.594	Durbin-Watson:	2.028
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2227.722
Skew:	1.231	Prob(JB):	0.00
Kurtosis:	6.155	Cond. No.	109.

Figure 11: OLS regression results

As our target variable is a discrete variable, the predicted values obtained from the model will be rounded off to obtain a valid prediction for the number of rings.

Calculated RMSE and MAPE values are as follows.

Training RMSE: 2.227
Test RMSE: 2.098

Training MAPE: 15.33%
Test MAPE: 15.78%

Correlation between the actual and predicted values of the test data is 0.73. It implies that the predicted values are bit deviated from the actual values.

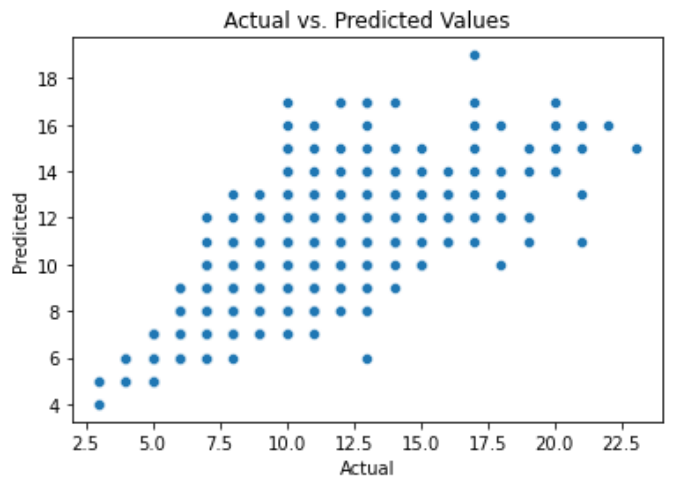


Figure 12: Scatterplot of correlation between Actual vs Predicted values of MLR

Checking assumptions of MLR Model

- ***Linearity - The relationship between the predictors and response is assumed to be linear.***

The residual vs fitted value plot displays a non-linear pattern resembling a funnel shape, which indicates the existence of a non-linear relationship between the response variable and predictor variables.
This suggests that the model may not be able to fully capture the underlying relationship between the variables, and that alternative models or transformations of the predictor and/or response variables may be necessary to improve the accuracy and fit of the model.

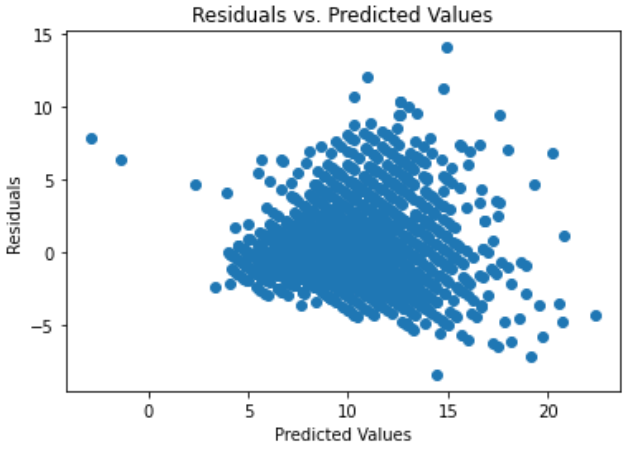


Figure 13: Scatterplot of residual vs fitted values

- ***Residuals are normally distributed with mean zero.***

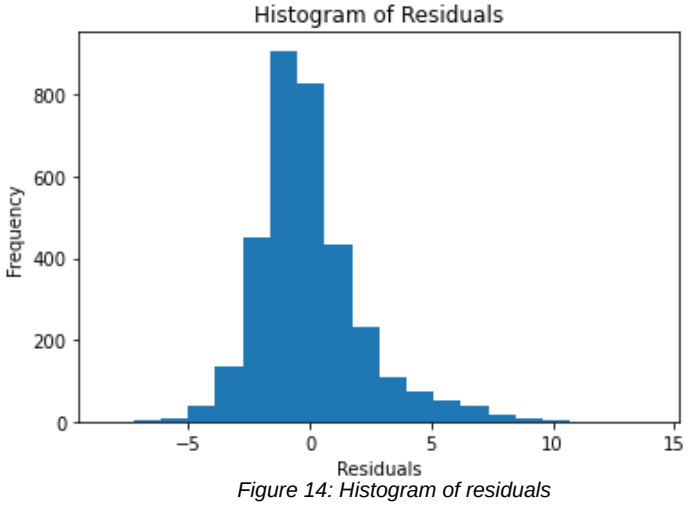


Figure 14: Histogram of residuals

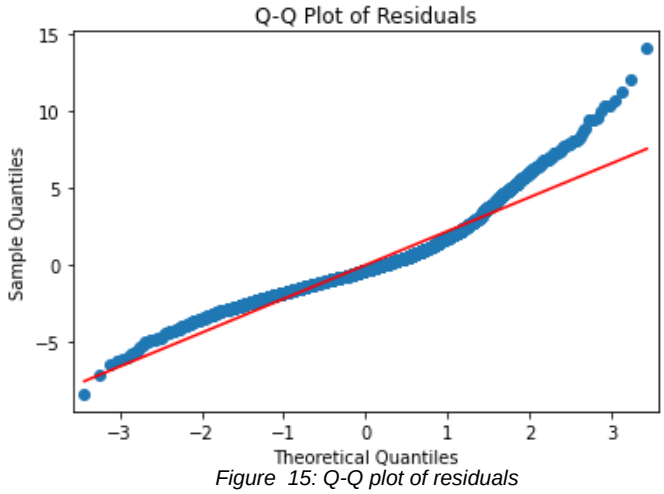


Figure 15: Q-Q plot of residuals

The histogram of the residuals indicates that they are mostly normally distributed around a mean of zero in the middle range, but with some skewness. However, there are few observations in the tails of the distribution. The QQ plot of the residuals suggests that a relatively small number of points lie on the reference line, indicating that the normality assumption of the residuals may not be met. This suggests that the model may not be the best fit for the data and that further investigation or adjustment may be necessary.

- ***Independence of the residual error terms***

We performed a Durbin Watson Test using python and it provided the Durbin Watson test statistic as 2.028 with a p-value of 0 which implies that the independence assumption is not satisfied.

- ***Homoscedasticity which implies the residuals have constant variance..***

Residual vs fitted value plot shows a funnel shape, with the variability of residuals increasing as the fitted values increase .
This can be due to heteroscedasticity, meaning that the variance of the residuals is not constant across the range of the predictor variable. This can lead to biased estimates and inaccurate predictions. Therefore, further investigation and potentially modifying the model is necessary to improve its performance.

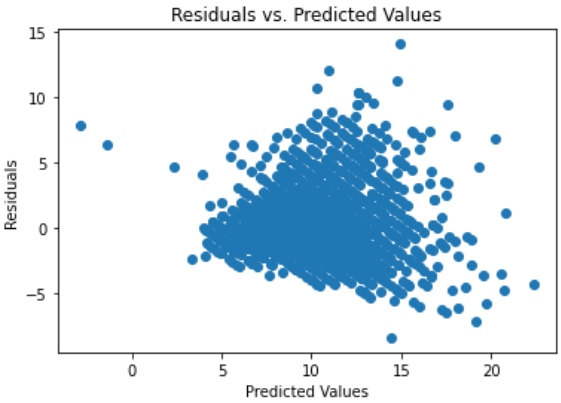


Figure 13: Scatterplot of residual vs fitted values

- ***Multicollinearity is not present***

Feature	VIF-Score
Diameter	9.3153
Height	6.5578
Whole weight	106.8587
Shucked weight	27.6457
Viscera weight	17.1481
Shell weight	21.4094
Sex_F	1.9990
Sex_M	1.8774

Table 3: VIF scores of each variable

Based on VIF values it is clear that multicollinearity does exits.
In particular, the variable with the VIF value of 106.8587 suggests that it may be highly correlated with one or more of the other predictor variables.

Remarks :

- We found out that some of the model assumptions are not satisfied. Violating the assumptions of multiple linear regression can lead to biased and inefficient estimates, incorrect conclusions, and inaccurate predictions. Therefore more different models will be fitted and evaluated.

Regularization Techniques

Regularization techniques such as Lasso, Ridge, and Elastic Net have become increasingly important in the field of machine learning and statistical modeling. These techniques are used to prevent overfitting, improve model accuracy, and handle multicollinearity in the data.

Lasso Regression

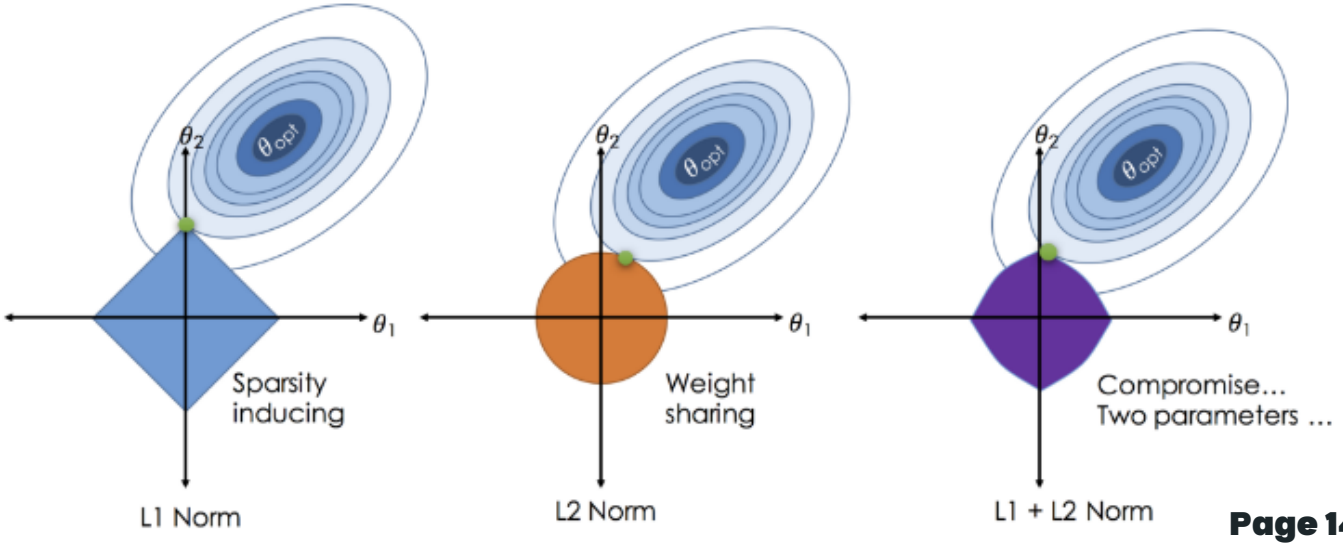
Lasso, or Least Absolute Shrinkage and Selection Operator, is a regularization technique used to perform variable selection in linear regression. Lasso adds a penalty term to the regression equation that forces some of the coefficients to be zero, effectively eliminating some of the variables from the model.

Ridge Regression

Ridge regression is another regularization technique used in linear regression. Similar to Lasso, Ridge adds a penalty term to the regression equation. However, the penalty term in Ridge is the sum of the squares of the coefficients, as opposed to the absolute values in Lasso. This results in Ridge shrinking the coefficients towards zero, but not necessarily to zero.

Elastic Net Regression

Elastic Net is a combination of Lasso and Ridge regression. Elastic Net adds both L1 (Lasso) and L2 (Ridge) penalties to the regression equation. This allows Elastic Net to handle multicollinearity in the data and perform variable selection at the same time. Elastic Net is useful when there are many correlated predictor variables in the model.



Model	Training RMSE	Test RMSE	R Squared Value	MAPE
Ridge	2.205	2.232	0.5398	16.308%
Lasso	3.206	3.291	-0.00027	27.338%
Elastic-Net	3.207	3.290	-0.00026	27.337%

Table 4: Ride, Lasso, Elastic Net Summary

Based on the findings from running regression techniques like Lasso, Ridge, and Elastic-Net, the results suggest that the machine learning algorithm applied to the dataset has produced precise outcomes. This is evident from the small disparity between the Training RMSE and Test RMSE, indicating that the models were not overfitted.

RMSE is the root mean square error and it measures the average difference between the predicted values and the actual values. The R squared value is the coefficient of determination and it speaks how much of a variation of a dependent variable is explained by the independent variables in the model. Both measures the goodness of a regression model.

Based on the given values, it can be concluded that the Ridge Regression model is performing better than Lasso and Elastic-Net Regression models. The training and testing RMSE values of Ridge Regression are comparatively lower than Lasso and Elastic-Net Regression. This indicates that the Ridge Regression model is able to predict the age of abalone with lesser errors.

However, the R-squared value of Lasso and Elastic-Net Regression models are negative, indicating that the models are not suitable for predicting the age of abalone as they are not able to explain the variance in the data. The MAPE values of Lasso and Elastic-Net Regression models are also higher than Ridge Regression model, indicating that the predictions made by these models are less accurate.

Even though the Ridge Regression model appears to be the better choice for predicting the age of abalone than Lasso and Elastic-Net Regression models the R square value is much lower that the predictive power would be lower. Therefore, further analysis and experimentation might be required to improve the performance of the models or to find a model using other regression techniques such as Random Forest or Artificial Neural Networks(ANN).

Random Forest Model

Random Forest is a robust ensemble learning method that has broad applications for solving prediction problems, including regression and classification. As an ensemble of decision trees, Random Forest offers the advantages of decision trees, such as high accuracy and ease of use without requiring data scaling.

In this study, we aim to forecast the rings of abalones using the abalone data set in kaggle by random forest model using python.

First we remove any records where the height of the abalone was zero, which may have been due to measurement errors or other issues. Additionally, one outlier was identified where the height was greater than 1 and was also removed from the dataset. Then, we have to split our data into training and test sets. As a training set, we will take 80% of all rows and use 20% as test data. One of the advantages of the Random Forest algorithm is that it does not require data scaling, as previously stated. To apply this random forest technique, all we need to do is define the features and the target we're attempting to predict.

Random forest models have several parameters that can be tuned to optimize their performance on a given dataset.
ex: n_estimaters, max_depth, min_samples_split,min_samples_leaf,max_features

To tune these parameters, we used technique of grid search method with cross-validation. Grid search involves exhaustively testing all possible combinations of parameter values within a specified range.

Using those best parameters, we have modeled a random forest regression model for our dataset. Then the model's accuracy is evaluated by comparing the predicted value to the actual values in the test data.

Calculated RMSE and MAPE values are as follows.

Train RMSE: 1.740
Test RMSE: 2.035

Train R^2: 0.715
Test R^2: 0.559

Testing MAPE: 14.403%

Artificial Neural Network Model

Artificial Neural Network (ANN) is a mathematical model that is driven by the functional feature of biological neural networks. A neural network contains an interconnected set of artificial neurons, and it processes information using a connectionist form to computation.
As a rule, an ANN is an adaptive system that adjusts its structure based on external or internal data that runs over the network during the learning process. ANN learning can be either supervised or unsupervised. Here we used supervised learning technique. As supervised training proceeds, the neural network is taken through several iterations, or epochs, until the actual output of the neural network matches the expected output, with a reasonably small error.

The 08 physical measurements of abalones were classified as input variables. The output variable represents the predicted age (rings) of abalone based on those inputs. Following steps are used to build the neural network.

1) Data Normalization

In this study we normalized the number of rings variable to be in the range between 0 and 1. We converted the categorical attribute Sex to numeric values then the numeric value was normalized to be between zero and one. We used the equation Xi = (Xi – Xmin) / (Xmax –Xmin) for the normalization and for that MinMaxScaler() function in python is used.

2) Building the ANN Model
Using JNN tool as well as using python we have build a multilayer ANN model.

The proposed model consists of five Layers: Input Layer with 8 nodes, First Hidden Layer with 5 nodes, Second Layer with 1 one, Third Layer with 7 nodes, and Output Layer with one node as can be seen in Figure 3. We have set the parameters of the proposed model as follows: Learning Rate 0.06 and the Momentum to be 0.08, and Average Error rate to be 0.01

3) Evaluating the ANN model
The abalone dataset consists of 4177 samples with 9 attributes. We imported the preprocessed CSV file of the abalone dataset into GoogleColab and divided the imported dataset into two sets randomly. The Training consists of approximately 80% (3340 observations) and the testing set consists of 20% of the dataset (835 observations). To avoid overfitting and find the best parameters for ANN model (ex: Batch size, Epochs), we have divided training set again as training and validation sets.
After making sure that the parameter control was set properly to best parameter values, we started training the ANN model and kept eye on the learning curve, error loss and validation accuracy. We trained the ANN model for 500 cycles.

After training the ANN model the **best accuracy we got was 85%.**

Summary of the proposed ANN model :

Training RMSE: 2.18
Testing RMSE: 2.22

Training R-squared: 0.54
Testing R-squared: 0.51

Testing MAPE: 15%
Accuracy : 85%

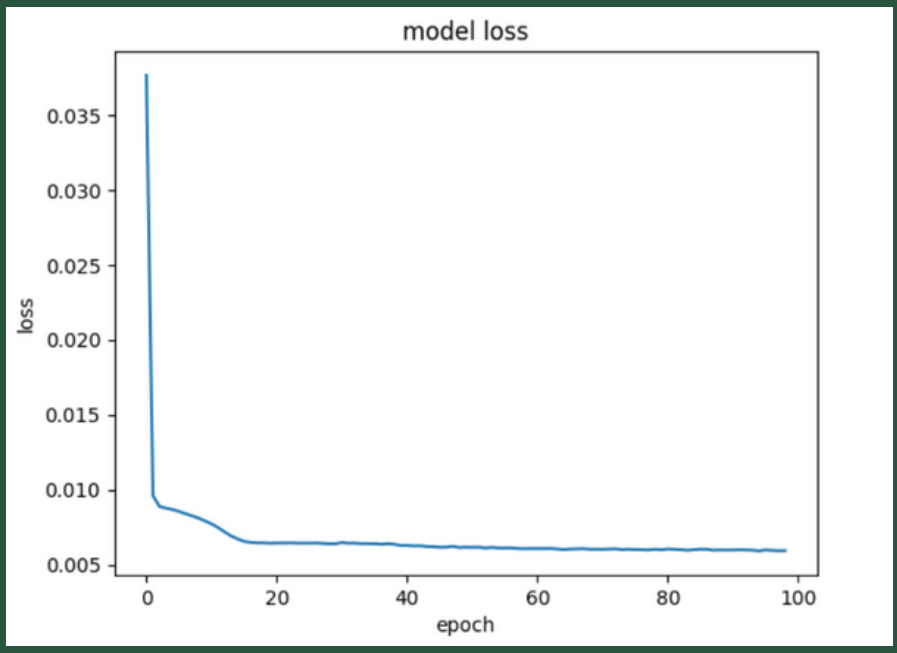


Figure 16:Summary of Model Loss- ANN

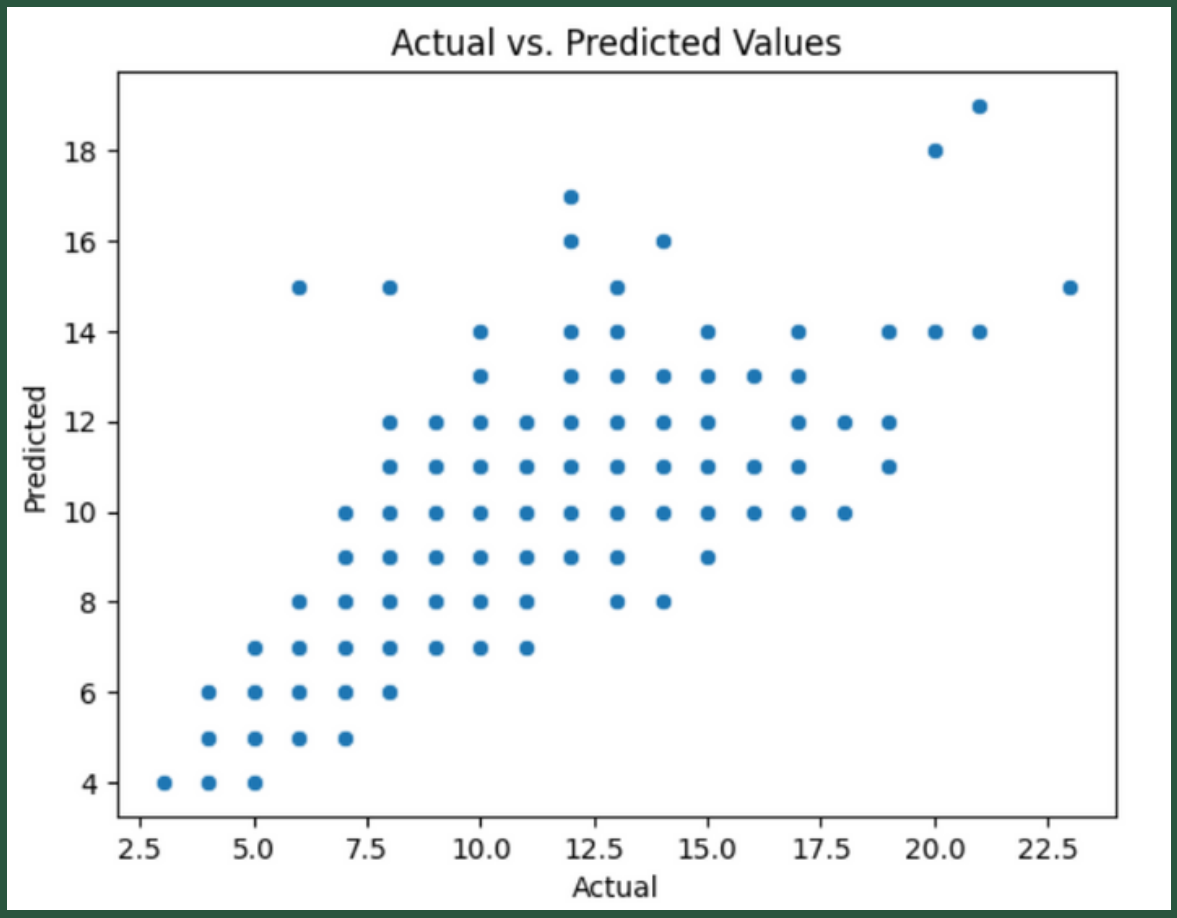
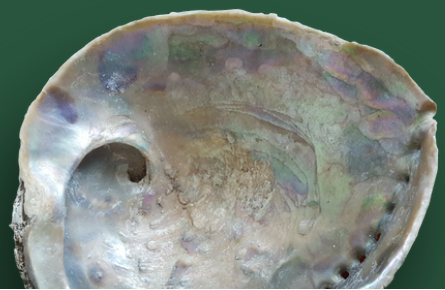


Figure 17: Actual vs Predicted values - ANN



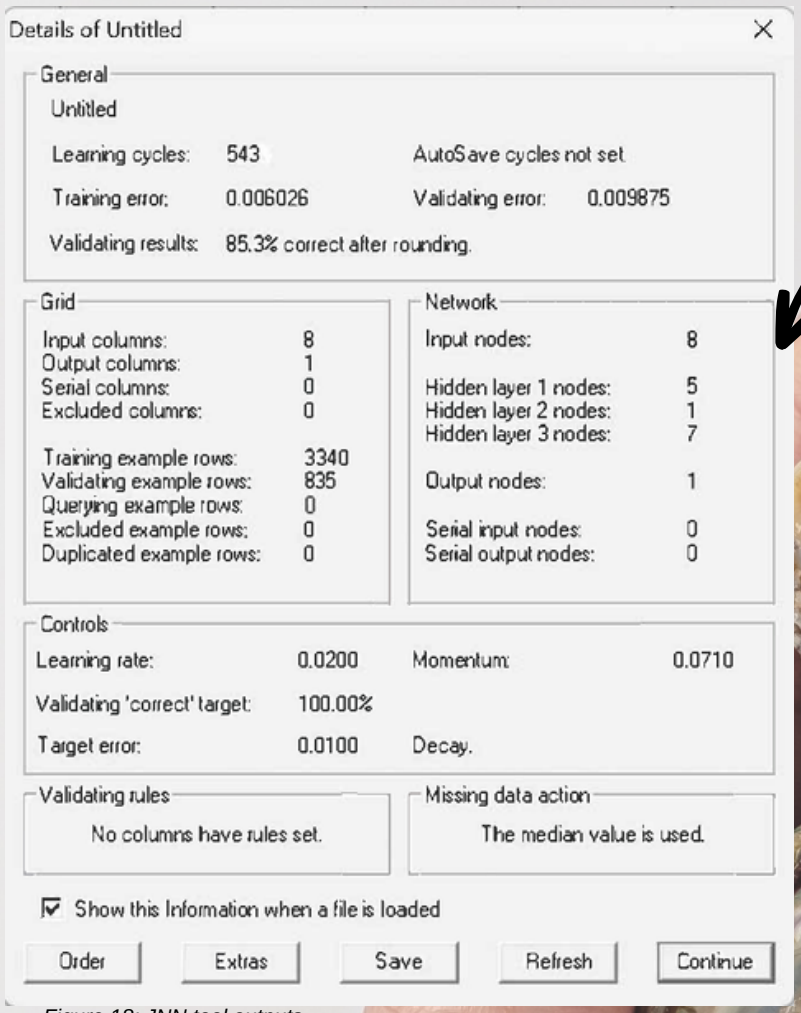


Figure 18: JNN tool outputs

JNN Tool outputs

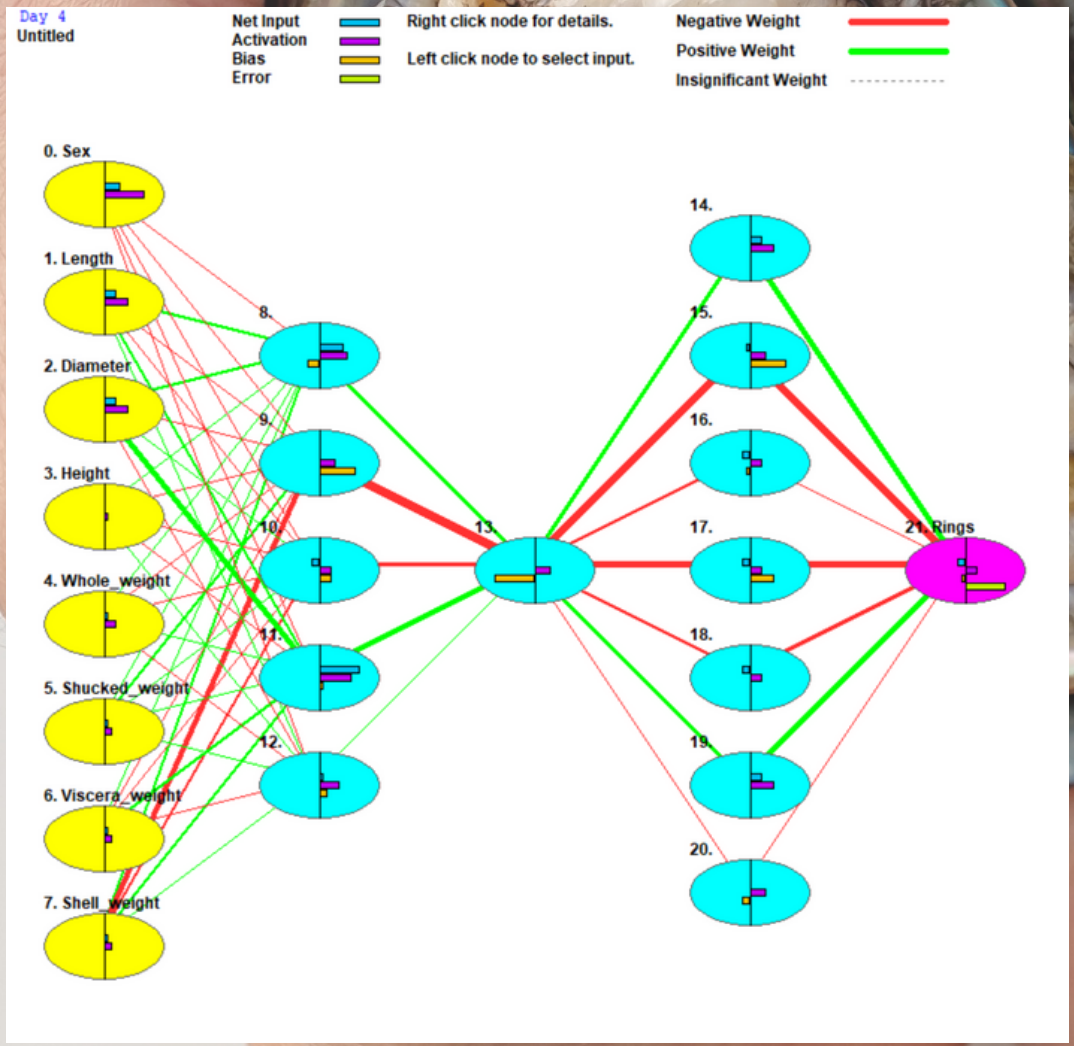


Figure 19: ANN model using JNN

Selecting the best model

	RMSE	R Squared	MAPE	Correlation y_test vs y_test_predict
MLR	2.10	0.54	15.78%	73.03%
LASSO	3.29	-0.00027	27.34%	7.54e-15%
RIDGE	2.23	0.54	16.31%	77.17%
ELASTIC NET	3.29	-0.00026	27.34%	7.54e-15%
RANDOM FOREST	2.04	0.56	14.40%	76.02% ✓
ANN	2.23	0.51	15.4%	72.04% ✓

Table 5: Overall summary of the predictive models

When comparing the above models, we can clearly see we can get better accuracy for predicting abalones' age (rings) by using random forest model and ANN model. (Accuracy is more than 85%)

Conclusions

- The measurements of abalones, including their length, diameter, and weight, can be utilized as predictors to estimate their age, eliminating the need for the labor-intensive and time-consuming process of counting rings. This has significant implications for the management and preservation of abalone populations, as it offers a more efficient and economical approach to estimating their age and monitoring shifts in population characteristics over time.

- It should be emphasized that accurately predicting the age of abalones based on their physical measurements may necessitate supplementary data, such as weather conditions and geographic location, to attain the utmost precision. Hence, it is crucial for further investigations to continue refining and exploring predictive models that incorporate both physical measurements and environmental factors to estimate the age of abalones more effectively.

Appendix

Python and R codes : st3082group10/Abalone_Age_Prediction (github.com)