TÉCNICO LISBOA

Advanced Automation

Masters in Mechanical Engineering

Abalone Age Prediction [EN]

Authors:

Margarida Gralha (100222) margarida.gralha@tecnico.ulisboa.pt
Tiago Neves (100272) tiagoneves540@tecnico.ulisboa.pt
Rodrigo Amorim (100702) rodrigo.j.amorim@tecnico.ulisboa.pt

Group 1

2023/2024 – 1st Semester, P2

Contents

1 Objective

The objective of this project is to estimate the age of the Abalone, a type of marine snail, as seen in Figure 1, by predicting its number of rings from physical measurements. The real way of determining the age of an Abalone revolves around cutting its shell through the cone, staining it, and counting the number of rings through a microscope. This process is very tedious and time-consuming; therefore, there's a benefit to trying to predict the age of the Abalone from physical measurements.

Figure 1: Abalone

2 Details of the dataset

2.1 List of features

The dataset used contains 4177 data points and 9 features; no missing values are present. A description of the features can be found in Table 1.

Table 1: Abalone dataset features

Feature	Type	Description	Units
Sex	Categorical	M, F, and I (infant)	/
Length	Continuous	Longest shell measurement	mm
Diameter	Continuous	Perpendicular to length	mm
Height	Continuous	With meat inside the shell	mm
Whole_weight	Continuous	Whole Abalone	grams
Shucked_weight	Continuous	Weight of the meat	grams
Viscera_weight	Continuous	Gut weight (after bleeding)	grams
Shell_weight	Continuous	After being dried	grams
Rings	Integer	+ 1.5 gives the age in years	/

Since the objective is to predict the number of Rings, this problem will be approached as a Regression one. As such, Multiple Linear Regression with Model Selection, Decision Trees and Neural Networks will be employed to tackle the task at hand.

Because our problem is a regression one, the categorical feature, Sex, will be one-hot encoded into three different features, Sex_M, Sex_F and Sex_I.

The correlation matrix of all the continuous features can be found in the annex, Figure 8.

2.2 Approach for the categorical feature

When investigating the influence of sex, it was inferred that this feature does not have an important influence on the number of Rings, save for whether an abalone is an infant or not. This can be better understood with the following results:

Table 2: Mean and median of abalone ring amount separated by sex

	Male	Female	Infant
Mean	11	11	8
Median	10	10	8

This conclusion is taken into account when choosing the relevant features for each of the methods used in the project, i.e., only Sex_I will be taken into consideration.

2.3 Data splitting

The data will be split similarly for each method: 10% of the dataset will be isolated from the rest, serving as the test set, which will be equal everywhere, so that models from different methods can be compared in the end. The remaining 90% will be used for training, hyperparameter tuning and validation. This way, the possibility of data leakage is eliminated, which could make the models fail to generalize to new, unseen data.

3 Regression

In this section, regression methods were used to predict the number of Rings. A single numeric predictor was first used to observe the behavior of the data, and multiple linear regressions were performed thereafter. Model selection methods were also used in an attempt to find the most adequate predictors, as well as their corresponding parameters.

3.1 Simple Linear Regression

Based on the correlation matrix of the numeric features, Figure 8, the predictor Shell weight was chosen to predict the number of Rings, since it was the feature with the highest correlation with the latter. After splitting the data, the following results were obtained:

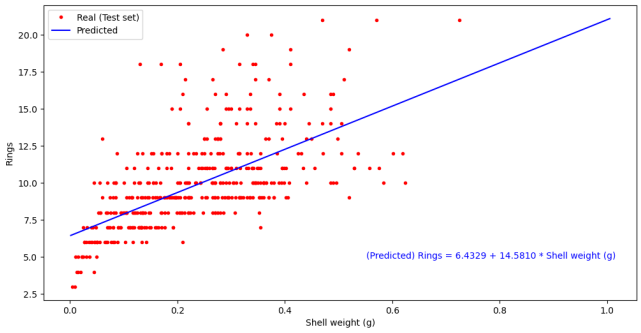


Figure 2: Simple linear regression and test data points (Test MSE: 6.2341; R^2: 0.3629)

The results show the model yields worse predictions, the higher the magnitude of the predictor. This suggests that predicting larger ring amounts may be even more challenging. Clearly, one predictor does not suffice to satisfactorily predict the amount of Rings.

3.2 Multiple Linear Regression

Following the previous conclusions, multiple linear regression with all the continuous features and the Sex_I feature was performed. The following expression was obtained:

$$\begin{aligned}\text{Rings} = {} & 3.8991 - 0.6801 \cdot \text{Length\_mm} + 11.3616 \cdot \text{Diameter\_mm} \\ & + 10.6141 \cdot \text{Height\_mm} + 8.6733 \cdot \text{Whole weight\_g} - 19.6357 \cdot \text{Shucked weight\_g} \\ & - 9.5101 \cdot \text{Viscera weight\_g} + 8.8391 \cdot \text{Shell weight\_g} - 0.8251 \cdot \text{Sex\_I} \\ \text{Test MSE}: {} & 4.4565 \\ R^2: {} & 0.5446\end{aligned}$$

Even though all continuous features (and the relevant categorical one) were considered, the performance of the model is still suboptimal.

Given these results, the influence of the interactions between predictors was analyzed. For this purpose, and to avoid an overly large number of model parameters, three predictors were chosen, once again based on the correlation matrix of Figure 8. Two models will be built upon those predictors – one with second-order interactions between predictors, and another without –, and their performances compared. Data was split in the same way it was done for the previous model. The expressions and results follow:

$$\begin{aligned}\text{Rings} = {} & 1.5352 + 2.5919 \cdot \text{Length\_mm} + 20.2847 \cdot \text{Diameter\_mm} \\ & + 49.0531 \cdot \text{Shell weight\_g} - 32.3857 \cdot \text{Length\_mm} \cdot \text{Diameter\_mm} \\ & - 40.5012 \cdot \text{Length\_mm} \cdot \text{Shell weight\_g} + 0.0985 \cdot \text{Diameter\_mm} \cdot \text{Shell weight\_g}\end{aligned}$$

$$\text{Rings} = 6.6502 - 13.1537 \cdot \text{Length\_mm} + 16.7210 \cdot \text{Diameter\_mm} + 13.9674 \cdot \text{Shell weight\_g}$$

Table 3: Test MSE and R^2 of multiple linear regression models with and without interactions

	Test MSE	$\boldsymbol{R^2}$
With Interactions	5.1502	0.4737
Without Interactions	6.1654	0.37

As it can be concluded from Table 3, the influence of the interactions between predictors is substantial, both in terms of Test MSE and R^2. In fact, the performance of the model that considers predictor interactions is fairly close to that of which considers all continuous features (and the relevant categorical one), while only taking into account three predictors. This suggests that, if further multiple linear regression models were to be conceived, interactions between predictors ought to be considered.

3.3 Model Selection

In order to select the best features, Best Subset Selection was employed, as well as Ridge Regression. For both of the methods, the data was normalized with a Z-score.

3.3.1 Best Subset Selection

Using BIC as the sorting performance index, the following results were obtained, Table 4:

Table 4: Best subset selection results (5 best combinations, sorted by highest BIC)

Feature number	Feature list	BIC
7	Diameter (mm), Height (mm), Whole weight (g), Shucked weight (g), Viscera weight (g), Shell weight (g), Sex_I	7840.03597
8	Length (mm), Diameter (mm), Height (mm), Whole weight (g), Shucked weight (g), Viscera weight (g), Shell weight (g), Sex_I	7848.14025
7	Length (mm), Height (mm), Whole weight (g), Shucked weight (g), Viscera weight (g), Shell weight (g), Sex_I	7863.29295
6	Diameter (mm), Whole weight (g), Shucked weight (g), Viscera weight (g), Shell weight (g), Sex_I	7876.779095
6	Length (mm), Diameter (mm), Whole weight (g), Shucked weight (g), Viscera weight (g), Shell weight (g), Sex_I	7881.242701

Based on these results, the model that will be considered for further performance analysis and comparison will be that of the first row of Table 4.

Performing the Validation set approach with the isolated 10% data, this model has a test MSE of 4.4546 and an R^2 of 0.5448.

The conclusion taken by inspection of the MSE and R^2 values is that these are very similar to the ones obtained when using all the numeric features and the Sex_I one, as seen in the Multiple Linear Regression, section 3.2. This is expected, since the only feature that wasn't selected in the Best Subset Selection was Length (mm), which is highly correlated with Diameter (mm), when observing the correlation matrix in Figure 8. Therefore, in reality, the model obtained in either case is approximately equal, yielding the similar values of test MSE and R^2.

In conclusion, the best model obtained using this methodology has the following structure:

$$\begin{aligned}\text{Rings} = 3.9253 &- 0.5523 \cdot \text{Diameter\_mm} + 11.2127 \cdot \text{Height\_mm} + 10.5243 \cdot \text{Whole weight\_g} \\ &+ 8.6553 \cdot \text{Shucked weight\_g} - 19.6213 \cdot \text{Viscera weight\_g} - 9.5360 \cdot \text{Shell weight\_g} \\ &+ 8.8415 \cdot \text{Sex I}\end{aligned}$$

3.3.2 Ridge Regression

The Ridge Regression model selection defines the coefficients for a multiple linear regression by determining a tuning parameter, λ. In that sense, the data is, once again, divided in the way presented in Section 2.3, and λ is selected as the highest value possible for the lowest Mean Squared Error, assuring a relatively simple model.

To identify the best λ, the MSE is calculated as the mean of 10-folds applied to the training data for a range of 100 values of λ between $[10^{-2}, 10^{10}]$. Once the section with the best results is defined, the process is repeated, so that there are 100 new values of λ in this specific range which, in this case, is $[10^{-1}, 10^{1}]$.

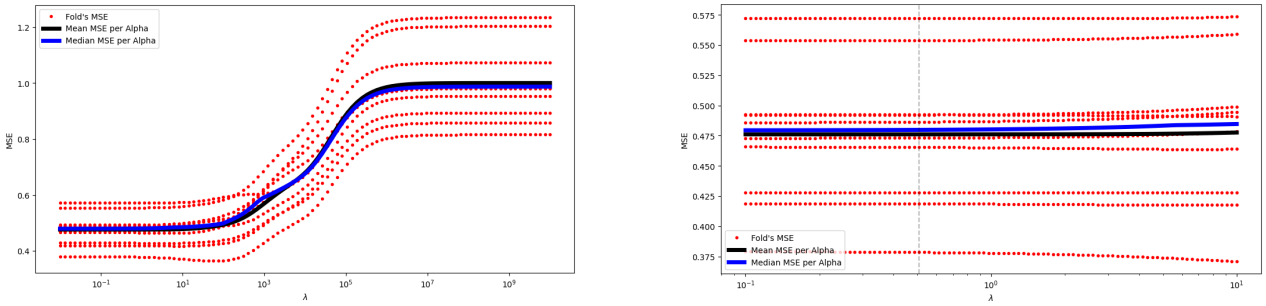


Figure 3: Mean Squared Error vs λ plots for: (Left) $\lambda \in [10^{-2}, 10^{10}]$, (Right) $\lambda \in [10^{-1}, 10^{1}]$

The acquired value of λ was 0.5094, which is desired, since, in order to prevent an over-simplification of the model or an overfitting, the value of λ should be between 0.01 and 10, as achieved [3].

The coefficients that correspond to this optimal value are:

Table 5: Table with feature coefficients for the Ridge Regression model

λ	Length (mm)	Diameter (mm)	Height (mm)	Whole weight (g)	Shucked weight (g)	Viscera weight (g)	Shell weight (g)	Sex I
0.5094	-0.0247	0.3450	0.1473	1.2885	-1.3338	-0.3169	0.3890	-0.1193

The value of the Length_mm feature has, once again, low significance for the definition of the model, and the sign of the parameters is consistent with the ones obtained in linear regression. This condition must be guaranteed when selecting λ to assure consistency when comparing models and with the meaning of the studied data.

When predicting the amount of Rings with the aforementioned model, the test MSE (on the original scale) is 4.4586.

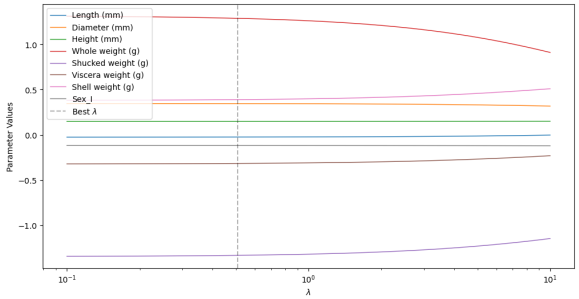


Figure 4: Coefficient values for $\lambda \in [10^{-1}, 10^{1}]$

4 Decision Trees

Regarding Decision Trees, Figure 9, three models, based on different techniques, were developed: Regression Trees, Random Forests and Gradient Boosting.

The data was split as previously stated. Hyperparameters were calculated for each technique, fitting the respective models afterwards. The test MSEs were calculated using the validation set approach, with a specific data partition reserved for this purpose, and the corresponding feature importances were plotted.

4.1 Hyperparameters

The values of the hyperparameters for every single technique were calculated by grid search with 5 folds[1], with a list of possibilities for each hyperparameter, in each specific technique. The results are found below, in Table 6.

Table 6: Hyperparameters

	max_depth	n_estimators	min_samples_split	learning_rate
Regression Tree	6	-	-	-
Random Forests	7	100	3	-
Gradient Boosting	3	1000	-	0.01

4.2 Test MSE

The values for the Test MSE for each technique can be found in Table 7.

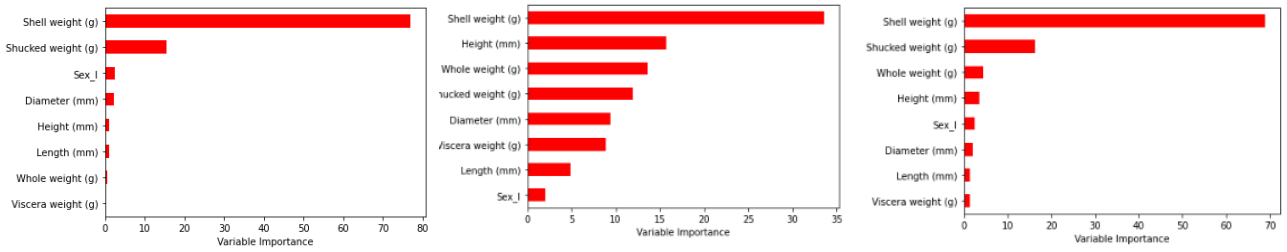
Table 7: Test MSE

	Test MSE
Regression Trees	5.0058
Random Forests	4.3683
Gradient Boosting	4.4812

The model with the lowest test MSE was the Random Forests one, meaning it is the best Decision Trees model to predict the number of Rings, and will be posteriorly used for comparison with models from other methods.

Further analyzing the values of Table 7, it can be concluded that the Regression Trees model performed the worst, which is to be expected, since this method, when compared with the other two, produces a simpler tree. Between Random Forests and Gradient Boosting, the difference in test MSE was minimal, which may be due to the fact that the latter is more sensitive to the hyperparameter tuning, when compared with the former. That being said, a less precise choice may lead to worse performances from Gradient Boosting.

4.3 Feature Importance



(a) Regression Tree **(b)** Random Forest **(c)** Gradient Boosting

Figure 5: Averaged Feature Importance for Decision Tree models

[1]Due to computational limitations

As seen in Figure 5, Shell weight is, by far, the most important feature in all three models. This may be because this feature has the highest correlation with Rings, when compared with the other features, based on the correlation matrix, from Figure 8. It's also important to note that the other weight features, Whole weight and Shucked weight, also have high feature importance in all the models, which suggests that weight as a whole is an important factor for predicting the number of Rings.

5 Neural Networks

In this section, neural networks were used to predict the number of Rings. Once the data was processed, taking into account some previous conclusions, a grid search was performed to identify the best values for some hyperparameters. Having run various neural network models, their performances and residuals were calculated for a final visualization of their quality.

5.1 Data Processing

As it has been previously concluded in this project, in terms of sex, it is more relevant whether an abalone is an infant or not, rather than whether it's male or female. With that in mind, a single dummy variable was used to take into account the aforementioned categorical feature. All the continuous features found in Table 1 were used to predict the amount of rings.

The data was split as previously mentioned in the report. In order to boost the convergence speed and generalization of the neural network models, both the input and output data were normalized for all cases.

5.2 Grid Search

Due to computational limitations, grid search was performed thrice, taking into account different sets of values for some hyperparameters in each case, as summarised below:

Table 8: 1st and 2nd Grid Search Summary (results for each case are displayed in bold)

Grid Search	Activation Function	Dropout Rate	Learning Rate	Regularization Parameter (Ridge, L2)
1st	Linear, **ReLU**	0.15, **0.20**	0.0001, **0.001**	**0.01**, 0.10
2nd	**ReLU**	0.15, **0.20**, 0.25	0.0001, 0.0005, **0.001**	0.001, **0.005**, 0.010

The existence of the second grid search is due to the fact that it was desired to test more values for the hyperparameters without creating too many different models, thus avoiding computational issues. In order to accomplish this, the activation function was arbitrated according to the previous grid search best result, and more values were added to the rest of the hyperparameters accordingly. On a side note, the learning rate was the only hyperparameter in which a value higher than the previous best result was not added in the second grid search. This is due to the fact that it was chosen for this parameter not to get too high, as it may lead to an overly fast convergence and, thus, instabilities [1].

In both cases, the neural network had 2 hidden layers with 100 and 50 neurons, respectively. 100 epochs and a batch size of 32 were chosen. In order to investigate further which architecture

for the neural network is more appropriate, a third grid search was conducted, in which all previous best results were used. The new hyperparameters, their possibilities and results follow:

Table 9: 3rd Grid Search Summary (results for each case are displayed in bold)

Grid Search	1st hidden layer neurons	2nd hidden layer neurons	Batch size
3rd	100, 150, **200**	25, 50, **100**	**32**, 64

5.3 Neural Network Models Comparison

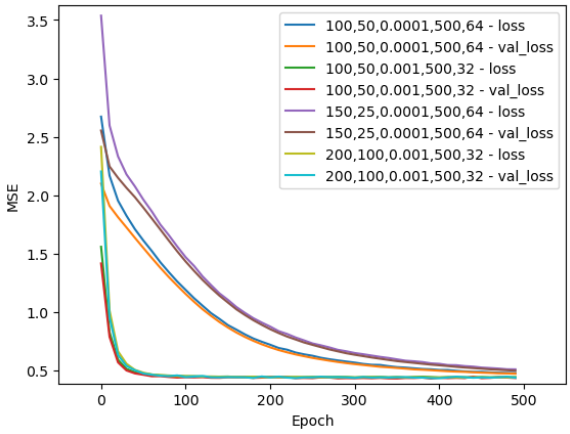


Figure 6: MSE of neural networks for validation and training data. Labels follow the sequence: Neurons in 1st hidden layer, Neurons in 2nd hidden layer, Learning Rate, Epochs, Batch Size

The previous results, Figure 6, consider a dropout rate and ridge regularization parameter of 0.2 and 0.005, respectively, as well as the ReLU activation function, as per the results of grid search. The amount of epochs was set to 500, since it was deemed that, beyond this point, the difference in the results is negligible, thus avoiding possible overfitting problems. The MSE is in relation to the normalized data.

By analyzing the results, it is concluded that the MSE of the models with a higher learning rate (of 0.001) converge much faster when compared with the rest. Though they may be appealing at first, these models must be used with caution, as they may bring problems of instability, as referenced before.

5.4 Residuals

The residuals for each of the previous models are found in Figure 10, using the test data was the reference. The respective test MSEs (on the original scale) are found on Table 10.

Table 10: Test MSE (on the original scale) of each Neural Network model (refer to Figure 6 for label meaning)

Model Label	100,50,0.0001, 500,64	100,50,0.001, 500,32	150,25,0.0001, 500,64	200,100,0.001, 500,32
Test MSE	3.9374	3.8739	3.9739	4.1738

As it was referenced earlier in this report, it was suspected that predicting larger numbers of Rings would be more challenging. This hypothesis is thus corroborated by the results of Figure 10. Despite a much better performance when compared to those of Regression and Decision Trees models, Neural Network models still struggle more in predicting larger numbers of Rings. For further comparison with models from other methods, the neural network model of label [100,50,0.001,500,32] will be used, since it has the best test MSE (on the original scale).

6 Comparisons and Conclusions

To summarize, a table with the test MSE of the best model for each method and the distribution of residuals for the best two models is showcased.

Table 11: Test MSE of the best models

	Test MSE
Multiple Linear Regression (8 features without Interactions)	4.4565
Best Subset Selection	4.4546
Random Forests	4.3683
Neural Networks	3.8739

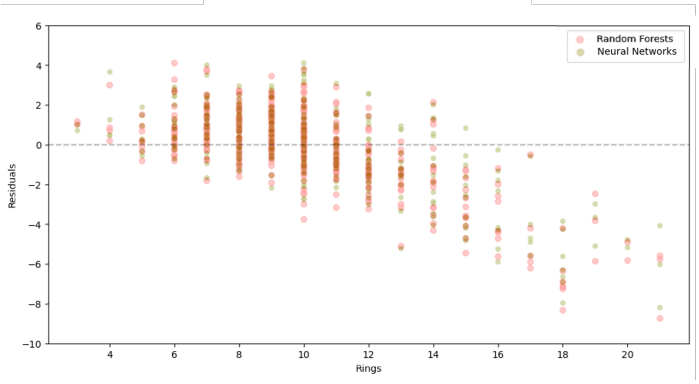


Figure 7: Model residuals (Random Forrests vs Neural Networks)

Looking into Table 11, it can be seen that the overall best model was the Neural Networks one, followed by the Random Forests and, finally, the models obtained with Multiple Linear Regression. This was expected since the Neural Networks model is the most sophisticated one, in which a relatively demanding grid search setup was used, leading to a more careful hyperparameter tuning, when compared with the rest.

From Table 11 and Section 3, it became clear that a linear regression model structure is not appropriate for the prediction of the number of Rings. Though interactions proved to be important, the performance was still poor, suggesting that its flexibility was not enough when compared to more complex models, such as the Random Forests or Neural Networks ones.

Comparing Random Forests with Neural Networks, the former had a worse test MSE, which may be due to the choice of the hyperparameters and to the relationships in the dataset.

In Figure 10, the performance of each model seems to deteriorate as the amount of Rings increases. This makes sense since there is way more data for lower values. In fact, the lower the amount of data, the worse the performance of the models, which was a trend that was also verified in linear regression methods.

Although the Neural Networks model may outperform the rest, it must be taken into account that its implementation is computationally more demanding and its interpretability more difficult. This means that, when choosing a model to predict the number of Rings, the available computational power should be taken into account, as well as the necessity for a clear interpretability of the results.

References

[1] Jason Brownlee. Understand the impact of learning rate on neural network performance.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning with Applications in Python*. Springer Cham, 2023.

[3] Charles Zaiontz. Estimating ridge regression lambda.

Appendices

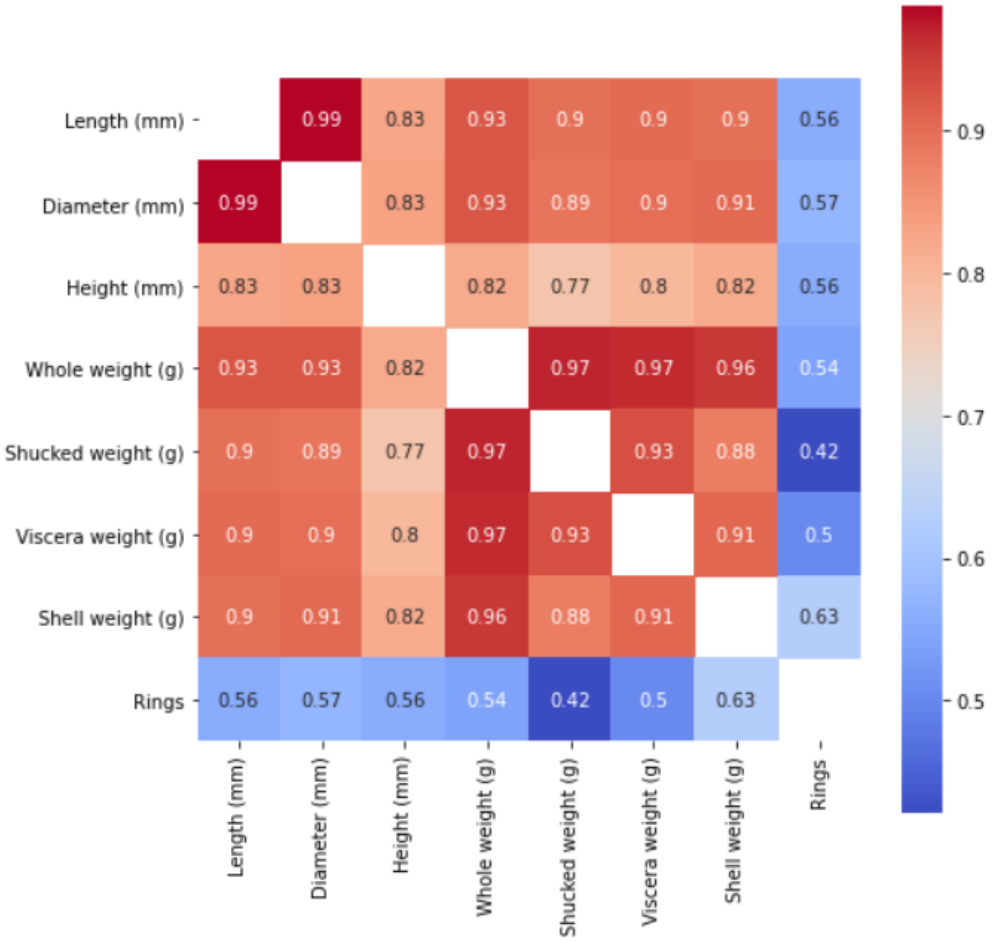


Figure 8: Correlation Matrix

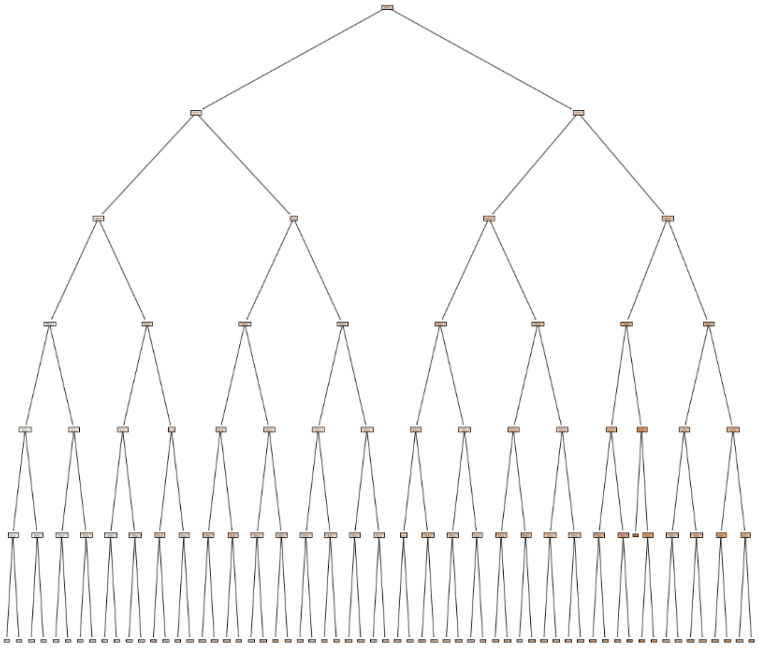


Figure 9: Regression Tree

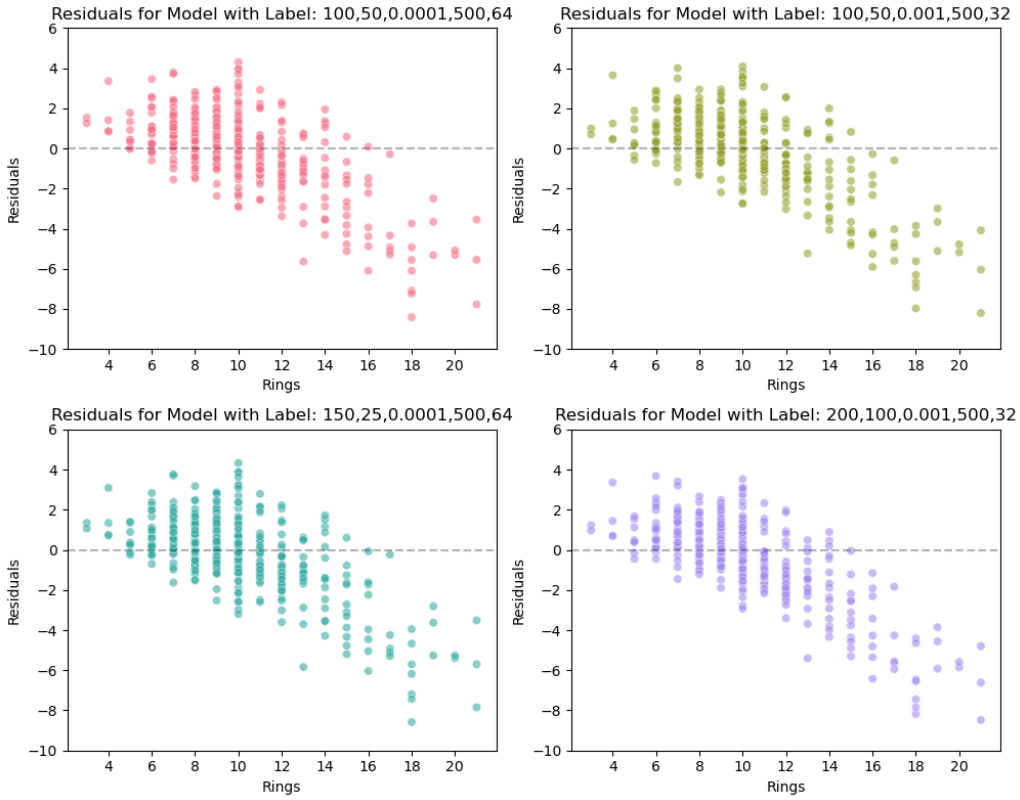


Figure 10: Residuals for each model previously obtained (refer to figure 6 for label meaning)