Abalone Age Prediction and Classification with labels

I. Introduction

a) As a very tasting marine creature, Abalone has a very high potential economic value, which is closely related to the age of Abalone. For instance, older abalone are fleshier and more fertile, and they are helpful to maintain species stability. On the contrary, younger abalones are more popular with consumers because their flesh is more tender and the dishes made from them are more flavorsome[1]. Furthermore, the study of the age of abalone is also of great value in the study of marine ecosystems. Consequently, knowing the age of the abalone appears to be very important. However, nowadays, when traditional fishermen catch abalone, the age of the abalone can only be known by cutting the shell of the abalone and counting the number of dark bands in the profile, and usually it is necessary to add 1.5 to the number of rings in order to eliminate the effect of uncertainty[2]. This process is very time-consuming and results in too much cost for fishermen to catch. With the help of machine learning, we can use the physical measurements of abalone, such as height and weight, to predict the age of abalone from the appearance.

b) Normally, it's useless to know the exact number of abalone's age. Consequently, I have divided the age of abalone into three categories based on their age range – young (0-8), medium (9-10) and old (older than 11). In figure 1 shows the basic information of the dataset. There're several physical measurements of abalone are introduced, which are the Sex, Length, Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight, Shell_weight and Rings.

Among all these variables, 'Sex' is a categorical variable, which has three categories 'Male', 'Female' and 'Infant'. 'Rings' is a variable that represents the age of abalone, ranging from 1 to 29. Besides these two variables, the others are continuous variables which represent the physical appearance of abalone. Within

[1] Hossain, M. M., & Chowdhury, M. N. M. (2019). Econometric ways to estimate the age and price of abalone.

[2] Runze Guo, et al. (2021). A New Method of Measuring the Age of Abalone Based on Data Visualization Analysis. Journal of Physics: Conference Series, pp: 1744.

these continuous variables, there are four variables, which are ‘Whole_weight’, ‘Shucked_weight’, ‘Viscera_weight’ and ‘Shell_weight’, have linear correlations with each other. According to the description, the whole weight of abalone is the aggregate of shucked weight, viscera weight, shell weight and an unknown evaluation of the bleed or water inside the abalone’s body.

Variable Name	Role	Type	Demographic	Description	Units
Sex	Feature	Categorical		M, F, and I (infant)	
Length	Feature	Continuous		Longest shell measurement	mm
Diameter	Feature	Continuous		perpendicular to length	mm
Height	Feature	Continuous		with meat in shell	mm
Whole_weight	Feature	Continuous		whole abalone	grams
Shucked_weight	Feature	Continuous		weight of meat	grams
Viscera_weight	Feature	Continuous		gut weight (after bleeding)	grams
Shell_weight	Feature	Continuous		after being dried	grams
Rings	Target	Integer		+1.5 gives the age in years	

Figure 1 Data Description in dataset

c) Modelling approaches:

There are three models that are used in training and fitting the dataset, which are Classification and Regression Trees and a suitable classification model using random forests. Moreover, this report contains a comparison from accuracy and efficiency of different models, and choose a final model for the age prediction of abalone.

II. Exploratory Data Analysis

According to the description in the original dataset, all the missing values have been removed, and the ranges of every continuous variable have been scaled by dividing by 200. However, all types of variables and possible erroneous data still need to be checked and reported to the model beforehand. According to the figure 2 showing below, we can see that there are 4176 observations contained in the dataset and 9 variables in total. However, 'M' variables should be treated as category variables of type 'chr'. As a result, it needs to be converted into factor variable. Furthermore, ‘M’ is divided into three categories: 'M' for male, 'F' for female, and 'I' for infant. The remaining variables are numerical, with 'Rings' being an integer type.

```
'data.frame':   4176 obs. of  9 variables:
 $ M             : chr  "M" "F" "M" "I" ...
 $ Length        : num  0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 0.525 ...
 $ Diameter      : num  0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 0.38 ...
 $ Height        : num  0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 0.14 ...
 $ Whole_weight  : num  0.226 0.677 0.516 0.205 0.351 ...
 $ Shucked_weight: num  0.0995 0.2565 0.2155 0.0895 0.141 ...
 $ Viscera_weight: num  0.0485 0.1415 0.114 0.0395 0.0775 ...
 $ Shell_weight  : num  0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 0.21 ...
 $ Rings         : int  7 9 10 7 8 20 16 9 19 14 ...
```

Figure 2 Data typo checking

Next step is to check for missing values to see if the dataset is complete. Figure 3 below shows the absence of missing values in each variable. The number of missing values for each column is zero, indicating that the dataset is complete.

A data.frame: 1 × 9

M	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	0	0	0	0	0	0	0	0

Figure 3 Check for missing data

After demonstrating that the dataset is complete, there is no evidence to suggest that all the data has been accurately recorded. Therefore, I check the quantile for each variable to see if all the variables have normal range. The result is shown in figure 4.

Except for height, the figure shows normal ranges for the variables. The minimum height value is 0, which is unlikely for an abalone. To investigate, I selected the rows from the dataset that have zero height, as shown in Figure 5. The figure

```
       M                 Length          Diameter          Height      
 Length:4176        Min.   :0.075   Min.   :0.0550   Min.   :0.0000  
 Class :character   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150  
 Mode  :character   Median :0.545   Median :0.4250   Median :0.1400  
                    Mean   :0.524   Mean   :0.4079   Mean   :0.1395  
                    3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650  
                    Max.   :0.815   Max.   :0.6500   Max.   :1.1300  
  Whole_weight    Shucked_weight   Viscera_weight     Shell_weight   
 Min.   :0.0020   Min.   :0.0010   Min.   :0.00050   Min.   :0.0015  
 1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.09337   1st Qu.:0.1300  
 Median :0.7997   Median :0.3360   Median :0.17100   Median :0.2340  
 Mean   :0.8288   Mean   :0.3594   Mean   :0.18061   Mean   :0.2389  
 3rd Qu.:1.1533   3rd Qu.:0.5020   3rd Qu.:0.25300   3rd Qu.:0.3290  
 Max.   :2.8255   Max.   :1.4880   Max.   :0.76000   Max.   :1.0050  
     Rings       
 Min.   : 1.000  
 1st Qu.: 8.000  
 Median : 9.000  
 Mean   : 9.932  
 3rd Qu.:11.000  
 Max.   :29.000  
```

Figure 5 Quantile of each variable

A data.frame: 2 × 9

	M	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1257	I	0.430	0.34	0	0.428	0.2065	0.0860	0.1150	8
3996	I	0.315	0.23	0	0.134	0.0575	0.0285	0.3505	6

Figure 4 Unnatural observations

indicates two observations with zero height. With the exception of the variable 'Height', almost all of the variables in these two observations fall below the 1st quantile. Additionally, based on the 'Rings' of these two observations, the number of rings is below 8, indicating that these two abalone are in the 'Young' category. It is reasonable to conclude that younger abalone are smaller and lighter than older ones. Therefore, it is possible that the recorded observations for these two characteristics were incorrect. As a result, these observations are removed from the original dataset.

After removed these observations, we're required to check the correlations among those variations, as shown in figure 6. The plot is colored by the category of the sex of abalone. In the figure, the data points in blue color are belonged to 'Male', while those in green color are belonged to 'Infant' and those in red color are belonged to 'Female'. The numbers show at the top-left corner of the figure are the correlations among the variables. Among those numbers, there is a high level

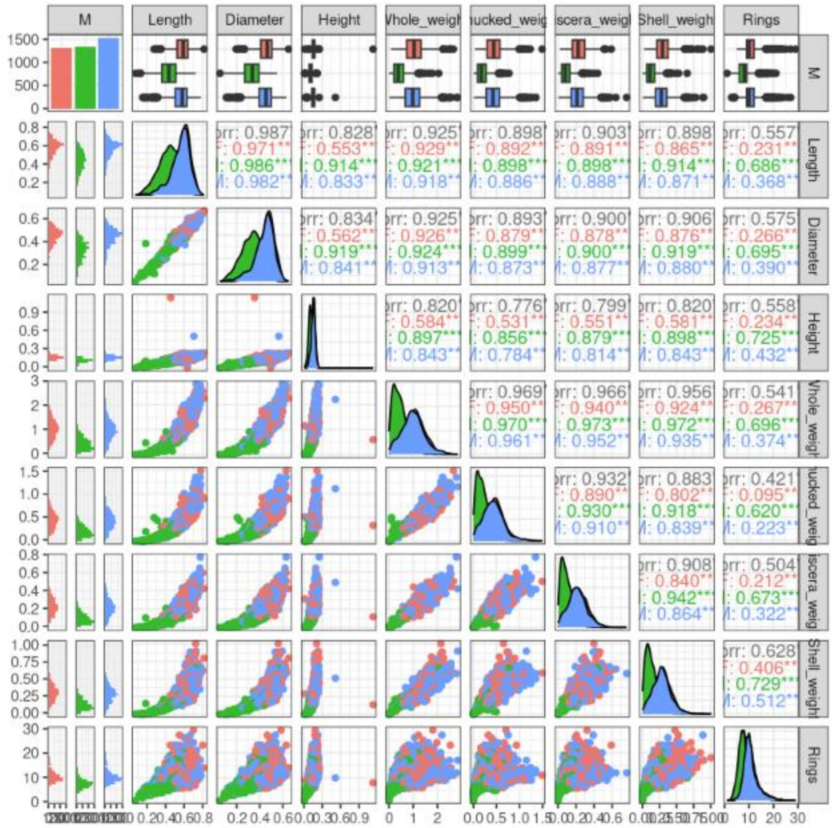


Figure 6 Correlations

of multicollinearity between the independent variables. Meanwhile, due to the possible linear correlations among ‘Whole_weight’ and the rest of the variables related to weight. Consequently, I have added a new variable 'Water' which is equal to 'Whole_weight' - 'Shucked_weight' - 'Viscera_weight' - 'Shell_weight' and only save ‘Whole_weight’ in the dataset. This variable represents the water and bleed contained in the abalone, so its value should be greater than 0. However, there are 153 observations that have a water value below 0, as shown in Figure 7. These observations are considered erroneously recorded data for no real reason that the water and bleed contained in the abalone is below 0. Furthermore, for a very small proportion of these observations account for the whole dataset, I consider removing these observations as well.

After having removed all the unnatural observations, I add a categorical variable ‘Age’, which has three categories: ‘Young’ for Rings number below 8, ‘Middle’ for Rings number range from 9 to 10, and ‘Old’ for the rest of the Rings numbers. Meanwhile, I assign numbers to the three categories in column 'M', 'I' to 0, 'F' to 1, 'M' to 2. After this work, I check the distribution of age and gender in the dataset to see if each category is evenly distributed. From Figure 8 we can see that all categories are almost evenly distributed.

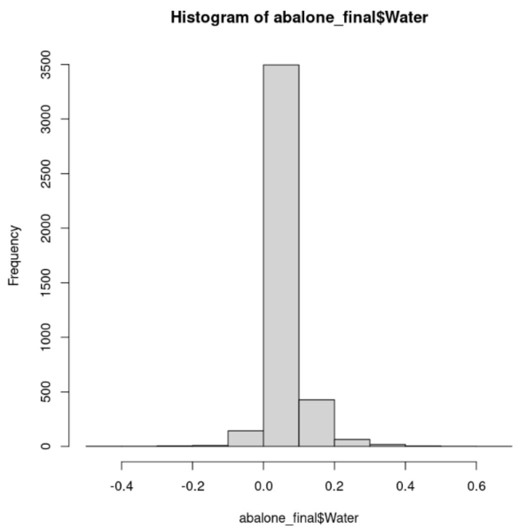


Figure 7 Distribution of Water

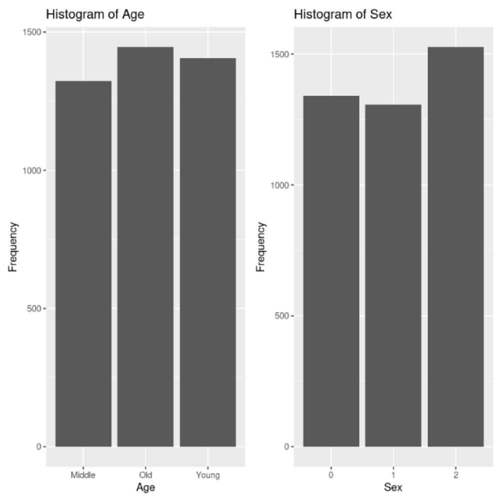


Figure 8 Distribution of Age and Sex

	M	Length	Diameter	Height	Whole_weight	Water	Age
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	2	0.350	0.265	0.090	0.2255	0.0075	Young
2	1	0.530	0.420	0.135	0.6770	0.0690	Middle
3	2	0.440	0.365	0.125	0.5160	0.0315	Middle
4	0	0.330	0.255	0.080	0.2050	0.0210	Young
5	0	0.425	0.300	0.095	0.3515	0.0130	Young
7	1	0.545	0.425	0.125	0.7680	0.0645	Old

A data.frame: 6 × 7

	M	Length	Diameter	Height	Whole_weight	Water	Age
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1004	1	0.605	0.490	0.150	1.1345	0.0485	Middle
623	1	0.500	0.385	0.105	0.4980	0.0390	Old
2694	1	0.650	0.535	0.175	1.2895	0.0595	Middle
934	0	0.450	0.355	0.105	0.4445	0.0210	Young
2949	2	0.635	0.515	0.160	1.2075	0.0420	Old
2147	2	0.460	0.375	0.135	0.4935	0.0530	Old

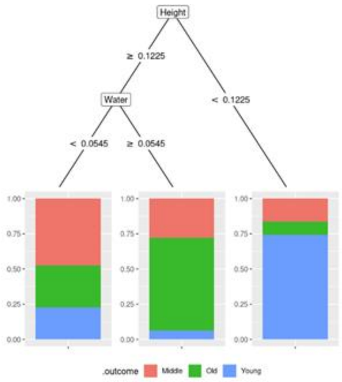
Figure 9 The training set and testing set

After that, I randomly split the datasets into training and testing sets according to a ratio of 75:25. The dataset described is shown in figure 9. Above is all the processing of the original dataset.

III. Modelling

1. Classification and Regression Trees

CART algorithm is based on a frame of decision trees, which uses a series of binary separation to build numerous decision trees[3]. It can chop up parameter hyperspace based on the score function, such as Gini Index, and predict a qualitative response[4]. There's no need to create dummy variables before using CART and it can be easily interpreted. However, there are some requirements for the dataset. For very low-robust ability of CART, the dataset has to be cleaned and easy to be separated. For abalone dataset, according to figure 6,



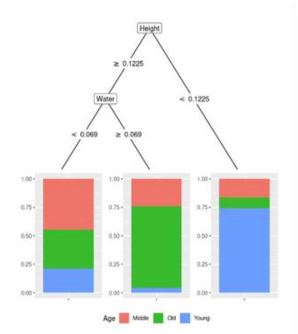

```
      cp
2 0.071
n= 3131

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 3131 2012 Old (0.30916640 0.35739380 0.33343980)
  2) Height>=0.1225 2162 1133 Old (0.37419056 0.47594820 0.14986124)
    4) Water< 0.069 1394  769 Middle (0.44835007 0.34289813 0.20875179) *
    5) Water>=0.069 768  217 Old (0.23958333 0.71744792 0.04296875) *
  3) Height< 0.1225 969  249 Young (0.16408669 0.09287926 0.74303406) *
```

Figure 10 CART predict result

we can see that the 'young' category of abalone is more easily segmented,




```
check_predict_rpart Middle Old Young
             Middle    224 153    91
             Old        69 145     8
             Young      62  29   262

[1] "Accuracy: 0.604985618408437"
[1] "Precision:"
   Middle       Old     Young
0.4786325 0.6531532 0.7422096
[1] "Recall:"
   Middle       Old     Young
0.6309859 0.4434251 0.7257618
[1] "F1_score:"
   Middle       Old     Young
0.5443499 0.5282332 0.7338936
```

```
check_fit_rpart Middle Old Young
         Middle    625 478   291
         Old       184 551    33
         Young     159  90   720

[1] "Accuracy: 0.605557329926541"
[1] "Precision:"
   Middle       Old     Young
0.4483501 0.7174479 0.7430341
[1] "Recall:"
   Middle       Old     Young
0.6456612 0.4924039 0.6896552
[1] "F1_score:"
   Middle       Old     Young
0.5292125 0.5839958 0.7153502
```

Figure 11 CART after pruning

[3] Breiman L, Friedman JH, Olshen RA, Stone CJ.(1984). Classification and regression trees (Wadsworth statistics/probability).

[4] James, Gareth. An introduction to statistical learning: with applications in R

whereas the other two categories are less easily segmented, and that there is a large difference in the distribution of abalone data in the 'I' category relative to the other two sexes. For reducing the uncertainties when building the model and also in order to improve the generalizability of the model, I include a 10-fold cross-test in the training phase. The figure 10 shows the result of the CART with 10-fold cross-validation.

The 10-fold cross-validation shows the best training model with 'cp', which corresponds to the minimum value of the cross-validation error. I then trim this model with cp equal to 0.071, as shown in Figure 10. The result is shown on the left in Figure 11. Accoding to the image shows on the right in Figure 11, we can see that this model fit almost 60.5% of the training data. Then this model is used to test the test set, and gets 60.49% accuracy as shows in the middle of Figure 11. As described above, 'young' abalone are easier to separate, resulting in higher accuracy in predicting 'young' abalone. However, the accuracy of the other two categories is lower, resulting in a lower overall accuracy.

2. Random Forest

Random forests are a collection of numerous decision trees, which composed of various simple models using binary splits on the predict variables[5]. Under this dataset, we are doing a classification with random forests. For there are 6 predictors in total we need consider, the number of predictors at each split of

```
Call:
 randomForest(x = x, y = y, ntree = 2000, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 2000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 36.63%
Confusion matrix:
       Middle Old Young class.error
Middle    444 347   177   0.5413223
Old       266 777    76   0.3056300
Young     183  98   763   0.2691571
```

```
check_predict_rf Middle Old Young
          Middle    156  86    55
          Old       133 208    25
          Young      66  33   281
[1] "Accuracy: 0.618408437200384"
[1] "Precision:"
   Middle       Old     Young
0.5252525 0.5683060 0.7394737
[1] "Recall:"
   Middle       Old     Young
0.4394366 0.6360856 0.7783934
[1] "F1_score:"
   Middle       Old     Young
0.4785276 0.6002886 0.7584345
```

Figure 12 random forest

[5] Breiman, L. (2017). Classification and regression trees. Routledge.

the trees is set to 2 according to $m = \sqrt{p}$, in which p is the number of total predictors and m is the predictors that need to be considered at each split[6]. Moreover, the number of trees is set to 2000 for stability of the error state. The result is shown in figure 12.

Figure 10 shows that there is a significant improvement in overall accuracy compared to CART. Random forests also have a good prediction for 'young' abalone, but lower accuracy for 'old' and 'middle aged' abalone. Furthermore, according to Figure 13, which is a plot showing the importance and contribution that all the predictors have in the model, we can see that 'Water' and 'Whole_weight' contribute a lot to the model, while 'M', which represents 'Sex', has very little contribution.

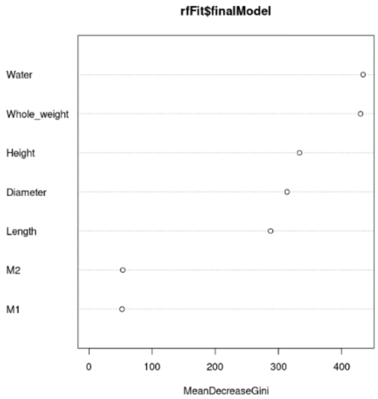


Figure 13 VarImpPlot of the model

However, based on the accuracy of the two models in classifying abalone age, prediction was better in the 'young' category, while the 'old' and 'middle' categories did not perform well in prediction. The 'Old' and 'Middle' categories did not perform well and the models need to be improved. The reason why the decision tree based categorical regression model was not accurate in these two categories is that Figure 14 shows that the distributions of these two categories, which ‘Old’ is colored in green and ‘Middle’ is colored in red, overlap almost identically on almost all variables, so the model may not be

[6] James, Gareth. An introduction to statistical learning: with applications in R.

able to discriminate between these two categories in detail.

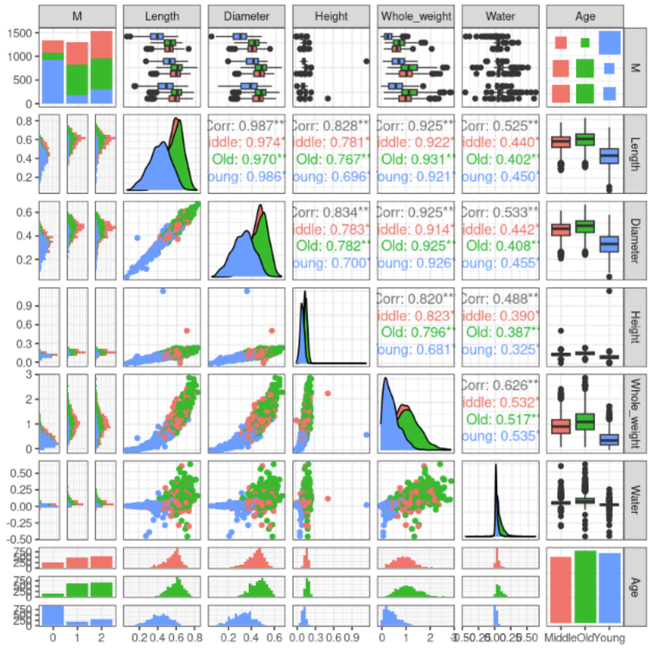


Figure 14 Dataset colored by 'Age'

Since the model does not perform well on the 'Old' and 'Middle' categories, I train a new random forest model based on these two categories of data only, by removing the 'Young' category from the test and training set as a new test and training set to be used in building and testing the model. The results of the training test of the reconstructed model are shown in Figure 15, and compared to Figure 12, we can see that there is a significant improvement in the accuracy of predicting the two classes.

```
Call:
 randomForest(formula = as.factor(Age) ~ ., data = abalone_train2,      mtry = 2, ntree = 3000, max_features = 50, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 3000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 32.73%
Confusion matrix:
       Middle Old class.error
Middle    622 346    0.3574380
Old       337 782    0.3011618


check_predict0 Middle Old
        Middle    223 114
        Old       132 213

[1] "Accuracy: 0.639296187683284"
[1] "Precision:"
   Middle       Old
0.6617211 0.6173913
[1] "Recall:"
   Middle       Old
0.6281690 0.6513761
[1] "F1_score:"
   Middle       Old
0.6445087 0.6339286
```

Figure 15 Improvement on randon forests

I then replaced the predictions of the first random forest model for the 'Old' and

'Middle' categories with the predictions of the new random forest model. The result is shown in figure 16. As show in the result, we can see that the total accuracy has raised to 68.55%

```
final_predictions2 Middle Old Young
            Middle    218 111    55
            Old       137 216    25
            Young       0   0   281

[1] "Accuracy: 0.685522531160115"
[1] "Precision:"
   Middle       Old     Young
0.5677083 0.5714286 1.0000000
[1] "Recall:"
   Middle       Old     Young
0.6140845 0.6605505 0.7783934
[1] "F1_score:"
   Middle       Old     Young
0.5899865 0.6127660 0.8753894
```

Figure 16 Final model's result

IV. Model Comparison

In figure 17, all results get from the three models above are put in the table. Moreover, all the models have used 10-fold cross-validation to reduce the uncertainty and raise the robust ability.

From the table shows in figure 18, there are two criterions being used for evaluation of all the models, which are the accuracy and running-time.

	M	Length	Diameter	Height	Whole_weight	Water	Age	cart	rf	imp	cart_right	rf_right	imp_right
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<fct>	<fct>	<fct>	<lgl>	<lgl>	<lgl>
1004	1	0.605	0.490	0.150	1.1345	0.0485	Middle	Middle	Middle	Middle	TRUE	TRUE	TRUE
623	1	0.500	0.385	0.105	0.4980	0.0390	Old	Young	Old	Middle	FALSE	TRUE	FALSE
2694	1	0.650	0.535	0.175	1.2895	0.0595	Middle	Middle	Old	Old	TRUE	FALSE	FALSE
934	0	0.450	0.355	0.105	0.4445	0.0210	Young	Young	Young	Young	TRUE	TRUE	TRUE
2949	2	0.635	0.515	0.160	1.2075	0.0420	Old	Middle	Middle	Middle	FALSE	FALSE	FALSE
2147	2	0.460	0.375	0.135	0.4935	0.0530	Old	Middle	Middle	Old	FALSE	FALSE	TRUE

Figure 17 Final Result

	Method	Accuracy	Running.time
1	CART	0.6049856	0.09215379 secs
2	Random Forest	0.6193672	0.38511801 secs
3	Improve	0.6855225	1.28300000 secs

Figure 18 Evaluation

In summary, the third model, which uses two random forest models to predict the age of the abalone, has the highest overall accuracy of 68.55%, but takes more time to run, about 1.283 seconds. The Pure Random Forests model ranked 2nd, with an overall accuracy of 61.94%, and took much less time to run than the third model. The CART model is ranked 3, which has the lowest accuracy but takes very little time to run. Although the third model takes 4 times longer to run than random forests, it has huge improvements in accuracy. Consequently, compared to the other two models, the third model is much more practical.

V. Results and conclusion

Although I used model integration methods in the third model and this resulted in a significant improvement in accuracy, the 68% accuracy is still relatively limited and has a lot of room for improvement. In addition, the model has used 10-fold cross-validation and obtained a good result, showing that it has a good robustness capability.

However, this model can only predict the abalone’s age approximately within three categories. However, in reality, fishermen do not need to know the exact age of the abalone, they only need to determine the approximate age, differentiate whether the abalone is young, mature or old, and judge whether it is worth catching. This model has a high accuracy in predicting whether an abalone is young or not, so it is already applicable to some extent.

Furthermore, the lack of variables and the high level of misreported data are also reasons that affect the accuracy of the model to some extent. In addition, the growth cycle of abalone varies in different geographical locations around the world, temperature also affects the growth of abalone, and different species also have different sizes and weights, so this dataset has major limitations when it comes to training the model. To improve the accuracy of the model, more factors that can affect the age of abalone need to be considered.

Consequently, I recommend few ways to improve the model:

a) Put more factors into the dataset, such as mean temperature of the growing sea area, geographical location.
b) Finding methods and variables that make it easier to distinguish between middle-aged and older abalone.
c) Using a more advanced method, such as XGBoost tree model and artificial neural network.

VI. Bibliography

Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository.

Hossain, M. M., & Chowdhury, M. N. M. (2019). Econometric ways to estimate the age and price of abalone.

Runze Guo, et al. (2021). A New Method of Measuring the Age of Abalone Based on Data Visualization Analysis. Journal of Physics: Conference Series, pp: 1744.

Breiman L, Friedman JH, Olshen RA, Stone CJ.(1984). Classification and regression trees (Wadsworth statistics/probability).

Breiman, L. (2017). Classification and regression trees. Routledge.

James, Gareth. An introduction to statistical learning: with applications in R.

Appendix

M	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7
I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.120	8
F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.330	20
F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.260	16
M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.320	19
F	0.525	0.380	0.140	0.6065	0.1940	0.1475	0.210	14

Table 1 Original Dataset

	M	Length	Diameter	Height	Whole_weight	Water	Age
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	2	0.350	0.265	0.090	0.2255	0.0075	Young
2	1	0.530	0.420	0.135	0.6770	0.0690	Middle
3	2	0.440	0.365	0.125	0.5160	0.0315	Middle
4	0	0.330	0.255	0.080	0.2050	0.0210	Young
5	0	0.425	0.300	0.095	0.3515	0.0130	Young
6	1	0.530	0.415	0.150	0.7775	0.0690	Old
7	1	0.545	0.425	0.125	0.7680	0.0645	Old
8	2	0.475	0.370	0.125	0.5095	0.0155	Middle
9	1	0.550	0.440	0.150	0.8945	0.1090	Old
10	1	0.525	0.380	0.140	0.6065	0.0550	Old

Table 2 Processed Dataset

	M	Length	Diameter	Height	Whole_weight	Water	Age	cart	rf	imp	cart_right	rf_right	imp_right
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<fct>	<fct>	<fct>	<lgl>	<lgl>	<lgl>
1004	1	0.605	0.490	0.150	1.1345	0.0485	Middle	Middle	Middle	Middle	TRUE	TRUE	TRUE
623	1	0.500	0.385	0.105	0.4980	0.0390	Old	Young	Old	Middle	FALSE	TRUE	FALSE
2694	1	0.650	0.535	0.175	1.2895	0.0595	Middle	Middle	Old	Old	TRUE	FALSE	FALSE
934	0	0.450	0.355	0.105	0.4445	0.0210	Young	Young	Young	Young	TRUE	TRUE	TRUE
2949	2	0.635	0.515	0.160	1.2075	0.0420	Old	Middle	Middle	Middle	FALSE	FALSE	FALSE
2147	2	0.460	0.375	0.135	0.4935	0.0530	Old	Middle	Middle	Middle	FALSE	FALSE	FALSE
3176	1	0.490	0.380	0.130	0.5390	0.0095	Old	Middle	Old	Middle	FALSE	TRUE	FALSE
2775	0	0.585	0.460	0.145	0.8465	0.0455	Middle	Middle	Middle	Middle	TRUE	TRUE	TRUE
2375	2	0.340	0.275	0.090	0.2065	0.0210	Middle	Young	Young	Middle	FALSE	FALSE	TRUE
1103	1	0.505	0.390	0.125	0.5445	0.0080	Young	Middle	Middle	Middle	FALSE	FALSE	FALSE
4048	1	0.625	0.495	0.160	1.2340	0.0585	Old	Middle	Middle	Middle	FALSE	FALSE	FALSE
3455	2	0.600	0.475	0.155	1.1385	0.0970	Middle	Old	Old	Old	FALSE	FALSE	FALSE
2233	2	0.595	0.475	0.170	1.0965	0.0985	Old	Old	Old	Old	TRUE	TRUE	TRUE
2623	1	0.705	0.560	0.205	2.3810	0.2650	Middle	Old	Old	Old	FALSE	FALSE	FALSE
3973	0	0.415	0.330	0.100	0.3905	0.0200	Young	Young	Young	Young	TRUE	TRUE	TRUE

Table 3 Final Result