Week 7 Assignment – Performance Measures

Zach Adair

Regis University

Abstract

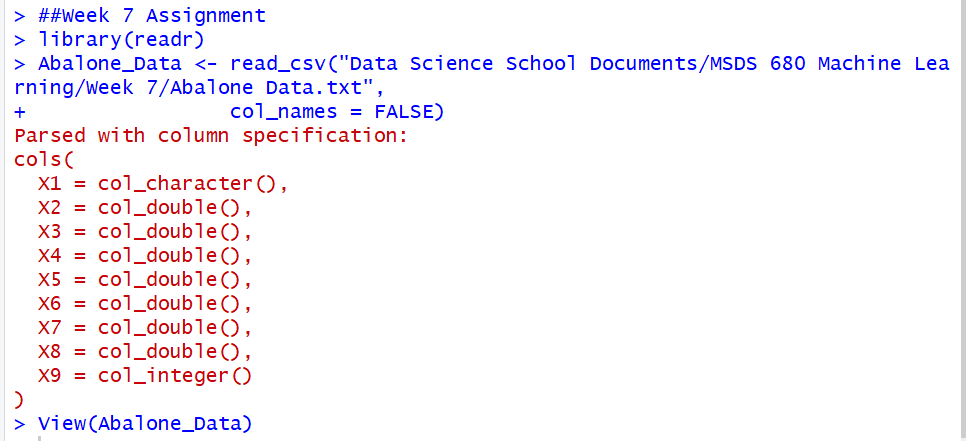
For week 7 our assignment is to take two classification techniques from prior weeks and use them on the data for the assignment. The data used in this assignment is the abalone data set, which can be tracked by their age based on the number of rings they have in their shell. This data doesn’t have an age field so it will need to be created before trying to predict what is the age of the abalone. The classification techniques I will be using are the KNN method and then the Naïve Bayes method, I believe the use of these methods will help construct a good predictor for the age of abalones.

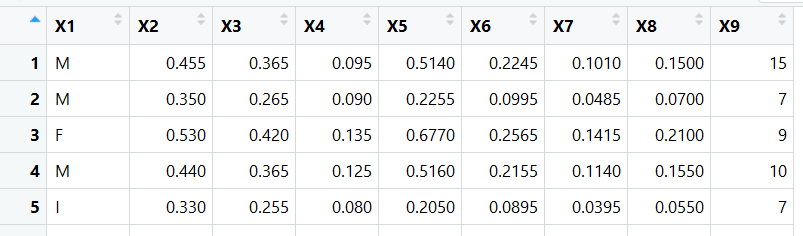
Week 7 Assignment – Performance Measures

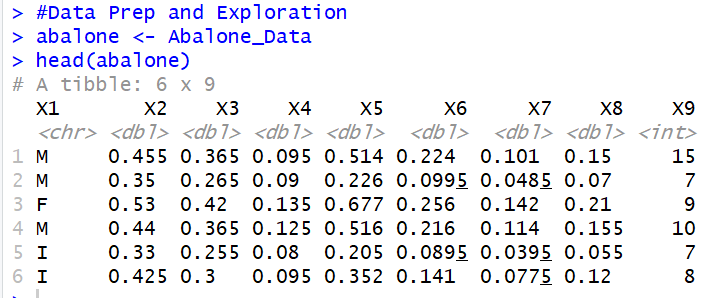
I will start by doing some data preparation, followed by some data exploration of the abalone data. After, I will be ready to start my analysis so I will start with the KNN method, followed by the Naïve Bayes method. Then I will compare and summarize the results of the two algorithms.

# Data Preparation

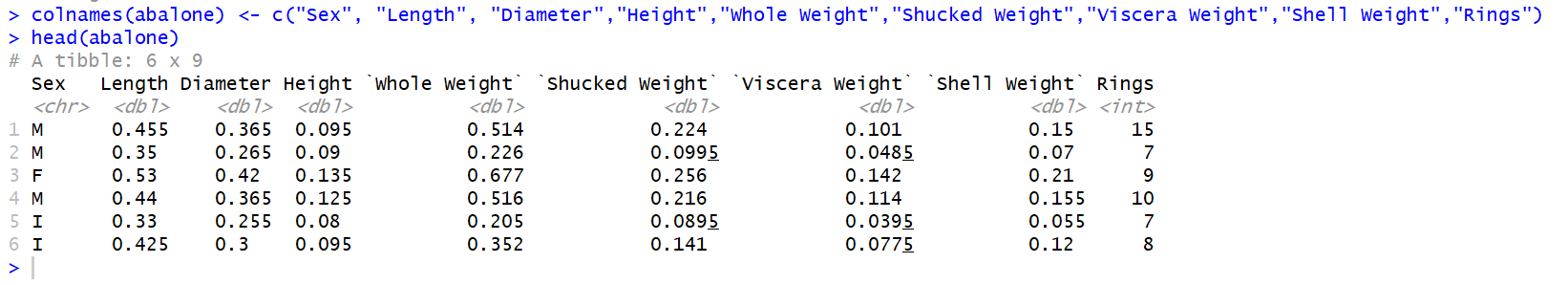
All good data preparation starts with first uploading your data into RStudio.



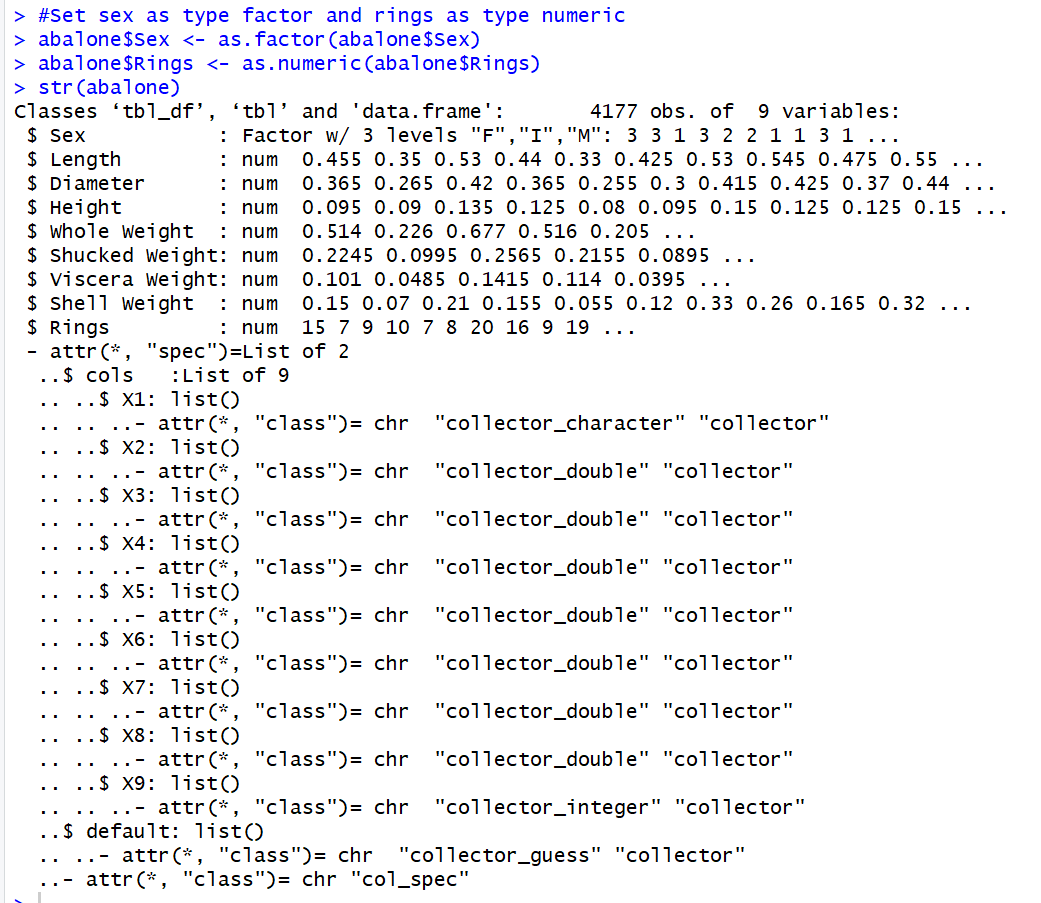




The data has come in pretty clean, but unfortunately there aren’t any column names, just the generic ones that RStudio gives, so I will need to make those for my dataset.



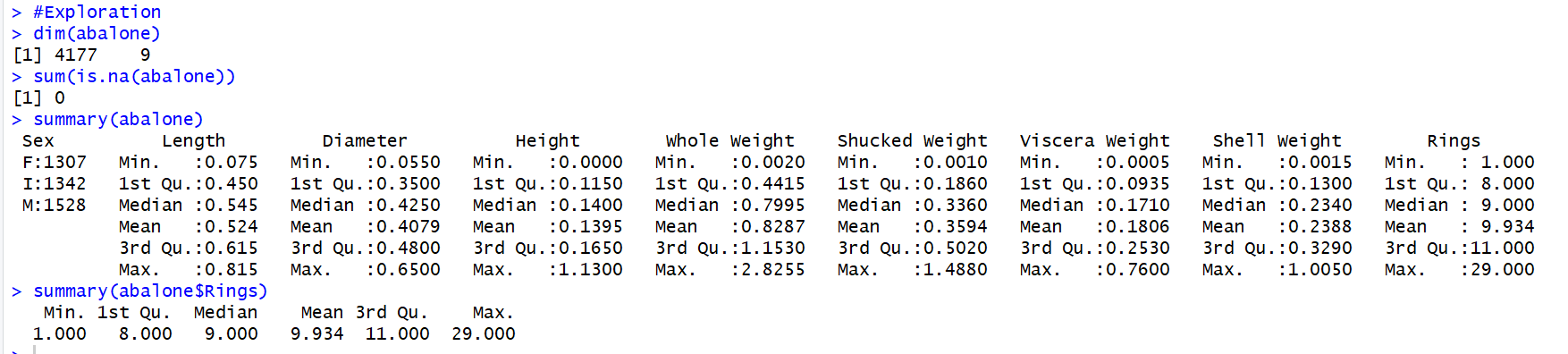
Now the data has column names. Now I need to make a couple changes to two of the columns, the sex column needs to go from character to factor, and the rings column needs to go from integer to numeric.



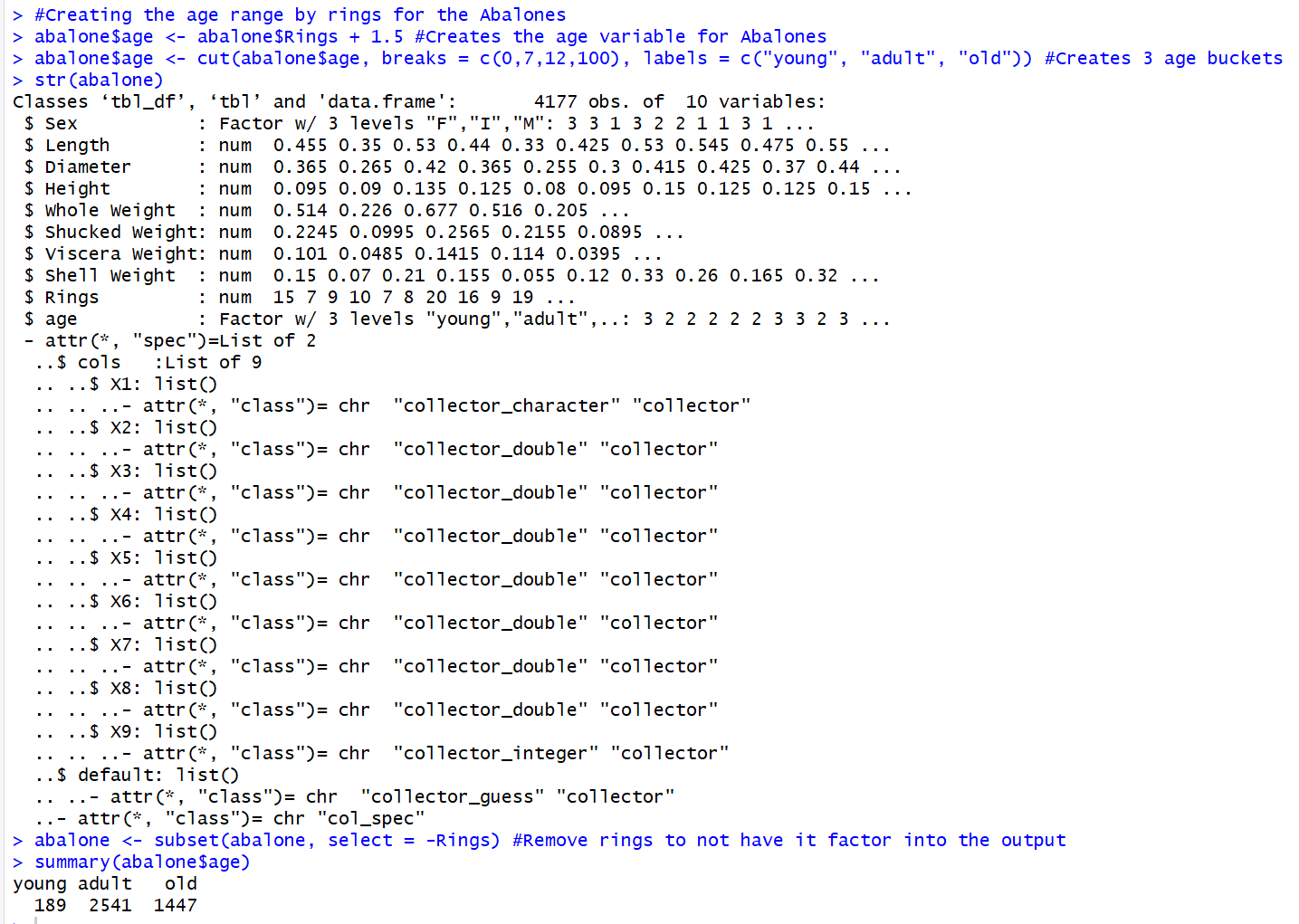
The data has the columns situated properly, now it is time to do some exploration of the data and set up how to measure the age of the abalones.

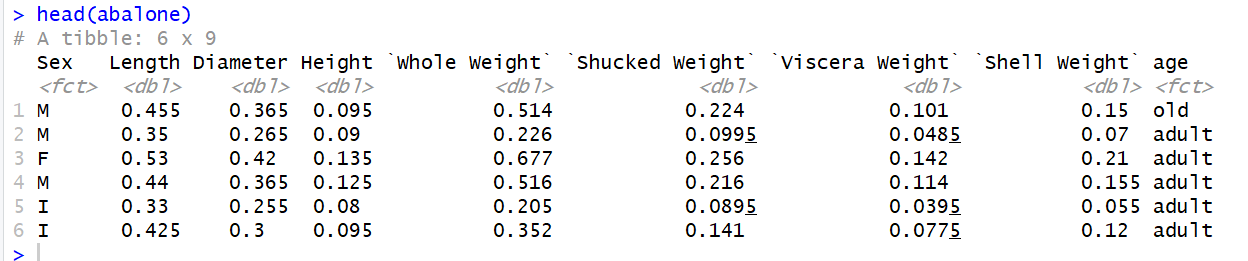
## Data Exploration

The data exploration part will also have a little bit more data preparation along with it, right now the data set doesn’t have a field which properly measures for age, and since age is primarily judged by the rings alternate the rings field to the age field instead. First I want to do some data exploration of the data set in its real form.



From our quick data exploration we know a few things about some fields that we need to make note of. One, we see that there are 9 fields within our data set. In particular, the field of sex, which was changed to a factor has three levels, F, M and I. Also we see our numeric field of most importance, rings, has a minimum value of 1 and a max value of 29 so that makes for a wide range of levels that an abalone can have coming in, so the best thing to do for this analysis will be to create an age range which brackets these ages into 3 categories: young, adult and old.

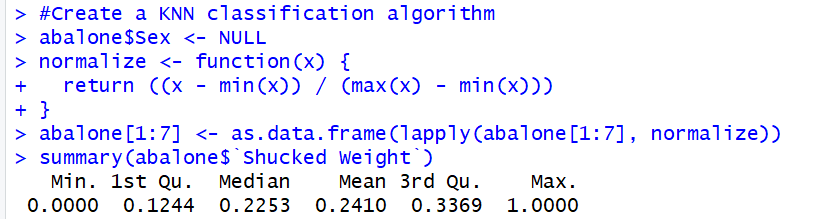




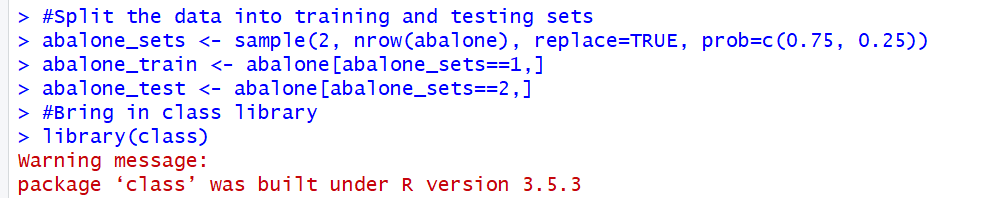
The age brackets have now been constructed, and I have gotten rid of the rings field so it wouldn’t dilute the analysis later on. I also took a look at the distribution of the abalones within the age brackets, Overwhelmingly they are adults, but there is a lot of old aged abalones and very few young abalones based on the distribution. With the necessary changes made, it is now time to start constructing the classifiers for this analysis.

### **KNN Classification.**

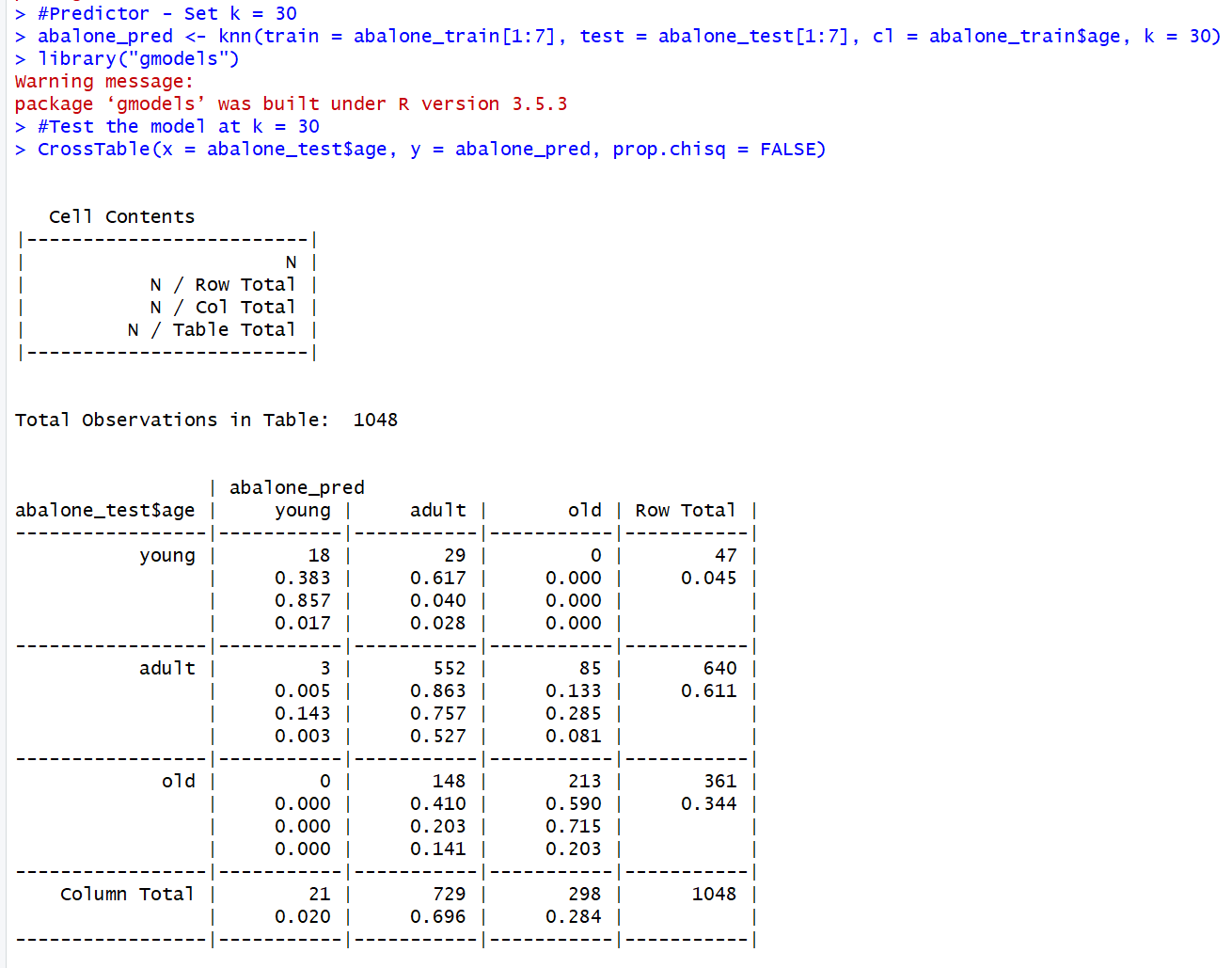
The first classification technique will be using KNN, to start I will be normalizing the data and preparing to use it for this specific algorithm.

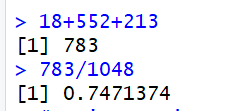


The normalization process is done and now all numeric fields have values that range from 0 to 1. Next I’ll need to set up training and test data sets. I will split them 75/25 and that will be how the data will train the model and then test it when it is trained.

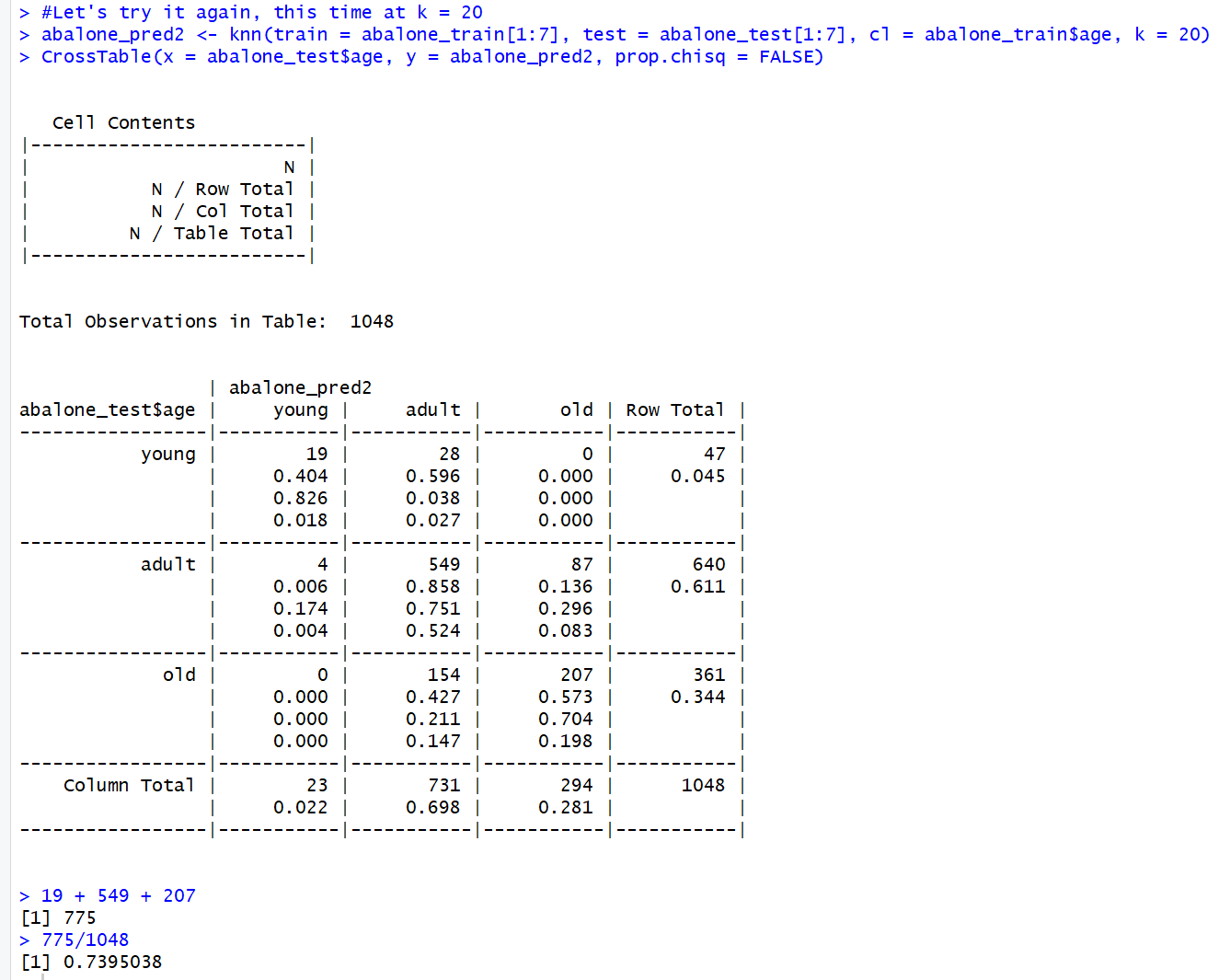


The two data sets are set up and now it is time to set up the predictor and see how the model test. I will start with it at k = 30 and see how the algorithm performs at that level.

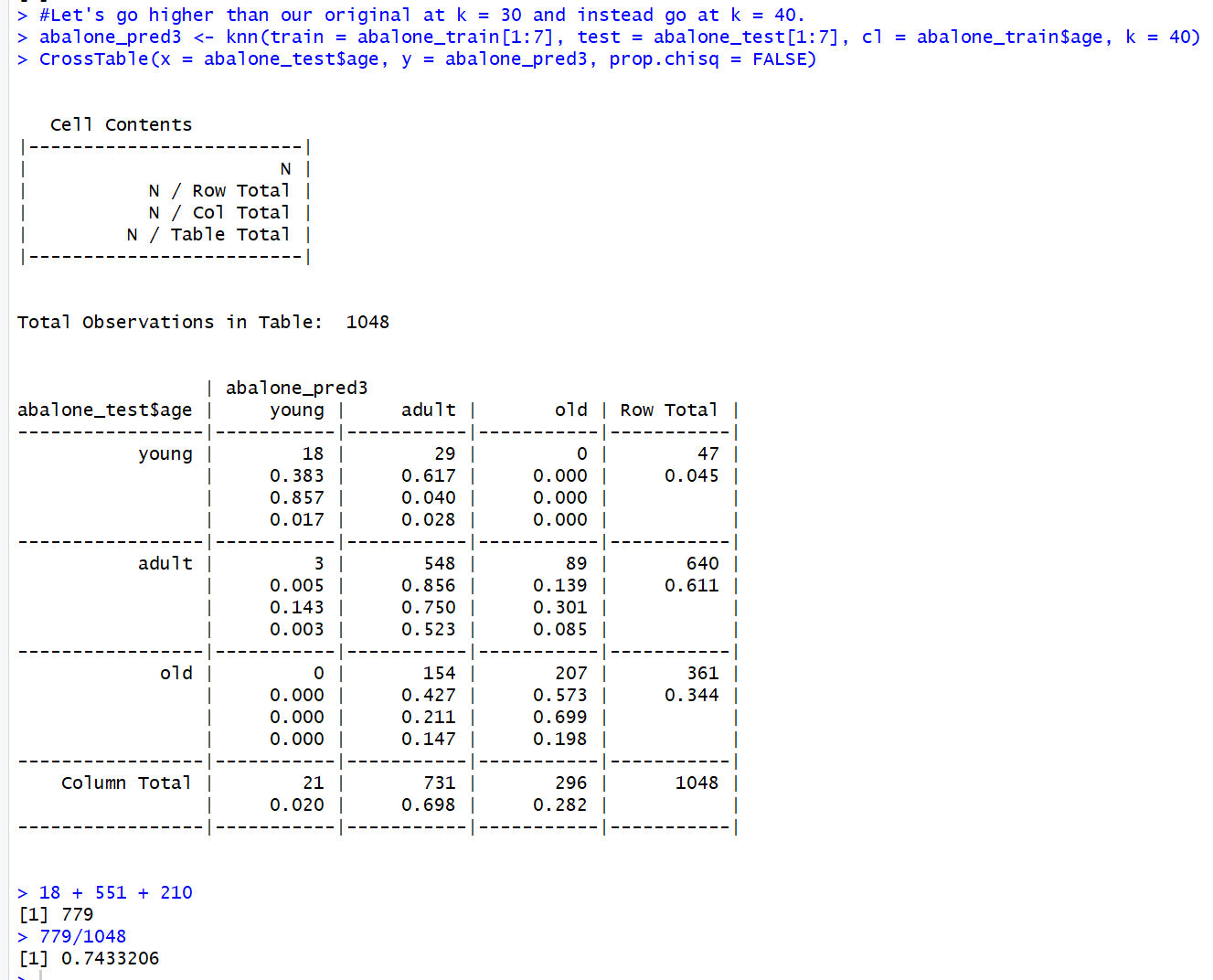




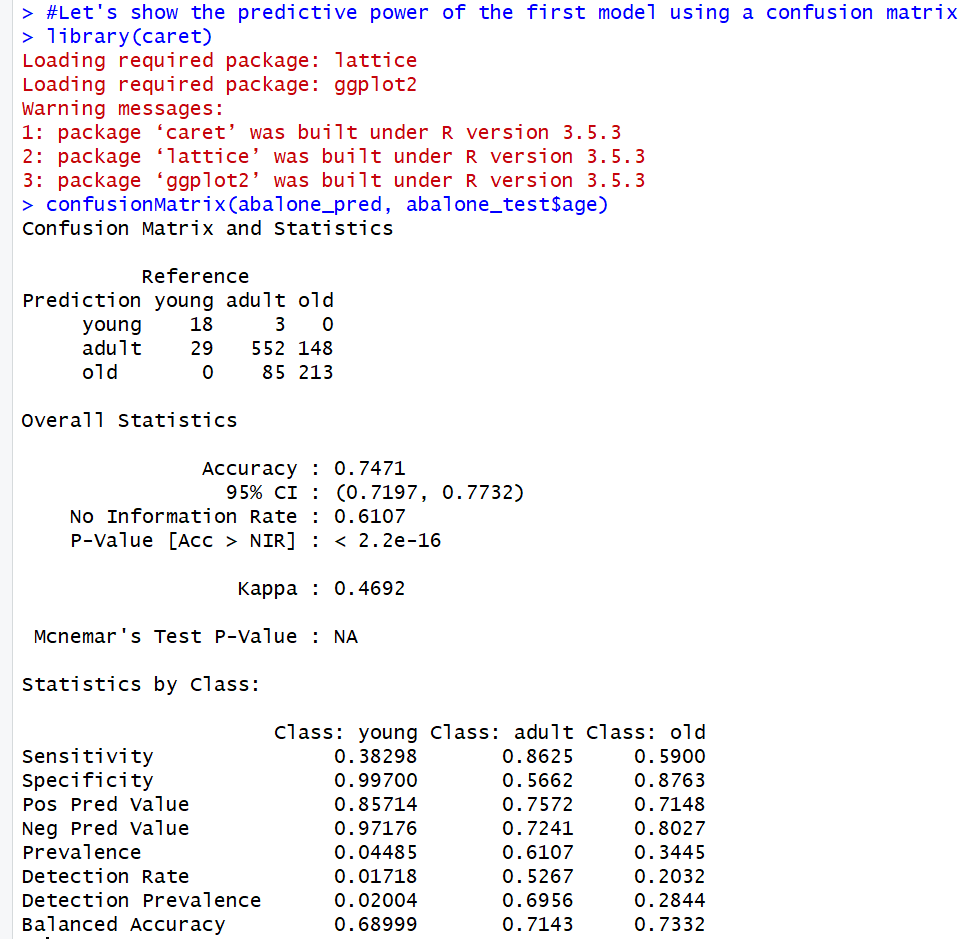
The model results were just under 75% which isn’t ideal for accurately measuring the age of the abalones, so let’s change the value of k to 20 and see how it performs.



The results unfortunately did worse when the k value was lowered to 20, so I will now try to run it when k = 40 and see if that improves the models accuracy.



Once again the model didn’t perform to the level I was hoping for, and though it did better than the second KNN model, it ended up doing worse than the original KNN model. I’m going to put the original KNN model into a confusion matrix and see what kind of information I can pull from it.

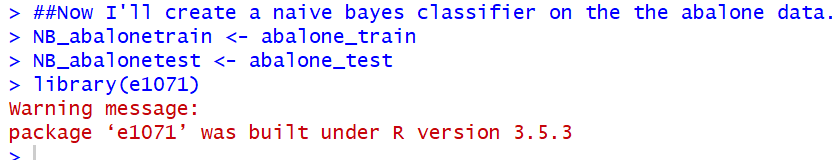


From the confusion matrix, we confirm that the model accuracy was indeed just under 75% and that we can validate the results of this model based on how low the p-value is, which isn’t promising because ideally we would want a p-value that is 95% and above.

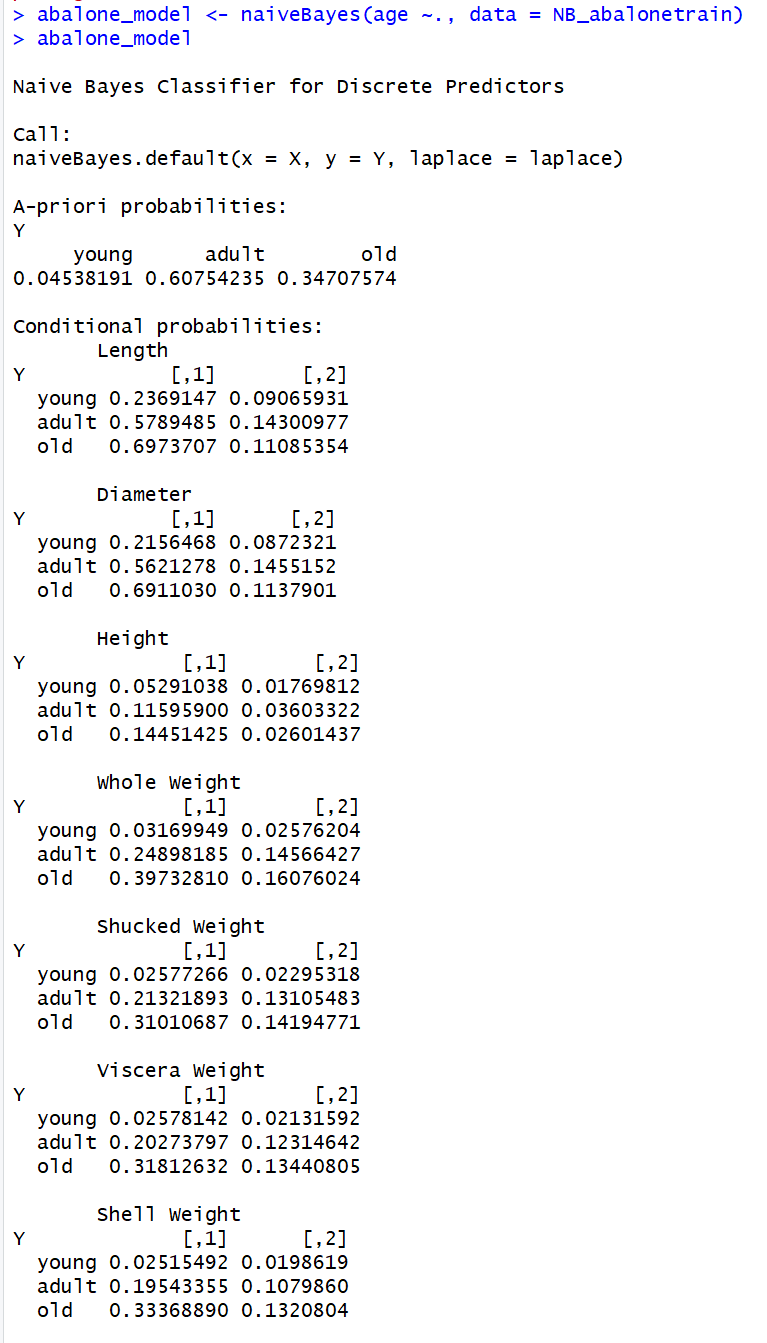
So I KNN might not be the best way to classify the abalone data. So next I will try and use the Naïve Bayes technique and find out if that yields better results.

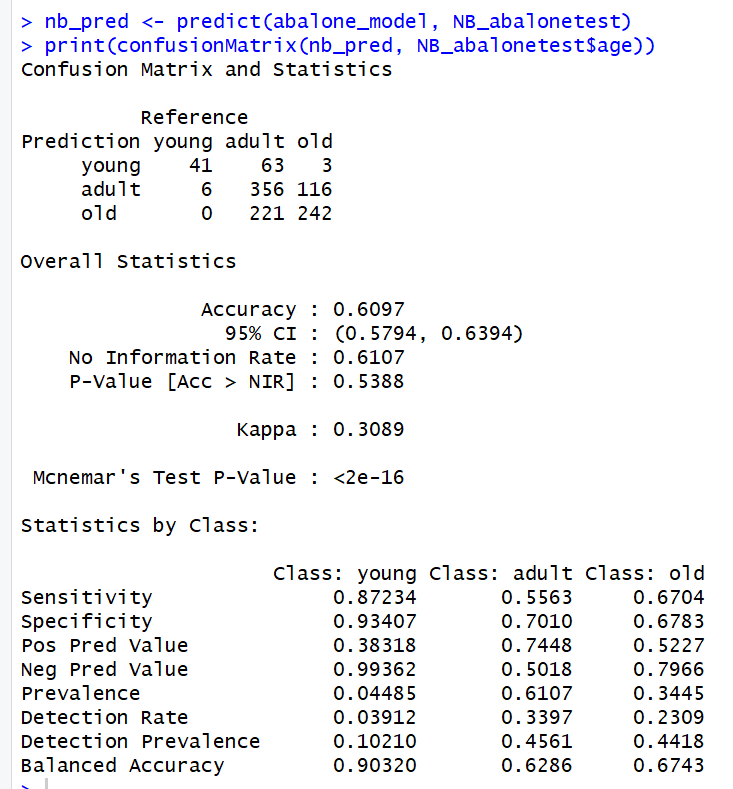
#### ***Naïve Bayes***

##### For the Naïve Bayes technique the hope is to get a better result than what I got from using the KNN method, but we will see. To start I will need to set up my training and test sets again as well as pull in the library e1071.

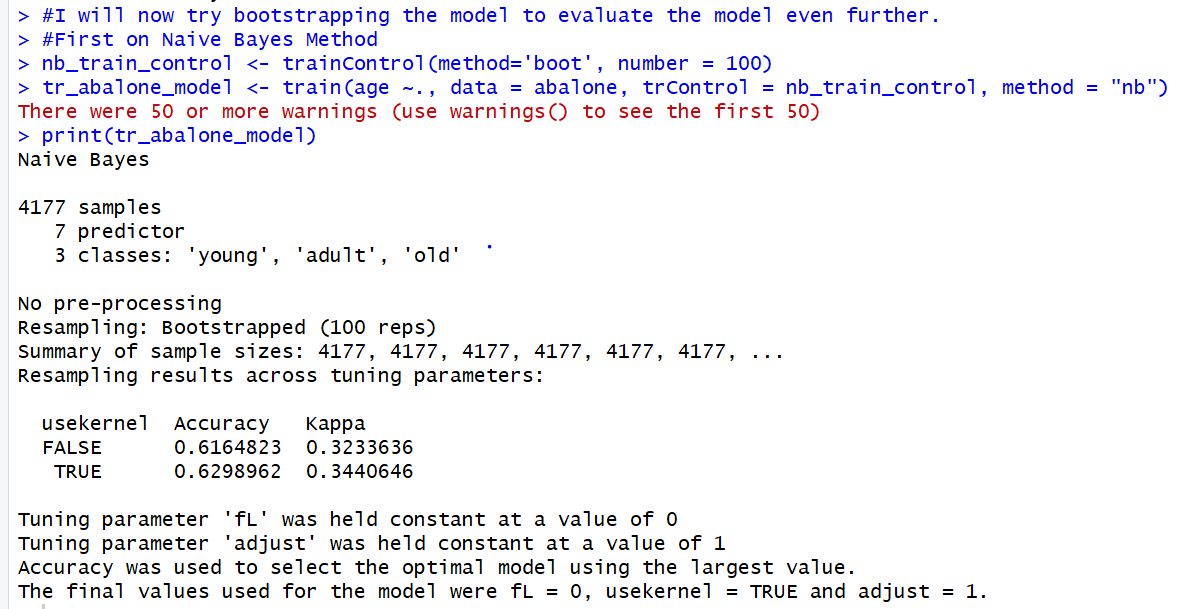


Now it is time to set up the NB model and try and accurately predict the age of the abalones.

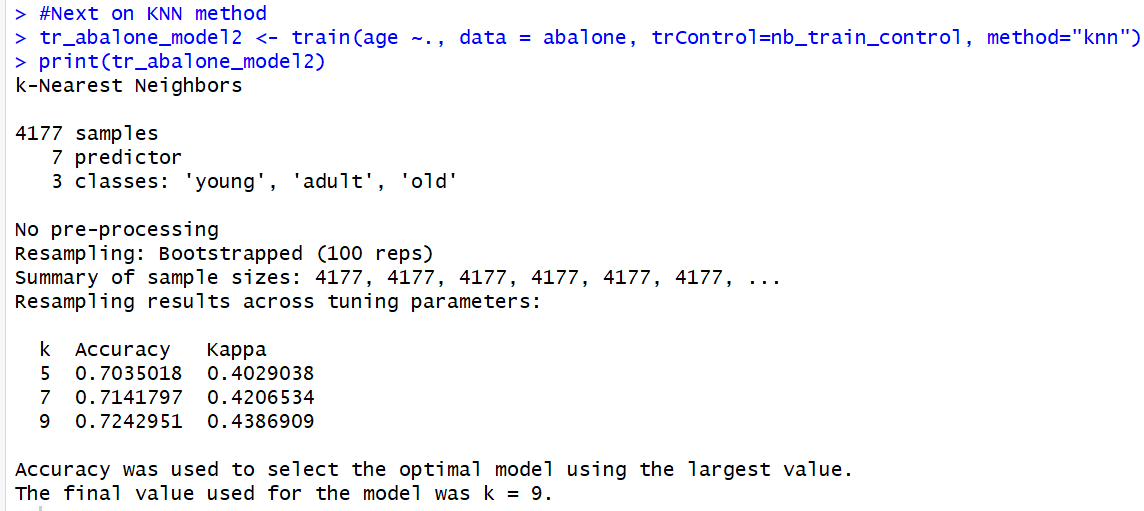




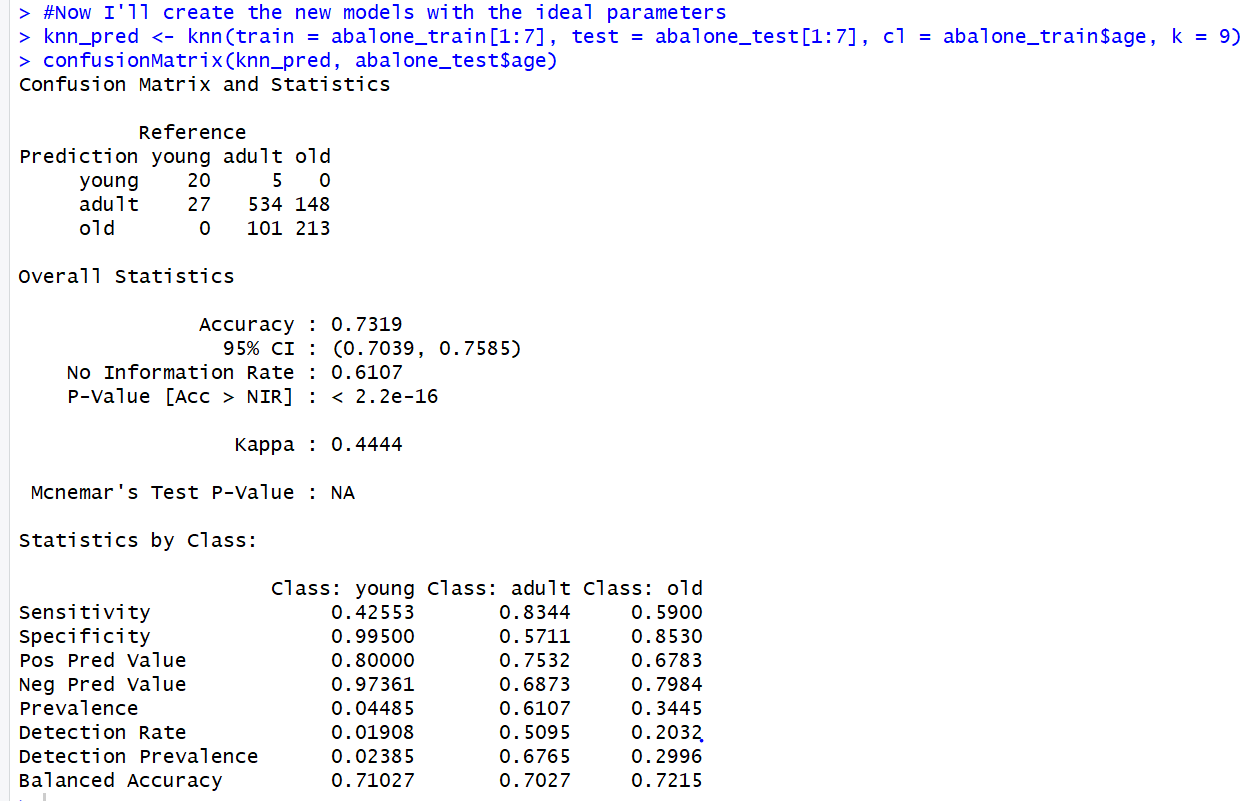
Unfortunately the NB model ended up performing much worse than the KNN model. The accuracy was just under 61% which is very unreliable when predicting the ages of abalones. Thus the KNN method is more accurate than the NB method. Before I move on though, I will bootstrap the models to evaluate their performance even further.



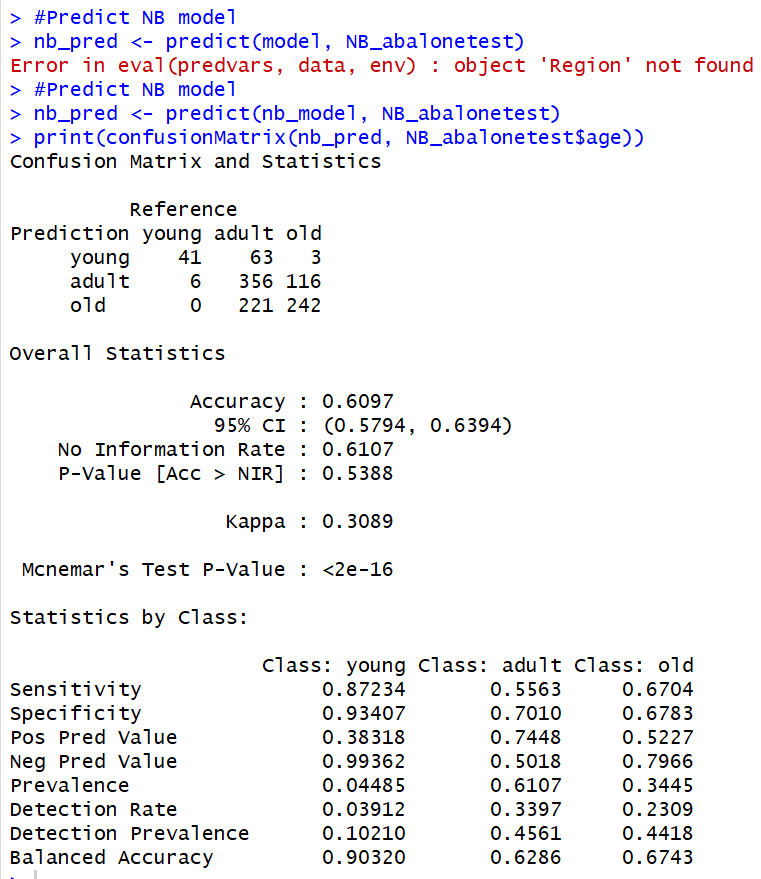
Next, the KNN method.



From bootstrapping both methods it was evident that KNN was the better of the two methods. Another thing we learned is that k = 9 would have been the ideal k value to use for that algorithm. The bootstrapping method also shows that at fL = 0 and usekernal = TRUE will yield better results from the Naïve Bayes model, so I will make a new model for each to see if these are true. Then the last step in the process before I end my analysis is doing a 10-fold cross validation for training on the classifiers.

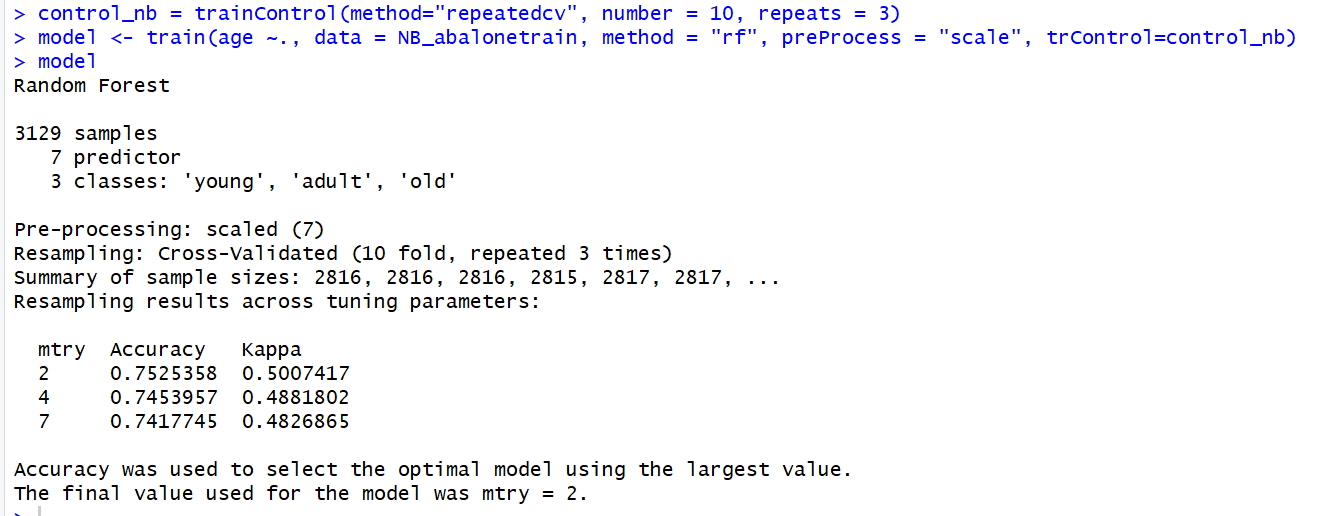


For the KNN model now, the accuracy is at 73% which is actually worse than when K = 30 so I’m not sure why the bootstrapping method told me it was ideal to have K=9 but I will move on with the NB method.



The new NB model with the bootstrap suggestions actually ended up giving me the same result that I had from the first model so there was no improvement the accuracy was still slightly below 61%. From the 10-fold cross validation there was no improvement in what the models outputs were.

I want to see if the model might now perform better if I chose a different classifier technique. The technique I’m going to test is the random forest.



The result of just testing the accuracy of the random forest yielded decent results of just above 75% but not anything that is notably better than the KNN method which was the better of my two performing methods.

##### *Summary.*

Unfortunately the performance of all my two methods as well as checking the RF method outputted results which were not as high as I was hoping for. The highest accuracies came in around 75% and ideally we would want to see that around 20% more to reliably push the model along and have it try to take in new abalones caught and used by restaurants. The way to improve on this issue is to possibly keep trying different classification methods to hopefully find one which performs better, but also another solution could be to get a larger data set so the models have more data to train on and hopefully yield better prediction results.