How accurately can you predict the age group of crabs based on physical properties and gender?

Larissa Bouwknegt

2023-11-01

Introduction

Many countries around the world eat crab, this creates a huge economic market for crab meat. Crab meat can be collected through fishing or through farming. But the production of crab meat through fisheries has shown a decreasing trend due to overfishing and/or changes in the coastal and marine ecosystems. With crab farming the production can be increased without placing undue pressure on the ecosystems population and keeping the food security for the local inhabitants. *(Oniam and Arkronrat 2013)*

So crab farming seems to be a solution for the decreasing trend in crab fishing. But how do you know you have reached the ideal size for farming? In this research the aim is to find the ideal time to farm crabs and build a machine learning algorithm to see if a crab of unknown age would be at the ideal farming stage. This way a farmer can optimize feeding costs and increase his profit.

The aim of this research is to try and train a machine learning algorithm and see how accurate the age of crabs can be predicted.

Materials and Methods

The dataset utilized for this research is the "Crab Age Prediction" dataset from Kaggle *(Sidhu 2021)*, a platform dedicated to open data and machine learning *("Kaggle: Your Data Science Workbench," n.d.)*. The dataset contains the physical properties and age of crabs farmed around the Boston area. The codebook below shows all the attributes and the corresponding measuring units.

Table 1: Codebook for the Crab Age Prediction Dataset, containing all the attributes, their respective measurements and a short description

Attribute	Units	Description
Sex	-	Gender of the crab, female(F) male(M) or interderminate(I)
Length	cm	Lenght of the crab in centimeters
Diameter	cm	Diameter of the crab in centimeters
Height	cm	Height of the crab in centimeters
Weight	gram	Weight of the crab in grams
Shucked.Weight	gram	Weight without the shell in grams
Viscera.Weight	gram	Weight that wraps around your abdominal organs deep inside body in grams
Shell.Weight	gram	Weight of the shell in grams
Age	months	Age of the crab in months

The initial phase involved performing an Exploratory Data Analysis (EDA) using Rmarkdown *(Xie, Allaire, and Grolemund 2023)* to delve into the dataset and prepare it for machine learning. The first step was to

verify if the data was loaded in correctly with the summarize function in R. Next the dataset was checked for missing values or implausible zero values with some simple R code*(Bouwknegt 2023)*, given that the dimensions of a crab cannot logically be zero. Additionally, a conversion from the imperial to metric measurements was made since the research is conducted in the Netherlands where the metric system is used. So foots were converted to centimeters and ounces were converted to grams. Next the distribution of the gender attribute was looked at with a histogram and the weight related and height related attributes were looked at with boxplots in ggplot2*(Wickham 2016)* to see if the data is not skewed. Two extreme outliers in the height attribute were removed since there were no corresponding outliers in the other attributes so it seemed unlikely that these were accurate measurements, the measurements of these crabs would have created some really deformed crabs. Next some boxplots illustrating the relationships between age and weight, age and length, age and height, as well as age and diameter were made to see if multiple categories showed a clear change compared to age. These boxplots showed a stagnation of growth around 10 months. The crab age attribute was split in 2 groups with this in mind, group 1 containing the crabs aged 1 to 10 months and group 2 containing the crabs aged 11 months and older. Next the distribution of these 2 groups was checked with a boxplot to make sure both groups are large enough to perform machine learning on. To see if there was correlation between the different attributes, which would be a positive trait for machine learning, a pairplot and a heatmap were made.

With the EDA completed the cleaned dataset was exported and loaded in to Weka a machine learning tool *(Frank, Hall, and Witten 2016)*. In Weka the following algorithms were used, based on their all around popularity and versatility, without any parameter modifications: Random forest, Bagging, Logistic, SMO, QDA (without the gender attribute since QDA doesn't take nominal values), J48, OneR, ZeroR, AdaBoostM1, IBk, Naïve Bayes, Classification Via Clustering. And the following algorithms were run with modifications: Voting with all of the above algorithms as learners except for QDA, Stacking with all of the above algorithms as learners except for QDA and with a J48 meta learner, Stacking with all of the above algorithms as learners except for QDA and with a random forest meta learner. Experiments in the Weka experimenter were done with smaller datasets with some attributes removed to see if it increases the best scoring algorithms accuracy. This was done based on CfsSubsetEva with exhaustive search forward and backwards, gain ratio with ranking and WrapperSubsetEval based on the best scoring algorithms. 4 subsets were made with different attribute combinations in all of them. The depth of the random forest was explored to see if modifying that will increase the score, 5 different depths were used, namely: 5, 6, 7, 8 and unlimmited. The amount of iterations with bagging were experimented with, 10, 100 and 1000. Different algorithms for the stacking algorithm were tried, this was based on random combinations of algorithms in different algorithm categories and then comparing them. A penalty for predicting group 2, age of 11 and older as group 1 the 1 to 10 months had been tried since there were more mistakes with the predictions in this category based on the confussion matrix, also to compensate for the smaller size of group 2. At last the ROC curves of the best scoring algorithms were plotted and compared. Stacking scored better then random forest when looking at a combination of accuracy and the area under the curve but the gain was minimal so with the time it takes to run in mind the random forest with depth of 7 was chosen as the final algorithm.

This algorithm was then implemented in a Java command line application with a Apache commons cli library to get the command line arguments and the Weka java library to implement the model and predict the outcome. Different options were created using the Apache commons cli library. These options contained the possible command line arguments. These options were further checked with a custom made Java class to collect the desired values and throw errors in case of wrong input formats. With the collected values the already existing model was loaded in or the newly supplied model was loaded in. All the code used can be found on the github. *(Bouwknegt 2023)*

Table 2: Table showing the programs and their corresponding versions for this research

Weka	3.8.3
Java	17.0.5
Apache Commons CLI	1.3.1
Weka java library	3.6.8
R	4.3.1

Results

After the data was loaded the dimensions of the dataset were looked at to get an first impression of the size and dimensions of the dataset.

Table 3: Dimensions of the crab data. Shows the amount of attributes and the amount of crabs measured

Column count	9
Row count	3889

Table 3 shows that there are 9 attributes measured in the dataset, which can be found in table 1 and 3893 rows which means that there were 3893 individual crabs measured for the dataset. The next step was determining if the data is complete by looking at the zero and NA values. This revealed no missing values but revealed 2 crabs with a height of 0, since there could not possibly exist a crab with a height of 0,0 feet were these 2 entries removed. A conversion of feet to centimeters was made by multiplying with 30.48 and from ounces to grams was done by multiplying with 28.3495231. Next the distributions of the attributes were looked at.

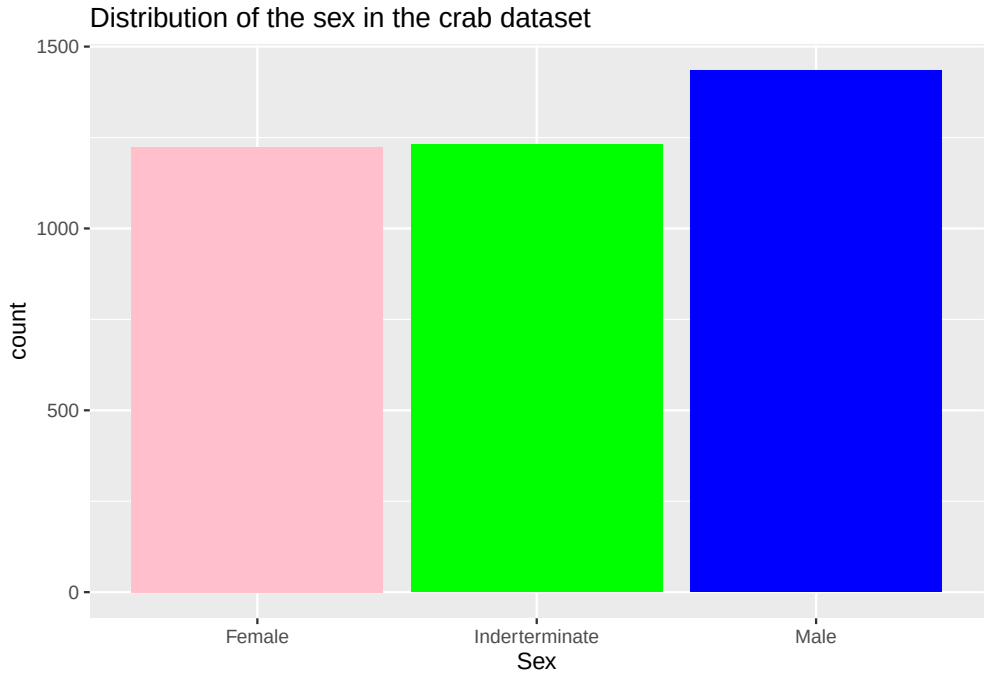

Figure 1: Distribution of the gender in the crab dataset for all 3893 crabs, pink for female, green for inderterminate and blue for male.

The distribution of the gender of the crabs was plotted in figure 1. It showed a fairly even distribution between the female and the indeterminate sex of the crabs. There seem to be slightly more males in the dataset. This distribution is of good use for machine learning.

Next all the attributes with numeric values were looked at.

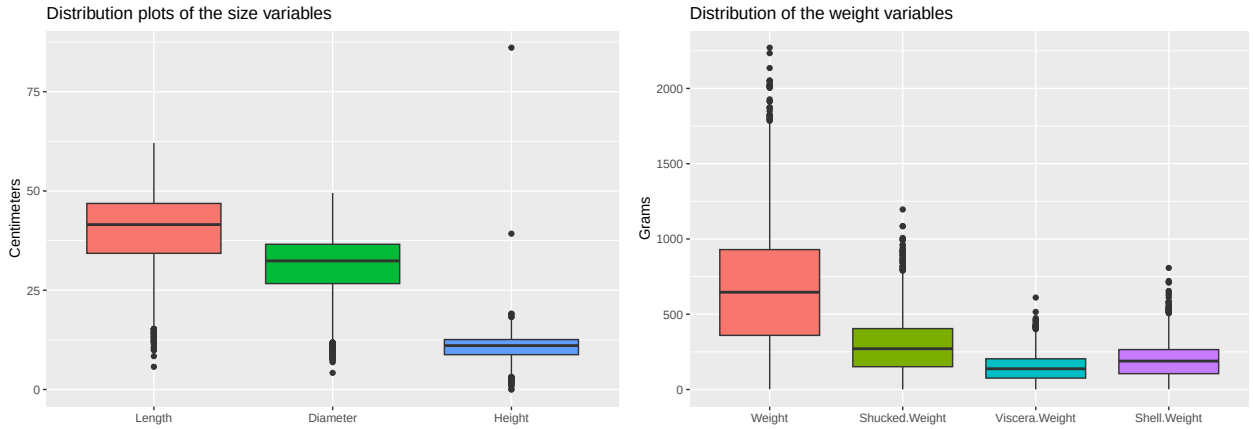

Figure 2: Distribution of the numeric atributes in the crab dataset. The left graph shows the attributes related to length and are in centimeters, the right graph shows the attributes related to weight and are in grams

Figure 2 shows that all the attributes seem to have outliers but there appeared to be 2 extreme outliers in the height attribute. With 2 outliers this far up in height corresponding outliers in the other attributes would have been expected, since this wasn't the case these 2 entries of the crabs have been removed. These 2 entries would otherwise be crabs with proportions which seem unlikly to be found in nature.

The age of the crabs was measured in months varying from 1 to 29 months. But a regression model was out of scope for this project. So the age of crabs needed to be split in groups. A plot with the age against some of the attributes has been made to see if there would be 1 or more logical points to split in to groups.

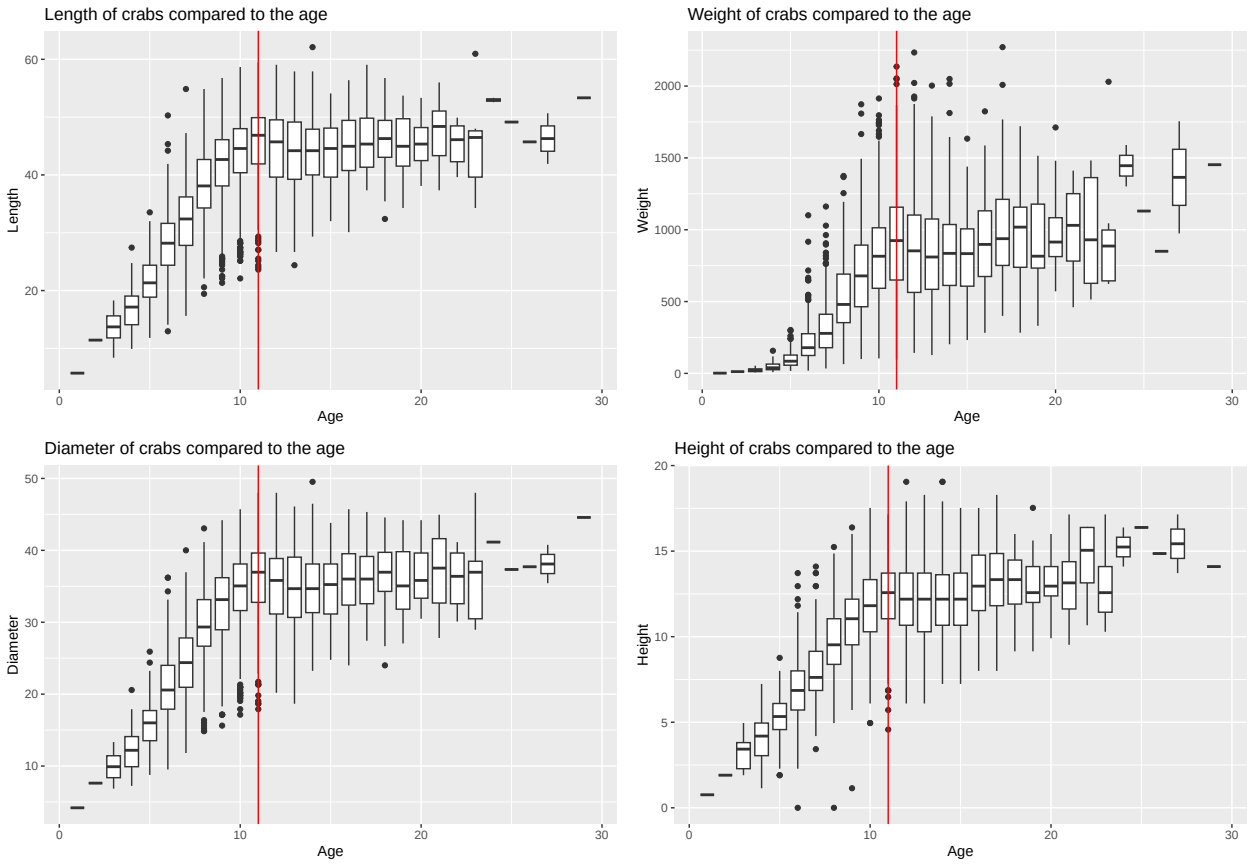

Figure 3: Atributes distribution based on the age of crabs. 4 different attributes were plotted as boxplots against the age to see if there could be groups. The red line is where the growth pattern of the crabs seems to change

Figure 3 shows the age attributed plotted against the lenght, the weight, the diameter and the height of crabs in a boxplot. This shows that the crabs show a steady growth until 10 months in all of the attributes. At 11 months and more they still seem to slowly grow but the pace has decreased a lot. This point, the 11 months, has been marked by a red line to show the sepperation in the growth curve. This seemed to be the ideal cutoff point to create 2 groups for machine learning. So 2 groups of crabs have been made based on the age. The first grouped, marked by a 1, is of the crabs until 10 months. The other group marked 2 is of the crabs of 11 months and above. This new distribution has been checked again.

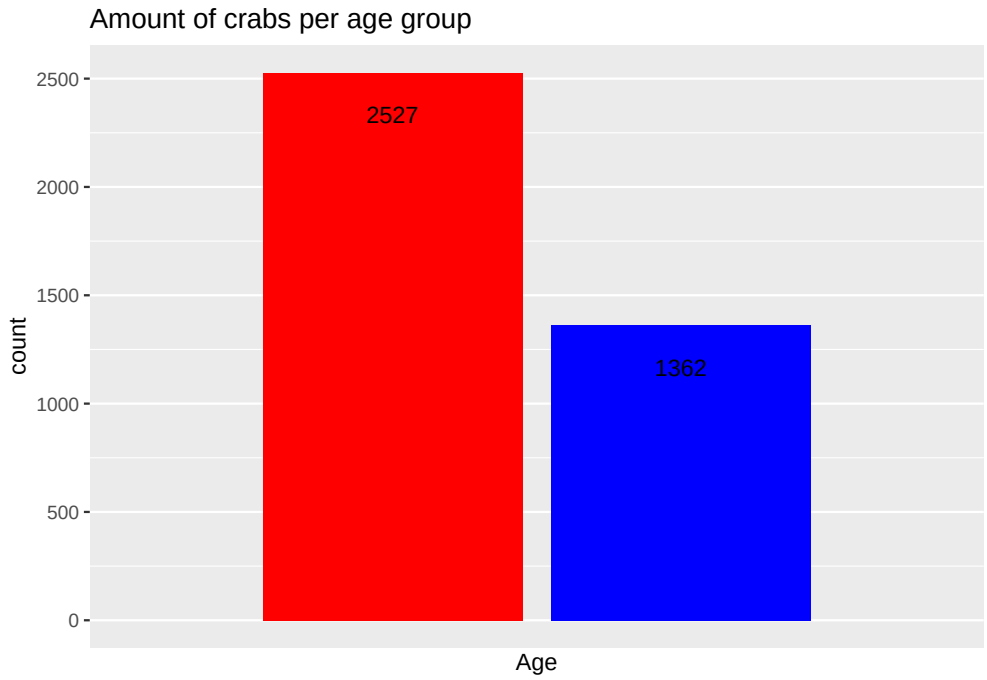

Figure 4: Distribution of the new age groups. Red shows group 1 which are the crabs until 10 months and blue shows the crabs of 11 months and older.

Figure 4 shows the new distribution of the age group. Group 1 is larger then group 2 but there seem to be enough entries left in group 2 to be able to use machine learning without any problems. The data appeared to be skewed with more young crabs.

To see if there was any correlation a pairplot was made.

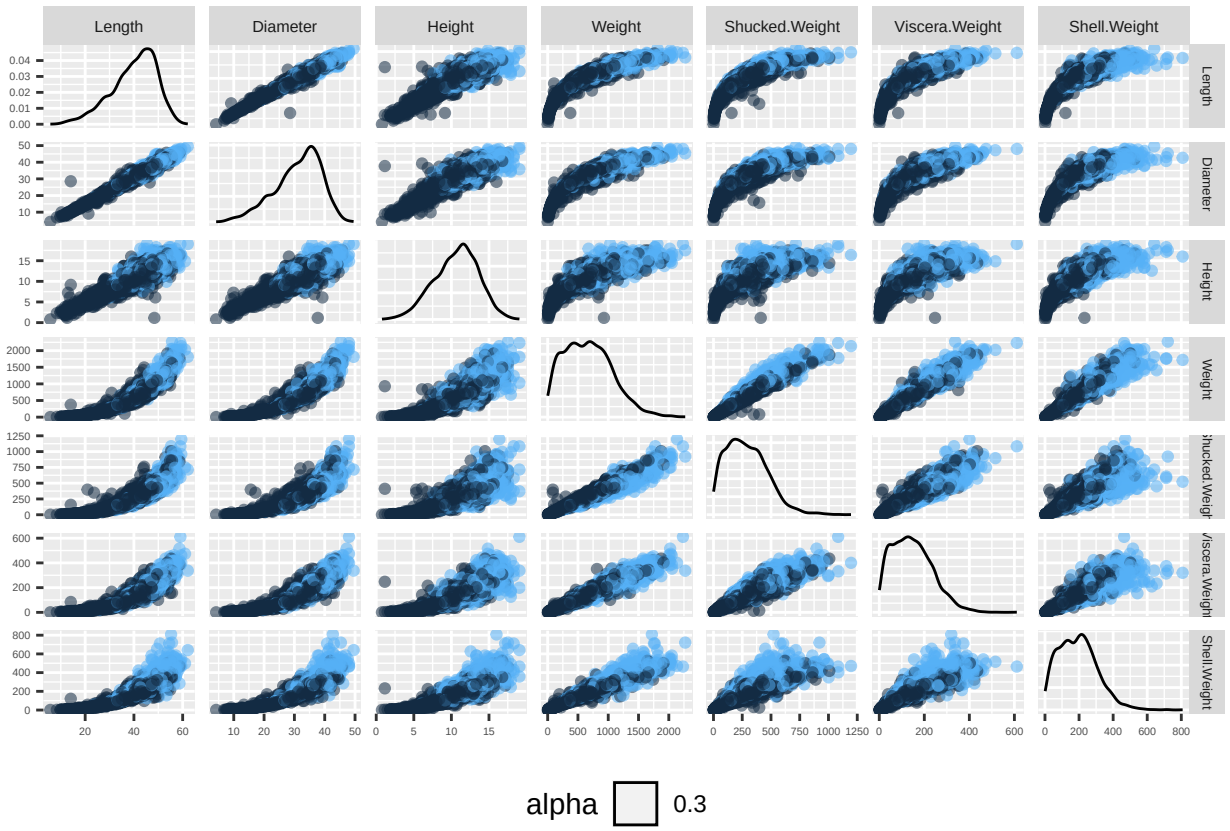

Figure 5: Pairwise plot of the continuous atributes with the age as color to see if there is any correlation. Darkblue is group are the crabs until 10 months, light blue are the crabs of 11 months and older

Figure 5 shows that there is correlation between the attributes and that most of age group 1 is on the left side of the pairplots and most of age group 2 is on the right side of the pairplots. The strenght of the correlation has been further investigated in a heatmap below. The sort of seperation between age group 1 and 2 shows promissing results for machine learning based on physical properties.

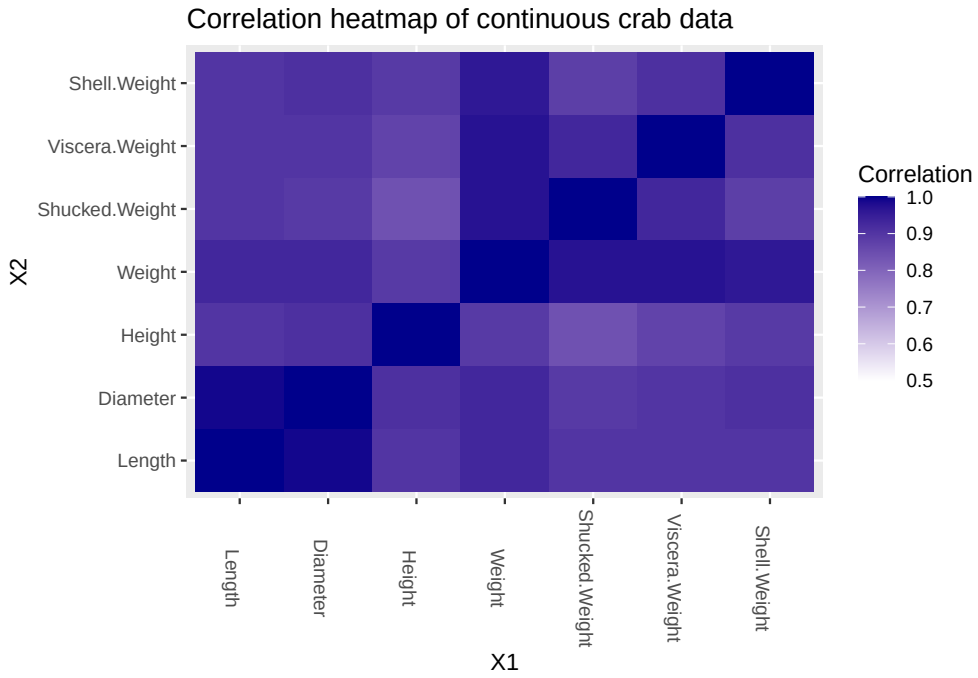

Figure 6: Heatmap of the continious variables and their correlation to each other

Figure 6 shows that the lowest correlation seems to be between the shucked weight and the height, but this is still around 0.8 which tells that event he lowest correlated attributes are highly correlated. The variable that seems to be correlating the most is the weight variable, not only does it get high correlation scores in the weight variables but also the length and diameter seem to be of importance.

Lastly before actual machine learning the mean difference between the age groups and the continuous variables was looked at with a T-test to see which categories are most likely to be used as a predictive variable in machine learning.

Table 4: Mean Differences between Groups calculated through a T-test to see which attributes are most likely the most predictive

Column	Mean_Difference
Length	-7.735194
Diameter	-6.694571
Height	-2.801867
Weight	-358.013809
Shucked.Weight	-126.216612
Viscera.Weight	-75.847967
Shell.Weight	-116.557100

Table 4 shows that the weight variables seem to have the biggest mean difference and will most likely be the best predictive variables in machine learning. This corresponds with the heatmap of figure 6 where the weight variable seemed to have the most correlation of all the attributes.

The next step was the actual machine learning. In Weka the following algorithms were used, based on their all around popularity and versatility, without any parameter modifications: Random forest, Bagging, Logistic, SMO, QDA (without the gender attribute since QDA doesn't take nominal values), J48, OneR, ZeroR, AdaBoostM1, IBk, Naïve Bayes, Classification Via Clustering. And the following algorithms were

run with modifications: Voting with all of the above algorithms as learners except for QDA, Stacking with all of the above algorithms as learners except for QDA and with a J48 meta learner, Stacking with all of the above algorithms as learners except for QDA and with a random forest meta learner. The results can be found in table 5.

Table 5: Table showing all the algorithms tried at the start with minimal changes to the algorithms. Ranked based on the scores achieved.

algorithm	Percentage_Correct	Percentage_Wrong
Stacking with random forest meta	78.272	21.728
Random Forest	77.9121	22.0879
Bagging	77.7578	22.2422
Logistic	77.7064	22.2936
Vote	76.6264	23.3736
SMO	76.4464	23.5536
Stacking with j48 meta	76.4207	23.5793
QDA (sex atribute removed)	75.7521	24.2479
j48	75.495	24.505
OneR	73.9779	26.0221
AdaBoostM1	72.1265	27.8735
IBk	70.5837	29.4163
Naive Bayes	68.1409	31.8591
ZeroR	64.9781	35.0219
ClassificationViaClustering	62.4839	37.5161

The best scoring algorithms are, as can been seen in table 5: Stacking with a random forest, Random forest, bagging, logistic and voting. These algorithms were further looked into to optimize the parameters and the rest of the algorithms were not used further.

The tree had been improvised by tweaking the depth size. The results of different depths can be found in table 6 below.

Table 6: The results of the different depths sizes in the random tree algorithm and their corresponding accuracy on the crab data.

tree_depth	percentage_corect
unlimited	70.6351
5	74.1322
6	75.0064
7	75.6236
8	74.7750

Table 6 shows the changes to the depth of the tree parameter and how they corresponded to the percentage correct with 10 fold cross validation on the crab dataset. Next a forest has been build with trees with a depth of 7 since this seemed to be the best parameter score wise. A forest with the depth of 7 gave a 78.5035 percent accuracy. By increasing the trees used under the iteration value to a 1000 instead of the standard 100 gave a 78.7349 percent accuracy. Increasing this to 10000 did take alot more time and did not improve the score at all.

The same had been done for bagging but with the iteration as can been seen in table 7 below.

Table 7: Different amounts of iterations in the bagging algorithm and their corresponding accuracies

itterations	percentage_correct
10	77.7578
100	78.2206
1000	78.1435

Changing the number of iterations improves it when it is set to a 100 but decreases when set to a 1000. A 1000 iterations also takes more time and it is getting on the long side there with a time of 20.8 to build the model. So bagging seemed to be optimal at a 100 iterations.

Stacking seemed to improve when using less and simpler algorithms. Also with j48 as a meta learner. After some more experimenting with different algorithms and settings the highest percentage seems to be 79.0177 correct with an AUC of 0,817. The settings used are the following base algorithms: Random forest with depth of 7, logistic, IBk with 3 neighbours, bagging with 100 iterations and OneR. The meta learner used was J48 without any changes to the settings. Using less algorithms from different categories and improving their settings led to this.

An ROC curve was made of the 4 best algorithms with the changed variables.

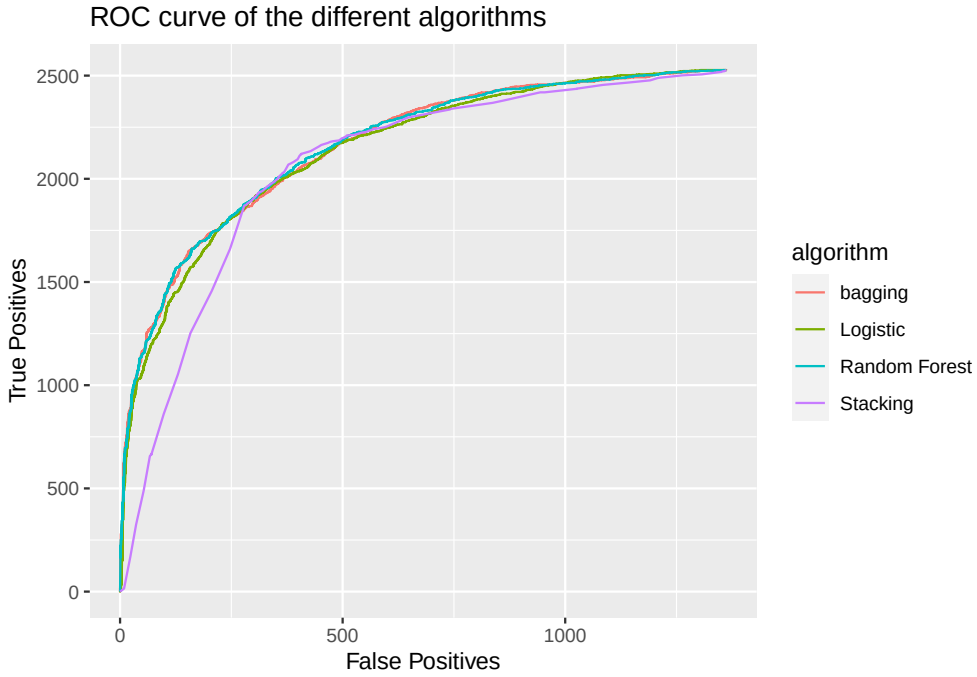

Figure 7: ROC curve of the 4 best scoring algorithms, based on age group 1 as outcome

The 2 best scoring algorithms were stacking and random forest, but the accuracy was slightly higher with stacking. But the graph in figure 7 revealed that stacking had a worse ROC curve, stacking was also more time consuming. For that reason a wrapper had been build with a random forest with the depth of 7 and a 78.43% accuracy.

Discussion and Conclusion

The dataset used in this research, sourced from Kaggle, lacks information regarding the specific crab species measured. This is something to keep in mind, as different crab species may exhibit different growth curves. The growth boxplots over time reveal a slight decrease after 10 months, indicating potential challenges in accurate age prediction for crabs with similar proportions in both age groups, since the crabs in group 2 can have the same porportions as the onces in group 1. The exploratory data analysis (EDA) showed correlations within different attribute categories. It can be noted that there is a change in crab growth patterns around 11 months, where growth appears to stagnate. This observation led to the seperation of crabs into two age groups (1-10 months and 11 months and above). Experimantation with different algorithms led to a random forest algorithm with a depth of 7 which gave a 78.43% accuracy. The answer to how accurate can you predict the age of crabs with machine learning based on physical properties, seems to be answered with that.

Yet I believe there is a lot to gain when more information is collected such as which species and are they farmed or found in the wild. Further research with this algorithm could be by training it on datasets with more information.

References

Bouwknegt, Larissa. 2023. "Thema-9: GitHub Repository." https://github.com/376308/Thema-9.

Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench*. Morgan Kaufmann.

"Kaggle: Your Data Science Workbench." n.d. https://www.kaggle.com.

Oniam, Vutthichai, and Wasana Arkronrat. 2013. "Development of Crab Farming: The Complete Cycle of Blue Swimming Crab Culture Program (CBSC Program) in Thailand." *Journal of Fisheries and Environment* 37 (2): 31–43. https://li01.tci-thaijo.org/index.php/JFE/article/view/80657.

Sidhu, Gursewak Singh. 2021. "Crab Age Prediction." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2834512.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2023. *R Markdown: Dynamic Documents for r*.