Crab body metrics - predicting the subspecies of crabs

IJsbrand Pool, 403589

Contents

1 Recapitulation

The goal of this project was to find out if the species of a *Leptograpsus variegatus* was predictable. The research question for this project is "Can the species of a *L.variegatus* be determined based on some morphological measurements of its carapace". The data needed for this contains 5 morphological measurements, measured in millimeters. To answer this question, the data was first explored and cleaned. Visualizations were created to show these explorations. Then, multiple machine learning algorithms were tested to find what algorithm could be used best. After these tests, the SimpleLogistics algorithm reached a perfect 100% accuracy. This could be because the dataset is very evenly divided, with exactly 100 blue and 100 orange crabs.

2 Introduction

The rock crab *Leptograpsus variegatus*, is recorded as occurring on a number of southern Pacific islands, the western coast of South America, and the coasts of Australia south of the Tropic of Capricorn. Mahon, using ecological studies which extended those of Shield, and a genetical analysis based on an electrophoretic study, established the specific distinctness of rock crabs of the blue and orange forms of the genus *Leptograpsus* which occur on the coasts of Australia. These colour forms were previously regarded as morphs of *L. variegatus.*

In an attempt to resolve this problem of identification, a morphological study of the Western Australian species was undertaken. This paper reports an exploratory data analysis of the data and a machine learning algorithm to predict the species of the crab based on this data.

The dataset used is the crab body metrics dataset by Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus *Leptograpsus*[1]. This data set contains multiple morphological metrics of the bodies of *L. variegatus.* crabs, and the gender and color of the crab. The measured metrics are the frontal lobe size, rear width, carapace length, carapace width and body depth. All these values are in millimeters. There are 100 orange and 100 blue crabs. 50 crabs of each gender per color of crab.

3 Materials and methods

This project researches if it is possible to predict the species of *L.variegatus* based on the morphological measurements of its carapace using machine learning. Multiple machine learning algorithms were used to find what algorithm has the best accuracy of predicting this. The data ws also described and visualized to help answering the research question.

3.1 Materials

The data used in this project was publicated by Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus *Leptograpsus*[1]. This paper contained useful information for this project. The dataset used was a csv file containing multiple morphological measurements for 200 crabs. Not all attributes are as important, so it was part of the reasearch question to determine what attributes could be best used. Multiple packages were used for this project, see table 1.

Table 1: The used packages and their versions.

packagelist	versionlist
kableExtra	1.3.4
ggplot2	3.3.5
factoextra	1.0.7
RWeka	0.4-43
reshape	0.8.8
FactoMineR	2.4
gridExtra	2.3

3.2 Methods

There were multiple methods used for this project. Most of the packages contained one or more methods that were used for the creating this paper and its graphs. This paper was made in Rstudio, using Rmarkdown. Multiple machine learning algorithms were also used for this project. These were all used in the weka software. The java wrapper was made in the intelliJ software.

4 Results

4.1 Exploratory data analysis

To give a better overview of the data, this exploratory data analysis was created. To do this, the initial data was first examined to see how the dataset is set up. Then, multiple graphs were made to determine if any of the attributes were connected. After that it is possible that the data might have to be changed or cleaned to be able to use it for machine learning algorithms.

4.1.1 Data description

After the data was loaded in, a codebook with the attribute names and their descriptions was generated, shown in table 2. To get a better view of the measurements in the dataset, a five number summary was created for each attribute. These values are shown in table 3. The table shows that the mean and the median are close in value for each column, meaning that they all are normal distributions.

Table 2: Codebook of the dataset

column	description
sp	Species
sex	Sex
index	Index
FL	Frontal lobe size (mm)
RW	Rear width (mm)
CL	Carapace length (mm)
CW	Carapace width (mm)
BD	Body depth (mm)

Table 3: Five number summary of the morphological measurements of the crab bodies.

	Frontal Lobe	Rear Width	Carapace Length	Carapace Width	Body Depth
Minimum	7.20	6.50	14.70	17.10	6.10
Q1	12.90	11.00	27.27	31.50	11.40
Median	15.55	12.80	32.10	36.80	13.90
Mean	15.58	12.74	32.11	36.41	14.03
Q3	18.05	14.30	37.23	42.00	16.60
Maximum	23.10	20.20	47.60	54.60	21.60

4.1.2 Data visualisation

To get a better view of how these attributes are correlated, multiple graphs were made. These graphs show the corralation between 2 attributes with the colored data points. These data points are colored to the color of the crab.

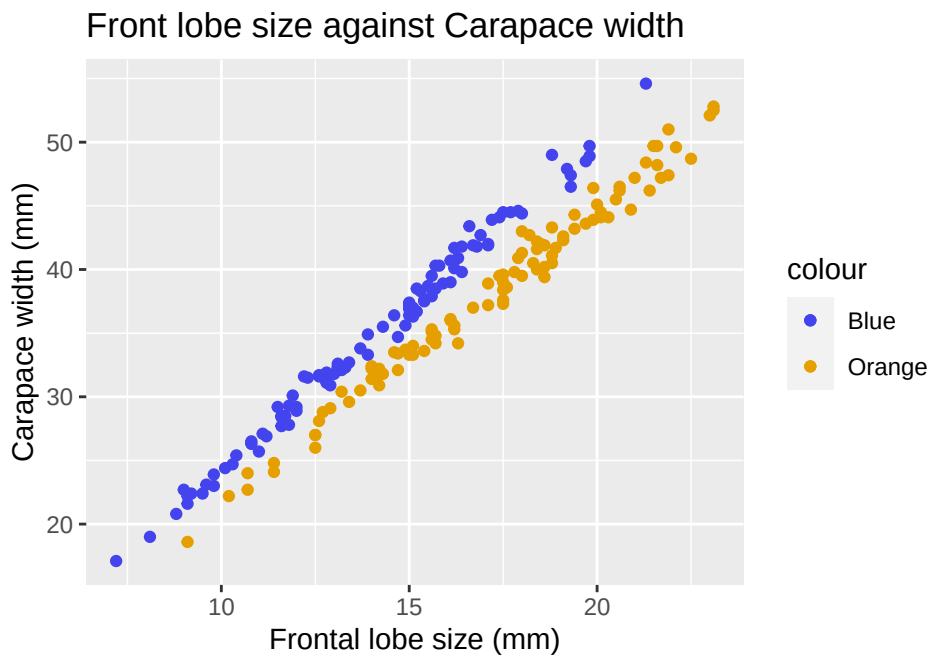


Figure 1: Spread of Front lobe size against Carapace width based on color

This plot shows that blue crabs on average have wider carapaces and shorter frontal lobes. This could be a good indicator to determine the subspecies of the crab.

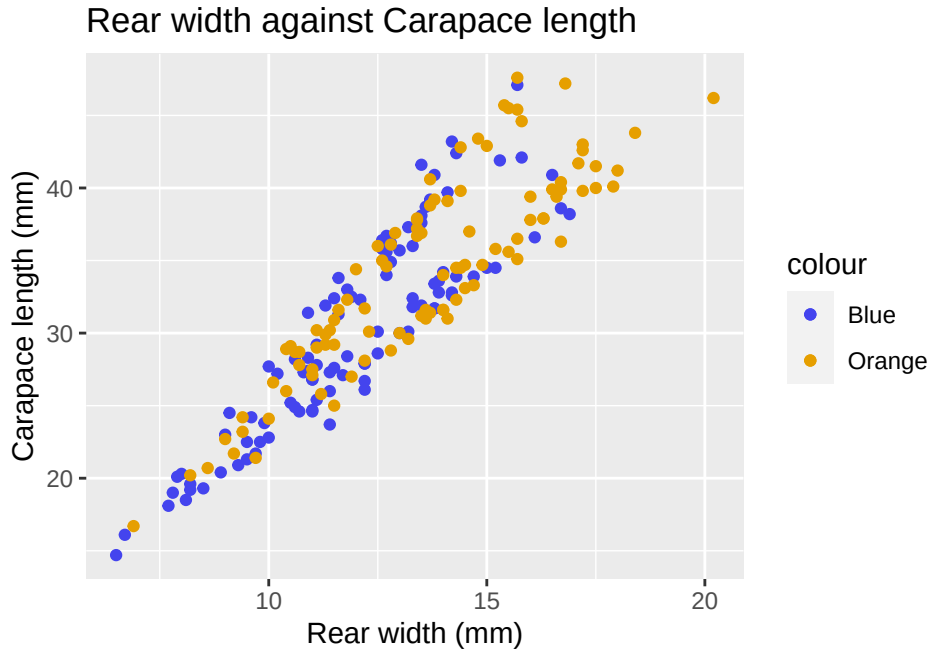


Figure 2: Spread of Rear width against Carapace length based on color

This plot does not show a clear difference between blue and orange crabs. Still there are 2 separated groups, this could be investigated further.

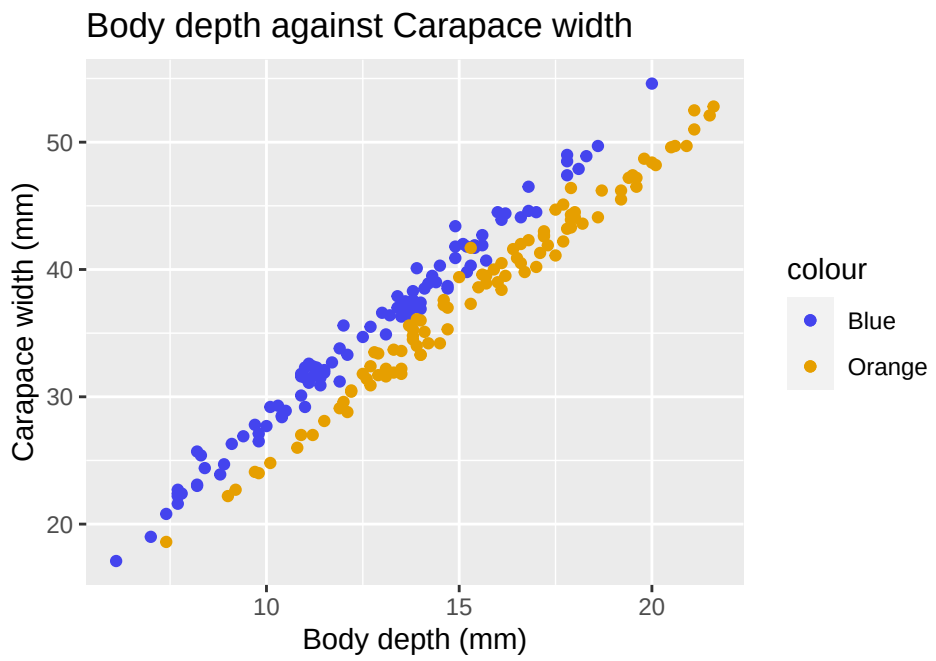


Figure 3: Spread of Body depth against Carapace width based on color

This plot shows again that blue crabs on average have wider carapaces, but that orange crabs tend to have deeper bodies. This could also be a good indicator to determine the subspecies of the crab, though its less clear than figure 1.

As shown in the figures one and three, the metrics of the orange and the blue crabs are in two separate groups. This could mean that the two species could be identifiable by their carapace width. However, figure two also has two clear groups but the colors of the crabs are mixed. This could mean that these two groups are separated by gender.

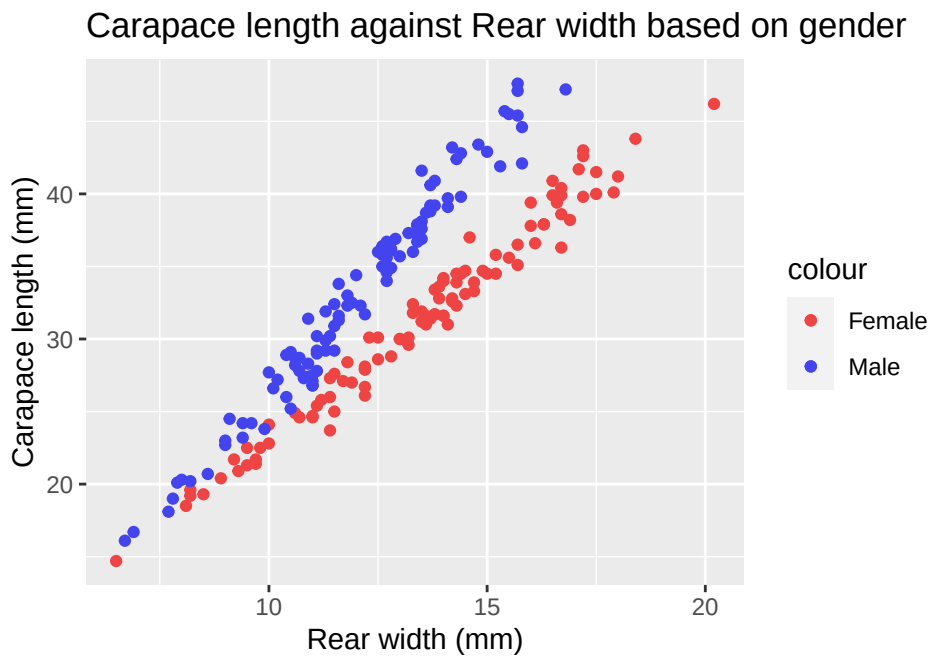


Figure 4: Spread of Carapace length against Rear width based on gender

As shown in figure 4, it is indeed two groups of the genders of the crabs. This means that the males have longer carapaces than the females. The difference in carapace length for both genders are explored further.

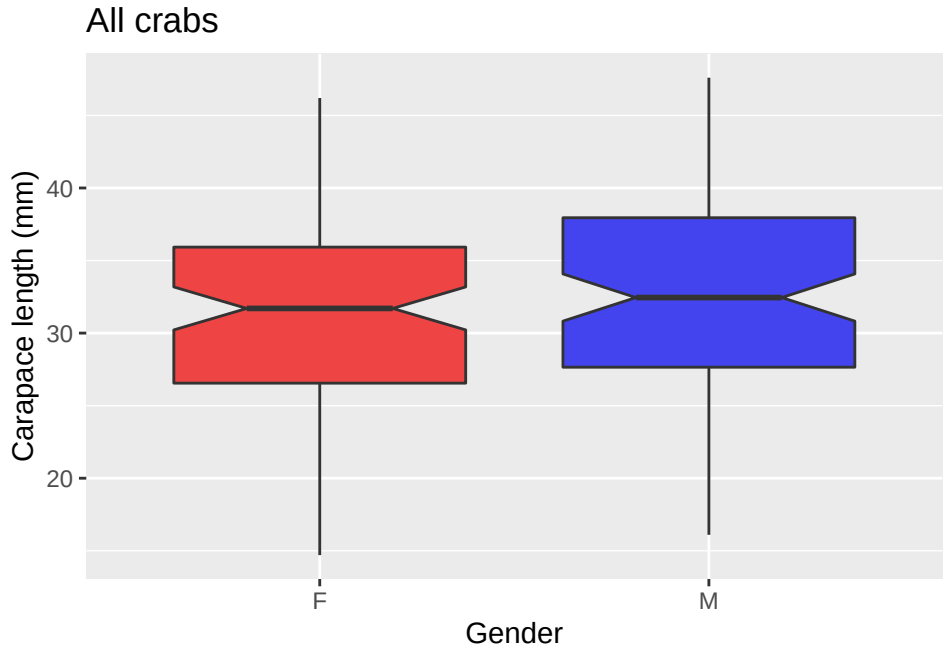


Figure 5: Distribution of Carapace length for all male and female crabs

This plot does not show a big difference between male and female crabs. The female crabs have a slightly shorter carapace, but it does not seem significant. This could be investigated further, but is not much

connected to the research question.

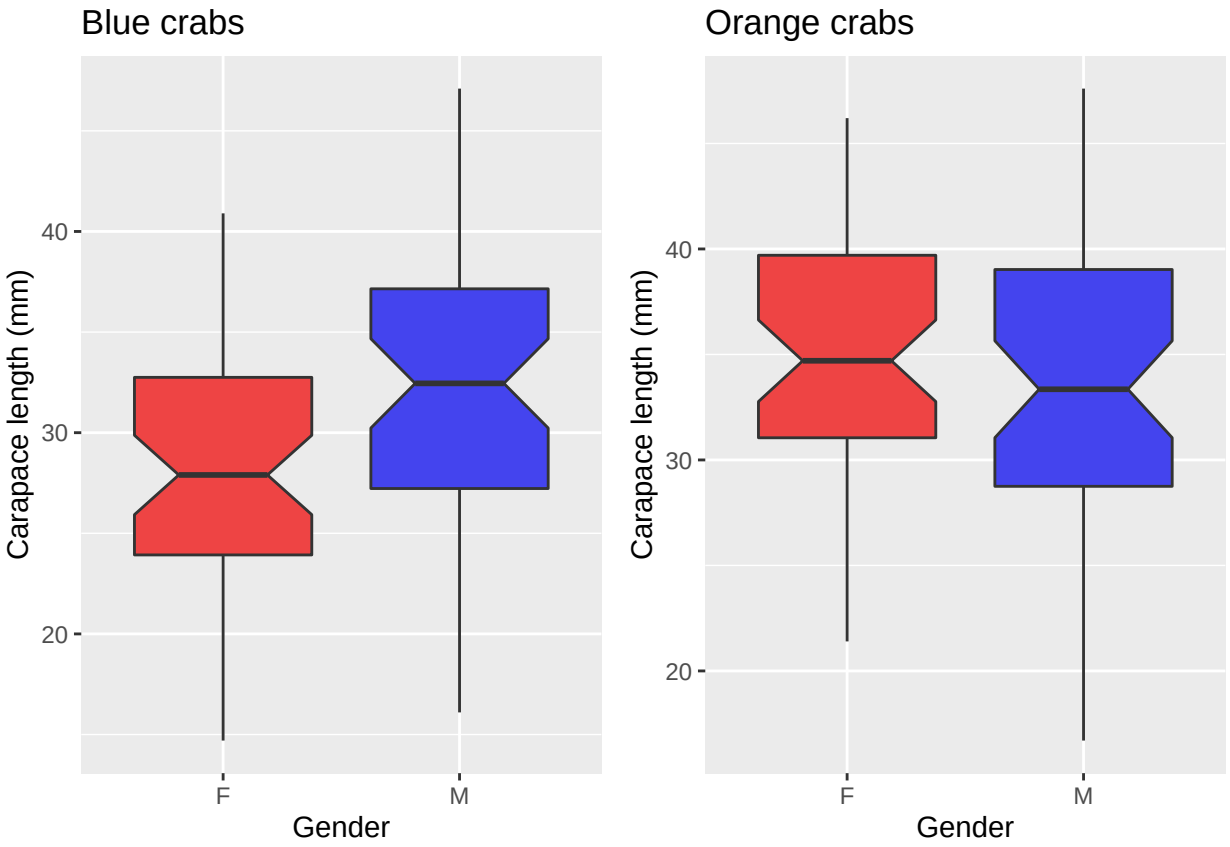


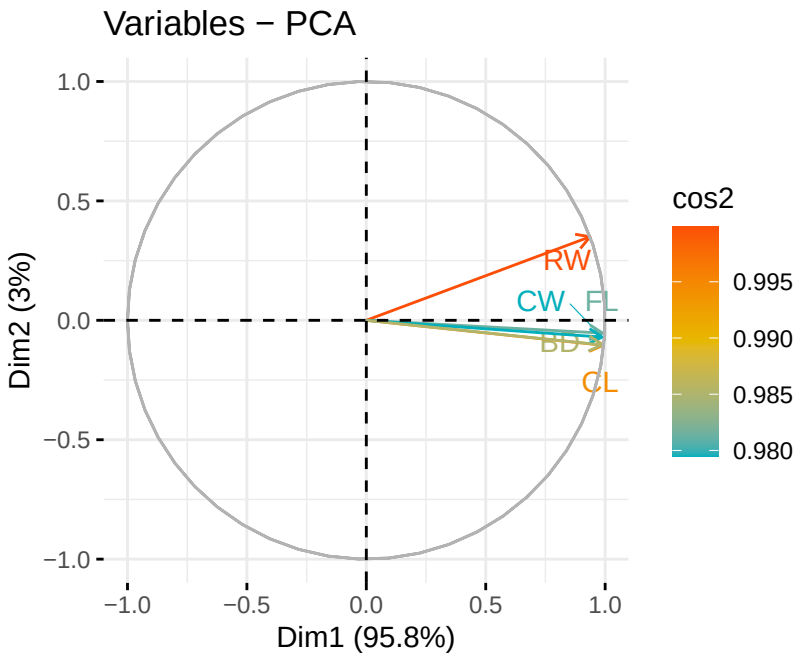
Figure 6: Distribution of Carapace length for the orange and the blue male and female crabs

This plot shows a small difference in the length of the carapace between male and female orange crabs. It shows that the male crabs have on average a shorter carapace, but the difference is not that big.

This plot shows a bigger difference in the length of the carapace between male and female blue crabs. Here, the difference is means does seem significant.

As shown in figure 7, it is mainly the blue crabs that have a significant difference in carapace length between genders. It shows that the male crabs have longer carapace lengths on average. In figure 6 its clear that the male orange crabs have shorter carapace lengths on average compared to the females.

A variables PCA plot was generated to see how correlated all the variables are.



This PCA plot shows that the variables are highly correlated. The least correlated variable is the Rear width. This vector has the largest angle with the Carapace length, the same two variables were used to discover the difference in carapace length distribution between the genders.

4.2 Cleaning of the data

In its current form, the data is not ready for machine learning. Because of the id column, some of the machine learning algorithms overfit their model by using this column. So this column should be removed first. The species columns was also moved to the last column so it would be used as the class attribute.

4.2.1 Machine learning

To find out what machine learning algorithm is the best for predicting the species of crab, multiple algorithms were tested. These algorithms are ZeroR, OneR, Simple logistic, Naive bayes, Random forest, J48, SMO and K-nearest neighbor. These algorithms were tested using 10 fold cross-validation. The highest quality metric for this dataset is the accuracy, since it does not matter whether a blue crab is predicted to be orange, or an orange crab to be blue. The software used to calculate the accuracy is weka. After the classification, the accuracy of these algorithms was saved in a csv file, and are shown in this barplot below.

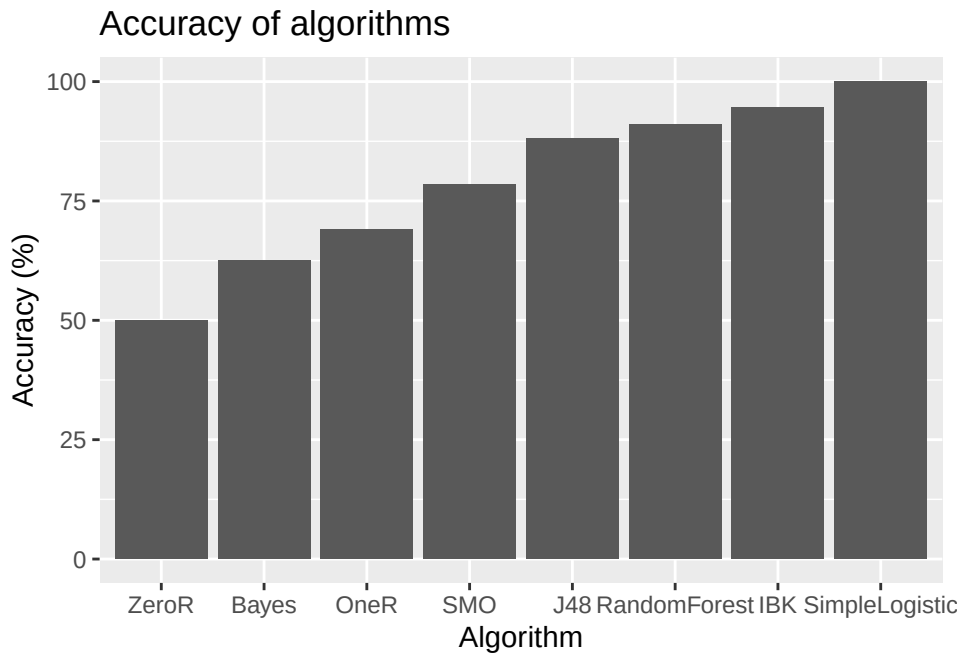
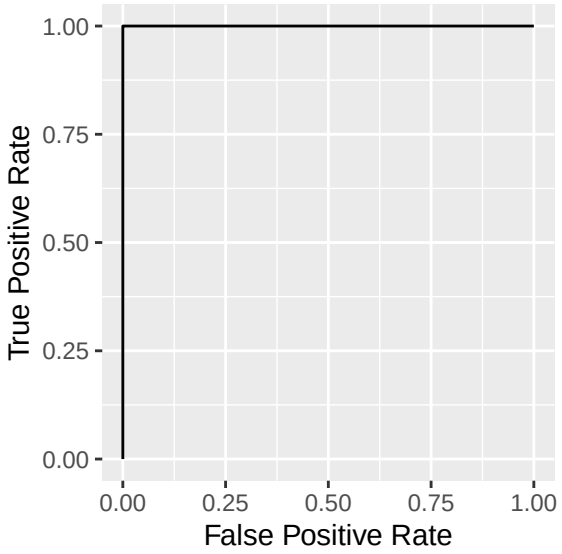


Figure 7: The accuracy of the machine learning algorithms ordered from low to high.

This plot shows a few interesting things, most notably the 100% accuracy of Simple logistic. The output in weka of the classification using Simple logistic with 10 fold cross-validation shows a model for each species. The model for the blue crabs shows 1.03 + [FL] * -0.6 + [CW] * 0.31 + [BD] * -0.22. The model for the orange crabs shows -1.03 + [FL] * 0.6 + [CW] * -0.31 + [BD] * 0.22. The values in the model of the orange crabs are the values of the model of the blue crabs times -1. The metrics used in the models are Front lobe size, Carapace width and Body depth. The barplot also shows an exact 50% accuracy for the ZeroR algorithm. This is expected since the data has the same amount of blue crabs as orange crabs.

Using the output from the simple logistic algorithm, the following roc curve can be made.



The curve is of course two straight lines because the accuracy is 100%.

5 Conclusion & Discussion

The goal was to get the dataset ready for machine learning. The data was analyzed and cleaned to make so it is ready to be used in machine learning algorithms. The data does not contain many outliers. The data points seem easy to classify since most plots show clear groups of blue and orange crabs. As shown in figure 4, the gender of the crab could also be a good attribute to help predict the species of crab. The data also had to be cleaned. This was done by removing the index column, since this column can not be used to help determine the species of crab. It might also be a problematic attribute for some machine learning algorithms. Then, the species column was moved to the last column, so the machine learning algorithms will use this column as the class index.

6 Sources

6.1 References

[1] Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.