Log of Crab body metrics - predicting the subspecies of crabs

IJsbrand Pool, 403589

Contents

1 Logbook EDA

1.1 Dataset information

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.

1.1.1 Research question

The goal of this project was to find out if the subspecies of a *Leptograpsus variegatus* was predictable.To find out if this is possible, a research question had to be formulated. The research question for this project is "Can the species of a *L.variegatus* be determined based on some morphological measurements of its carapace". To answer this question, the data was first explored and cleaned. Then, multiple machine learning algorithms were tested to find what algorithm could be used best.

1.2 Data analysis

```
#load in the data
myData <- read.csv("datafiles/data.csv")
```

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.

```
#create the codebook
column <- colnames(myData)
description <- c("Species", "Sex", "Index", "Frontal lobe size (mm)", "Rear width (mm)", "Carapace leng
codebook <- data.frame(column, description)
kable(codebook, caption = "Codebook of the dataset") %>%
kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: 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)

```
sumdat <- summary(myData[4:8])
sumdat <- sub(".*:", "", sumdat)
rownames(sumdat) <-  c("Minimum", "Q1", "Median", "Mean", "Q3", "Maximum")
colnames(sumdat) <- c("Frontal Lobe", "Rear Width", "Carapace Length", "Carapace Width", "Body Depth")
kable(sumdat, caption = "Five number summary of the morphological measurements of the crab bodies.") %>%
kable_styling(latex_options = "hold_position")
```

Table 2: 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

1.3 Visualisation

1.3.1 Scatterplot for frontal lobe size against carapace width

```
#plot the Front lobe size against Carapace width
ggplot() +
  geom_point(data = myData[myData$sp == "B",], mapping = aes(x = FL, y = CW, color = 'Blue')) +
  geom_point(data = myData[myData$sp == "O",], mapping = aes(x = FL, y = CW, color = 'Orange')) +
  scale_color_manual(values=c("#4444EE", "#E69F00")) +
  labs(x = "Frontal lobe size (mm)", y = "Carapace width (mm)", title='Front lobe size against Carapace
```

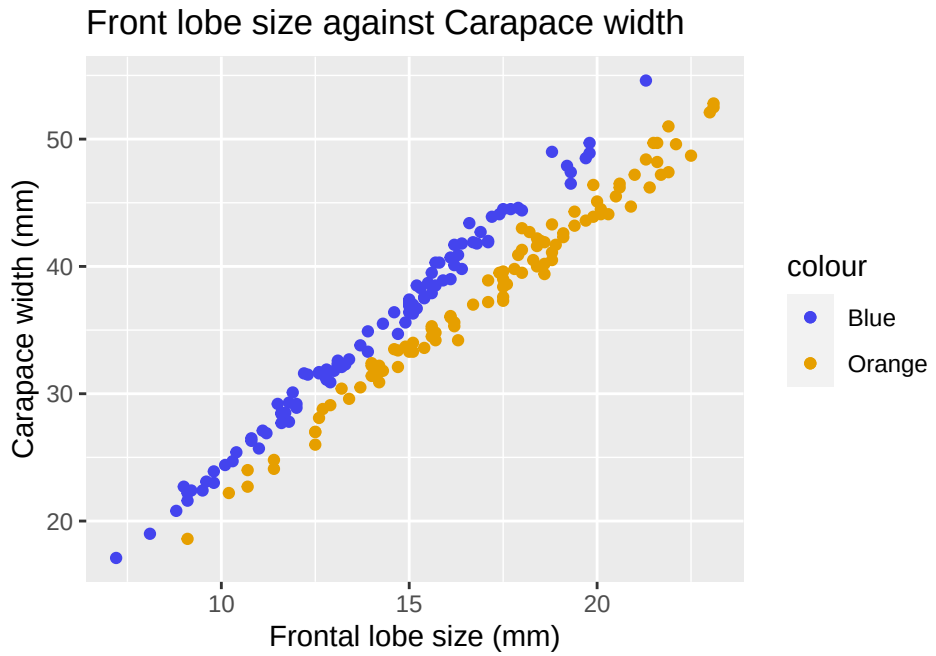


Figure 1: Spread of Front lobe size against Carapace width based on color

This plot plots the frontal lobe size against the carapace width of the blue crabs as blue dots, and of the orange crabs as orange dots. It shows that blue crabs on average have wider carapaces and shorter frontal lobes, and that these attributes are somewhat correlated. This could be a good indicator to determine the subspecies of the crab.

1.3.2 Scatterplot for rear width against carapace length

```
#plot the Rear width against Carapace length
ggplot() +
  geom_point(data = myData[myData$sp == "B",], mapping = aes(x = RW, y = CL, color = 'Blue')) +
  geom_point(data = myData[myData$sp == "O",], mapping = aes(x = RW, y = CL, color = 'Orange')) +
  scale_color_manual(values=c("#4444EE", "#E69F00")) +
  labs(x = "Rear width (mm)", y = "Carapace length (mm)", title='Rear width against Carapace length')
```

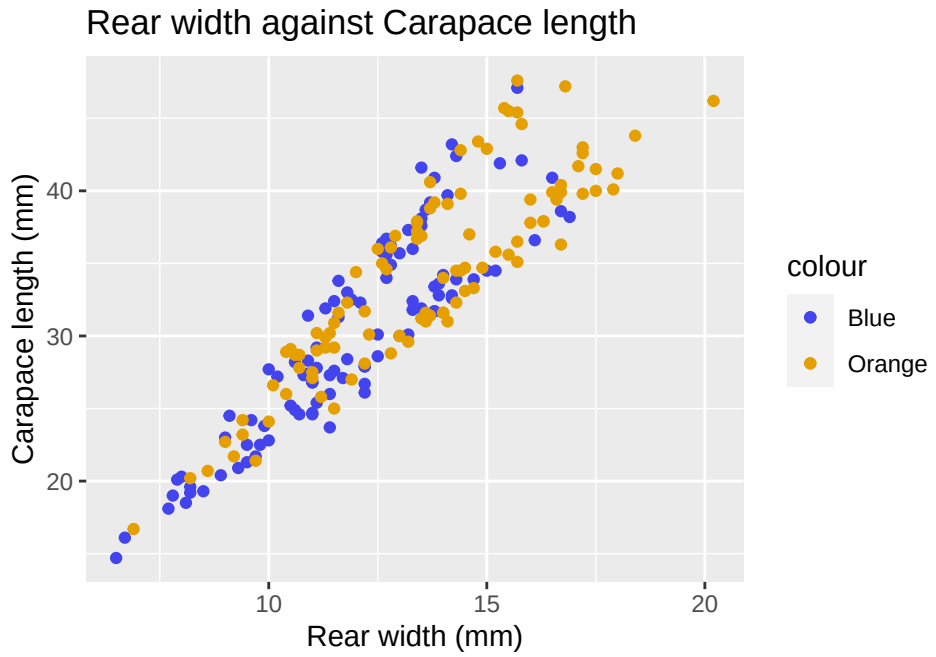


Figure 2: Spread of Rear width against Carapace length based on color

In this plot, the datapoints of the blue crabs are colored blue, and the datapoints of the orange crabs colored orange again. This plot however, does not show a clear difference between blue and orange crabs. Still there are 2 separated groups, this means that these attributes are also somewhat correlated and this could be investigated further.

1.3.3 Scatterplot for body depth against carapace width

```
#plot the Body depth against Carapace width
ggplot() +
  geom_point(data = myData[myData$sp == "B",], mapping = aes(x = BD, y = CW, color = 'Blue')) +
  geom_point(data = myData[myData$sp == "O",], mapping = aes(x = BD, y = CW, color = 'Orange')) +
  scale_color_manual(values=c("#4444EE", "#E69F00")) +
  labs(x = "Body depth (mm)", y = "Carapace width (mm)", title='Body depth against Carapace width')
```

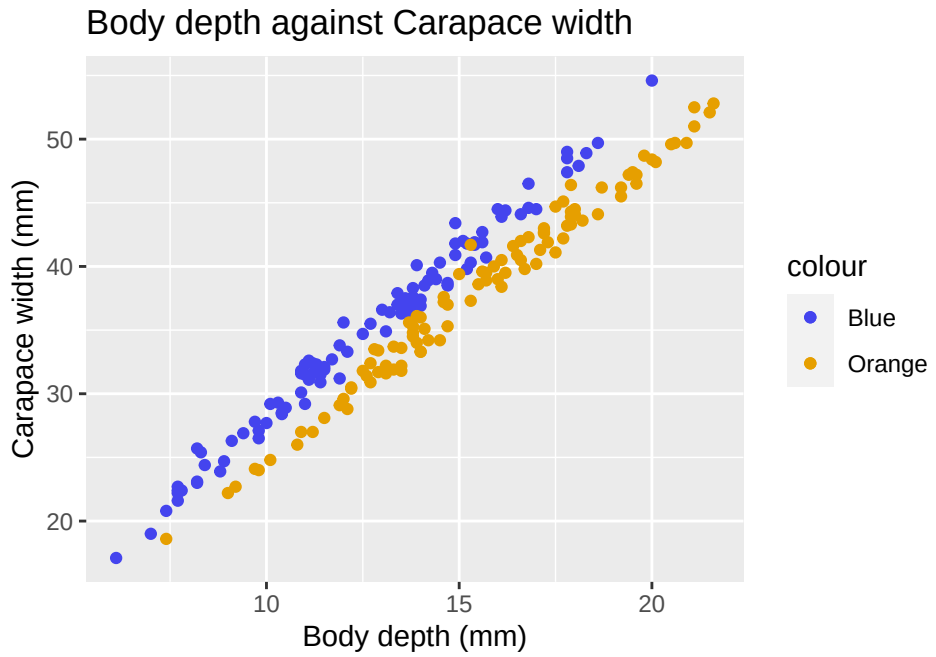


Figure 3: Spread of Body depth against Carapace width based on color

In this plot, again the datapoints were colored to represent the color of the crab. This plot also shows that blue crabs on average have wider carapaces, but that orange crabs tend to have deeper bodies, and that these attributes are somewhat correlated. This could also be a good indicator to determine the subspecies of the crab, though its less clear than figure 1.

1.3.4 Scatterplot for rear width against carapace length

```
# plot the rear width against the carapace length based on gender
ggplot() +
  geom_point(data = myData[myData$sex == "F",], mapping = aes(x = RW, y = CL, color = 'Female')) +
  geom_point(data = myData[myData$sex == "M",], mapping = aes(x = RW, y = CL, color = 'Male')) +
  scale_color_manual(values=c("#EE4444", "#4444EE")) +
  labs(x = "Rear width (mm)", y = "Carapace length (mm)", title='Carapace length against Rear width bas
```

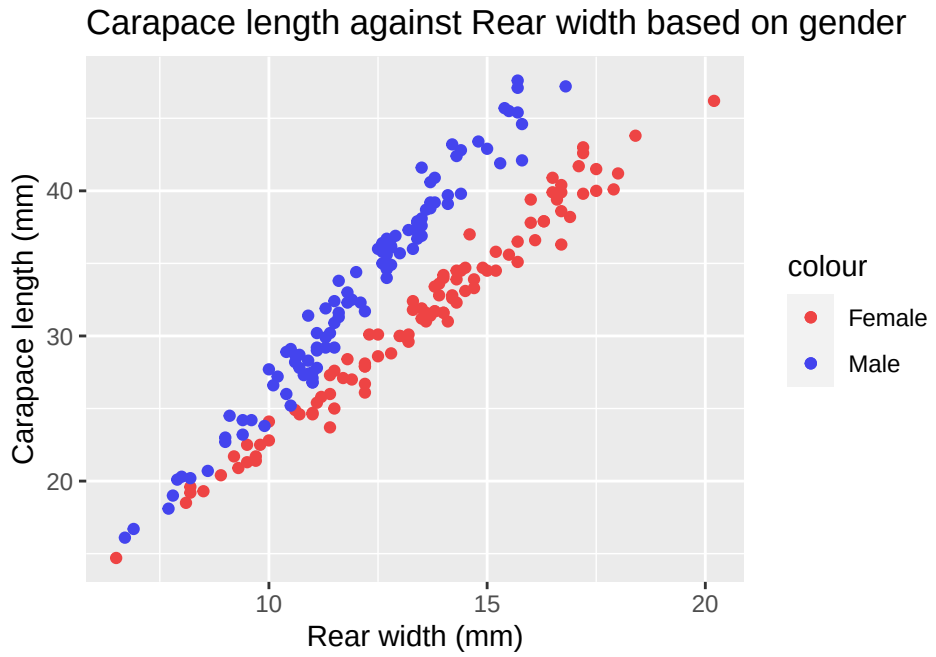


Figure 4: Spread of Carapace length against Rear width based on gender

This plot was created to further investigate the two groups of figure 2. Instead of coloring the datapoints blue and orange based on the color of the crab, the datapoints in this plot were colored according to the gender. Blue for male crabs and red for female crabs. 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.

1.3.5 Boxplot of carapace length for the genders

```
# sepperate the orange and blue crabs
orangecrabs <- myData[myData$sp == "O",]
bluecrabs <- myData[myData$sp == "B",]

ggplot(data=myData, mapping = aes(x=sex, y=CL)) +
  geom_boxplot(notch=TRUE, fill=c("#EE4444", "#4444EE")) +
  labs(x = "Gender", y = 'Carapace length (mm)', title='All crabs')
```

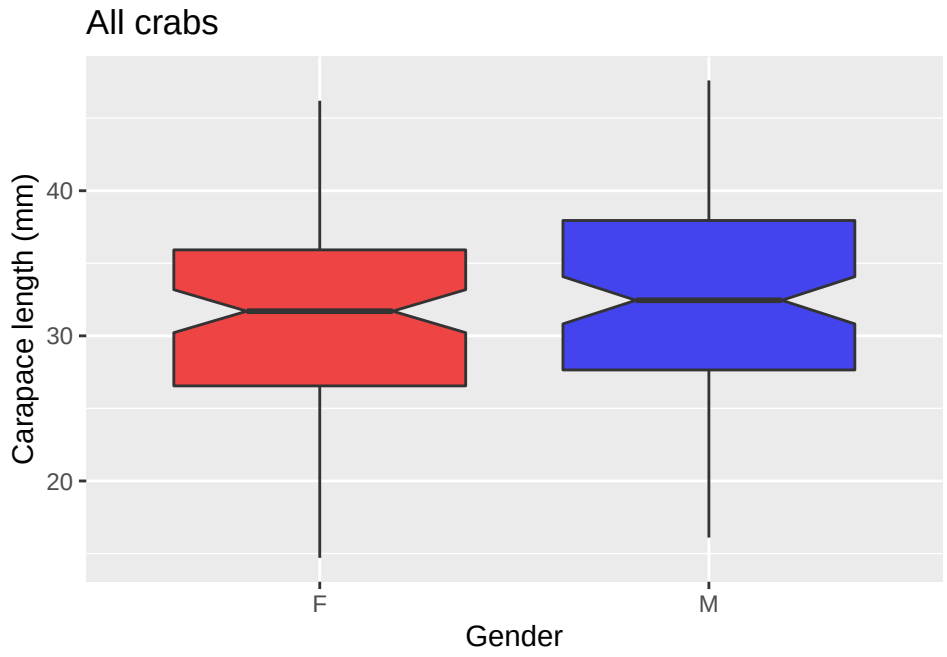


Figure 5: Distribution of Carapace length for all male and female crabs

This plot was created to further investigate the difference in carapace length for the genders instead of the color of the crab. 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.

1.3.6 Boxplots of the carapace length for each gender

```
require(gridExtra)

ocrabs <- ggplot(data=orangecrabs, mapping = aes(x=sex, y=CL)) +
  geom_boxplot(notch=TRUE, fill=c("#EE4444", "#4444EE")) +
  labs(x = "Gender", y = 'Carapace length (mm)', title='Orange crabs')

bcrabs <- ggplot(data=bluecrabs, mapping = aes(x=sex, y=CL)) +
  geom_boxplot(notch=TRUE, fill=c("#EE4444", "#4444EE")) +
  labs(x = "Gender", y = 'Carapace length (mm)', title='Blue crabs')

grid.arrange(bcrabs, ocrabs, ncol=2)
```

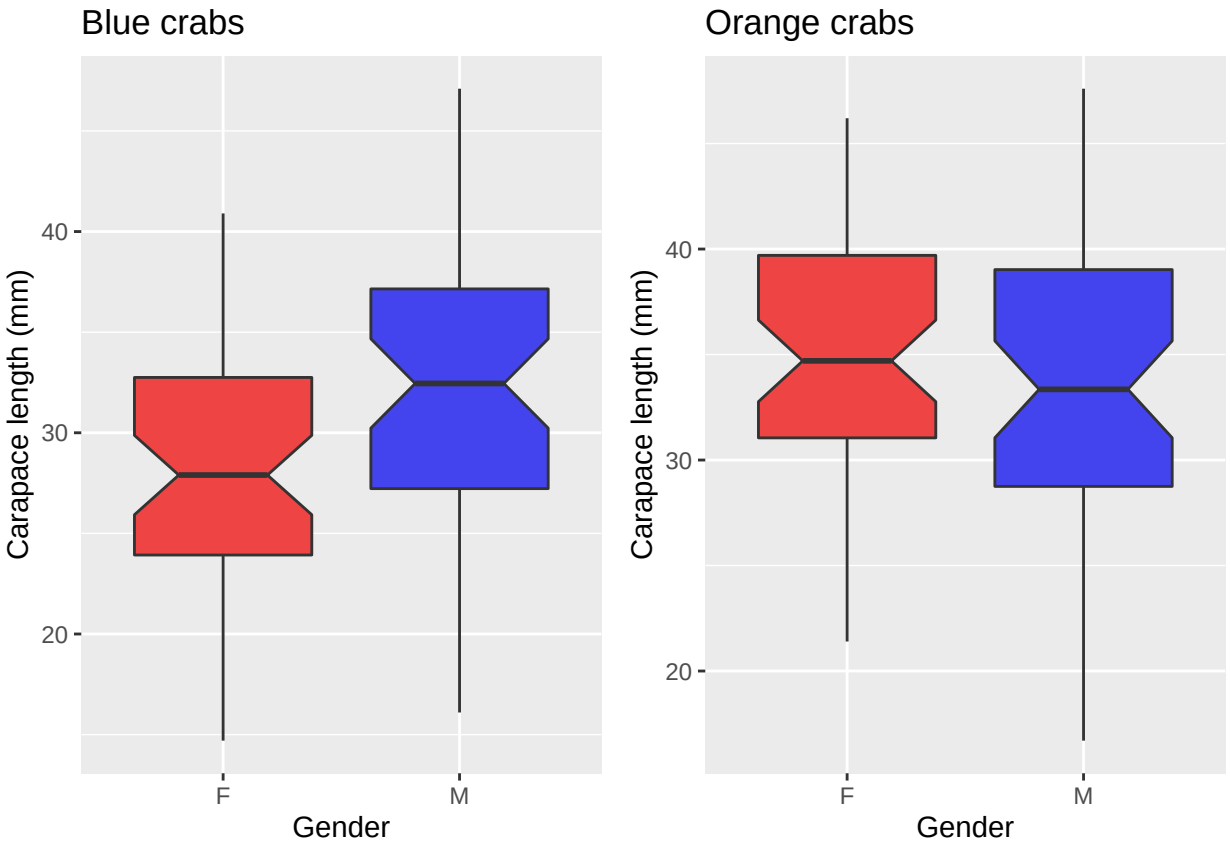


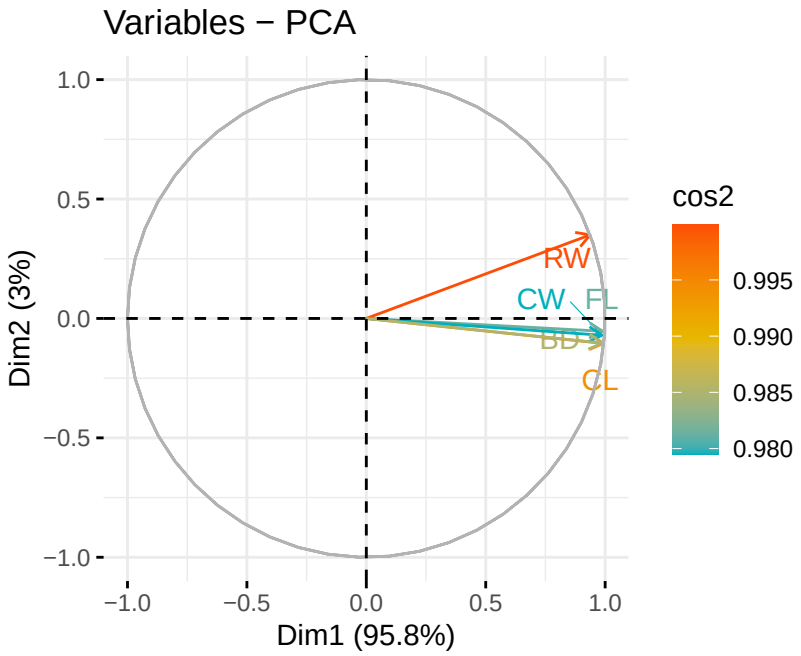
Figure 6: Distribution of Carapace length for the orange and the blue male and female crabs

To investigate further upon the difference in carapace length based on the gender, these two boxplots were created to show the difference in carapace length for each gender for each color. Figure 6 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. Figure 7 shows a bigger difference in the length of the carapace between male and female blue crabs. Here, the difference in 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.

1.3.7 PCA plot

```
res.pca <- PCA(myData[4:ncol(myData)], ncp = 5, graph = FALSE)

fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



This PCA plot was created to show the correlation between all variables. 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 that were used to discover the difference in carapace length distribution between the genders.

1.4 Machine learning

1.4.1 cleaning

Before machine learning can be used, the data must first be cleaned. 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.

```
clean_data <- myData[,c(2,4:ncol(myData), 1)]
write.csv(clean_data, "datafiles/cleanedData.csv", row.names = F, col.names = F)
```

1.4.2 WEKA

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.

```
algorithms <- read.csv("datafiles/ml.csv")
algorithms <- data.frame(algorithms)

ggplot(algorithms, aes(x = reorder(Algorithm, desc(Accuracy)), y = Accuracy, fill = Algorithm)) +
  geom_bar(stat="identity", alpha=.6, width=.4) +
  coord_flip() +
  xlab("") +
  theme_bw()
```

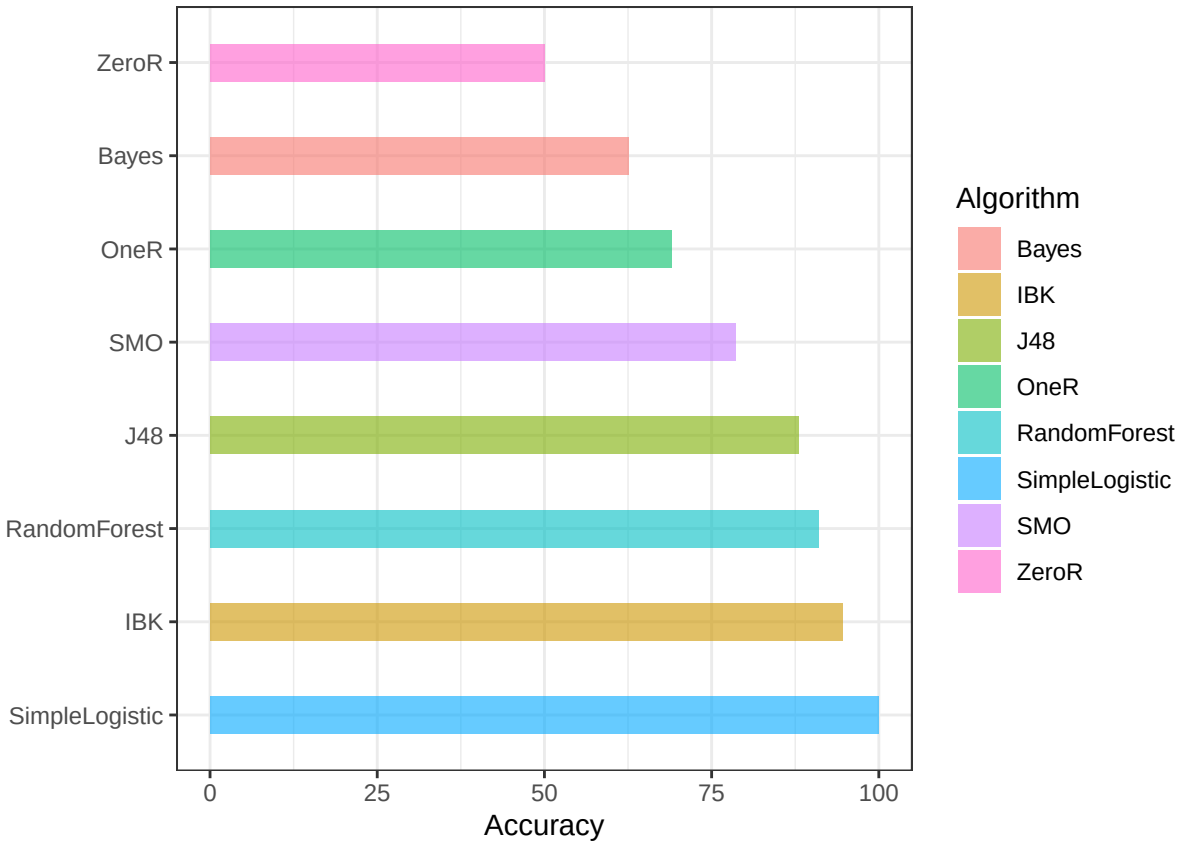
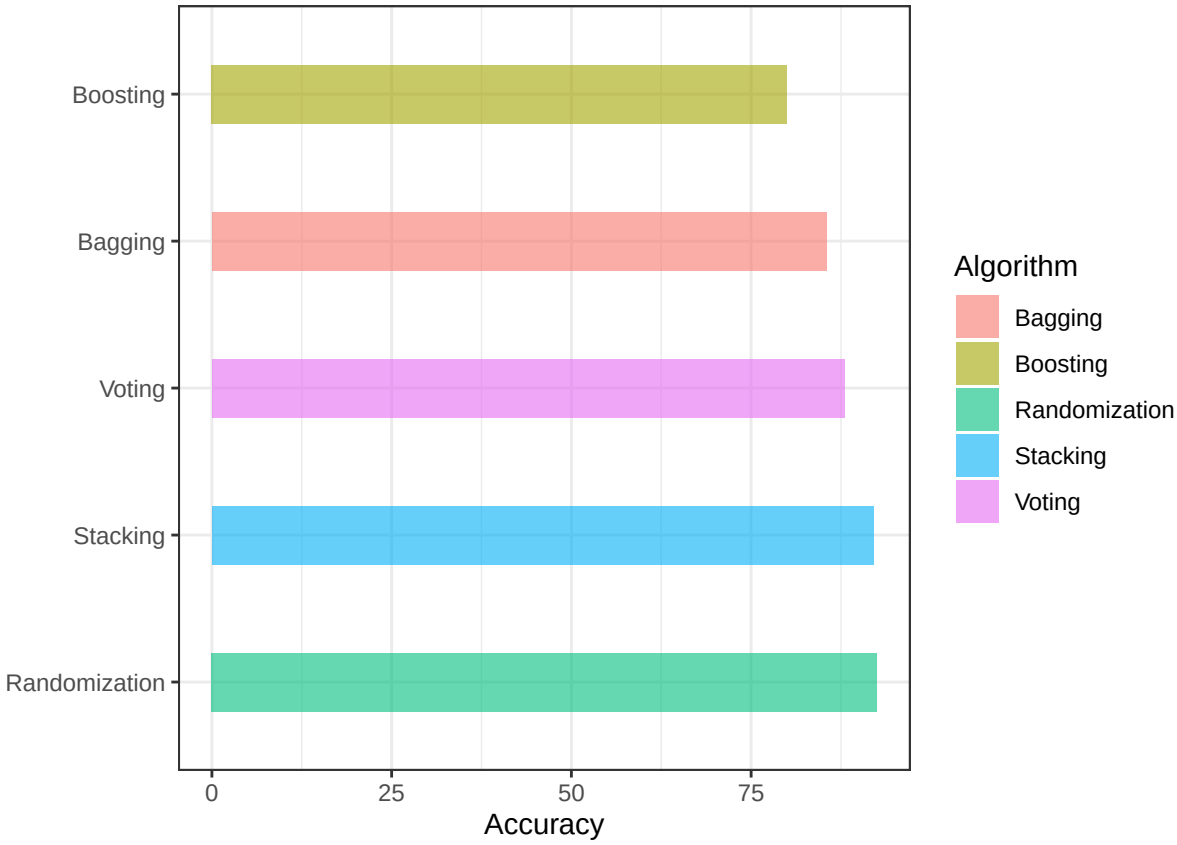


Figure 7: The accuracy of the machine learning algorithms ordered from low to high.

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.

```
metaalgorithms <- read.csv("datafiles/meta ml.csv")
metaalgorithms <- data.frame(metaalgorithms)

ggplot(metaalgorithms, aes(x = reorder(Algorithm, desc(Accuracy)), y = Accuracy, fill = Algorithm)) +
  geom_bar(stat="identity", alpha=.6, width=.4) +
  coord_flip() +
  xlab("") +
  theme_bw()
```



1.5 Discussion and future research

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.

Future research can be used to show more correlation between other measurements. It can be researched whether or not the gender of the crab could be predicted using these morphological measurements. Machine learning use could also be improved by expanding the dataset, or getting different amounts of blue or orange crabs, with different amounts of male and female crabs.