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 that required answering for this project was: "Can the species of a *L.variegatus* be determined based on some morphological measurements of its carapace".

The data that was delivered with this project contained five morphological measurements, these were measured in millimeters. To be capable of answering this question, the data was first explored and cleaned and visualized to be able to draw conclusions from the data. Then, multiple machine learning algorithms were tested along with some meta-learning algorithms to see if there would be an optimal algorithm for the answering of the research question.

The SimpleLogistic algorithm reached a accuracy of 100% This is due to the build-up of the data set, since its an evenly distributed data set. This was the best accuracy gained, but since biologists and zoologists will be using the program a tree would be preferred but since that would mean a decrees in accuracy of 12% a SimpleLogistic algorithm was chosen with a strong ReadMe.

2 Introduction

The rock crab *L. variegatus*, has been recorded to occur on a number of southern Pacific islands, along the western coast of South America, and the coasts of Australia south of the Tropic of Capricorn. Mahon.

By 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*. These colour forms were previously regarded as morphs of *L. variegatus*..

In an attempt to resolve this issue of identification and possible misidentification of *L.variegatus* a morphological study of the Western Australian species was undertaken.

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*[@Crabdata]. 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

For this project research has been done to be able to investigate and predict whether or not 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 and meta-learners have been tested and checked for their accuracy. The data was also used to be able to visualize every aspect of the *L.variegatus* its morphological measurements, to be able to answer the research question

3.1 Materials

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

For the research of this project, the original publication by Campbell, N.A. and Mahon, R.J. (1974) "A multivariate study of variation in two species of rock crab of genus" was used. Along side the original data the website for crabs was used to investigate why crabs are important for the environment.[@CrabImportance]

Based upon the literary information the research question was formed, this after rechecking the data set and finding out what would be an interesting and possibly important research question.

For the data and the visualizations certain libraries have been used, see table one for these libraries

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
forcats	0.5.1
dplyr	1.0.7
ggcorrplot	0.1.3
gridExtra	2.3
grid	4.0.5
ggpubr	0.4.0

The packages were used to be able to manipulate and visualize the data as well as load in data that was gathered from the WEKA program. Furthermore, one column of the data set was deleted, the index column since there was no apparent use for this column.

3.1.1 Java wrapper

To be able to run this project via the command line, a wrapper was created. This wrapper was made in java. It can classify instances given by the user using the SimpleLogistics model.

3.1.2 Github repositories

Two github repositories were used for this project. A repository for the report and log with the datafiles, and a repository for the java wrapper.

Link to the report and log repository: https://github.com/IJsbarnd/CrabWrapper

Link to the java wrapper repository: https://github.com/IJsbarnd/thema9_2

3.2 Methods

Before the data could be worked with, it had to be examined and its quality needed to be checked, there were no missing values and only one column that did not seem useful, so only one correction had to be made in order to understand the data better.

There are methods by which someone can determine the species of a *L.variegatus* based on some morphological measurements of its carapace. This project tries to see if these methods can be improved upon and if a better method can be developed

The goal is to be able to determine the the species of a *L.variegatus* based on some morphological measurements of its carapace. So researchers and biologist have a better understanding of the crab they are dealing with. The parameters used for this can be found in table two

Table 2: Codebook of the crab data set

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)

There were no further calculations needed for this project. The used script was CrabProject.Rmd, the data files can be found within the same area as the project file.

4 Results

The results from this research project are placed within two different categories, there being the exploratory data analysis and the machine learning algorithm analysis. This choice has been made since both sections have different outcomes and goals, but both are equally important.

4.1 Exploratory data analysis

The goal of the exploratory data analusis was to be able to give an overview of the data, which included a look at the original data and figuring out whether or not this data is suitable for machine learning and if not, then making it suitable for machine learning. Another goal here was to be able to check if the research question was capable of being answered.

Table three shows the five number summary of all the numerical attributes within the crab data set, Since the instances within this data set differ from each other but do seem correlated to each other the mean and medians are somewhat close together. There are no missing values present since they are absent from the data set.

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

4.2 Cleaning the data

The data in its current form, is not ready for the use of a machine learning algorithm, this is due to the Index column, some algorithms will over fit their model by making use of this column, so this column needs to be deleted to be able to verify the quality of the data gathered. The species columns was also moved to the last column so it would be used as the class attribute.

4.3 Visualisations

To gain a better understanding of the correlation between the attributes in the data set, some visualizations were made.

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```

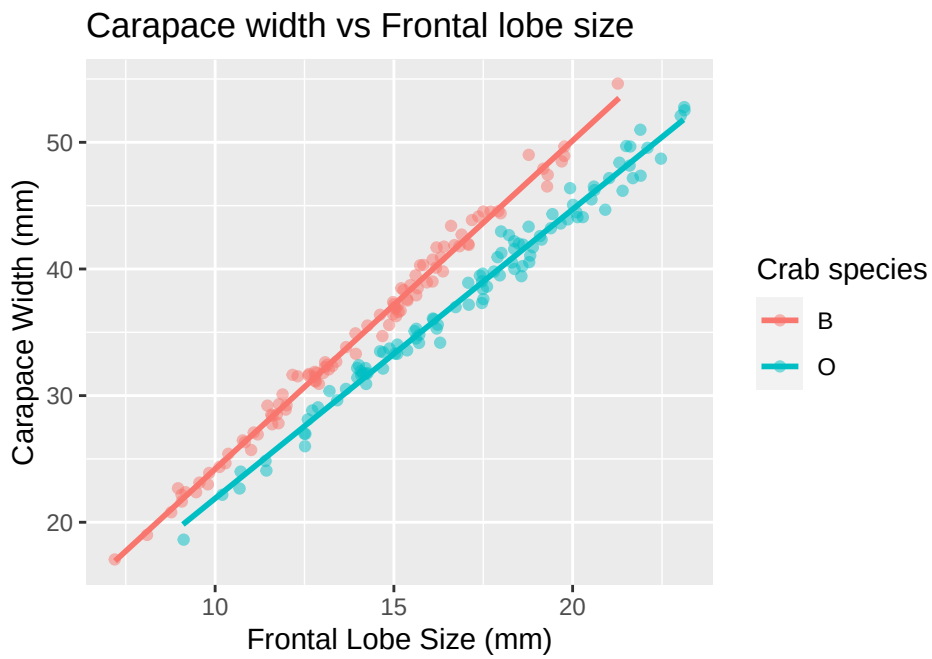


Figure 1: Spread of Front lobe size against Carapace width based on color

The image above shows the spread of front lobe size against Carapace width based on color, this visualization has been made to be able to see if there was any correlation between these two attributes. The image shows that the blue crabs have wider carapaces on average and shorter frontal lobes, this could be a genetic trait for the color of this crab and therefore this could be a good indicator to determine the subspecies of the crab by the use of a machine learning algorithm.

Since there were only 200 crabs in this study this could of course be a coincidence, but for the continuation of this report the assumption of data correctness due to differences in species has been accepted.

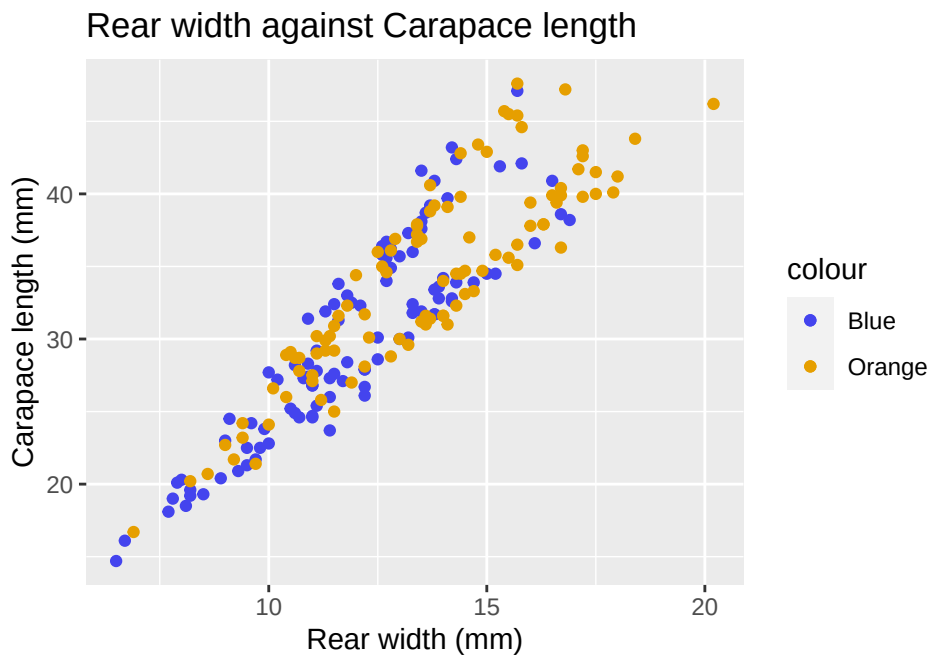


Figure 2: Spread of Rear width against Carapace length based on color

Similar to the previous plot, this data also seems to show some correlation with each other, this makes the question "is all the data correlated" rise, this will have to be investigated if the data continues to seem correlated. In the plot above the differences between Rear width against Carapace length are visible for both the different crabs, this being the blue group and the orange group.

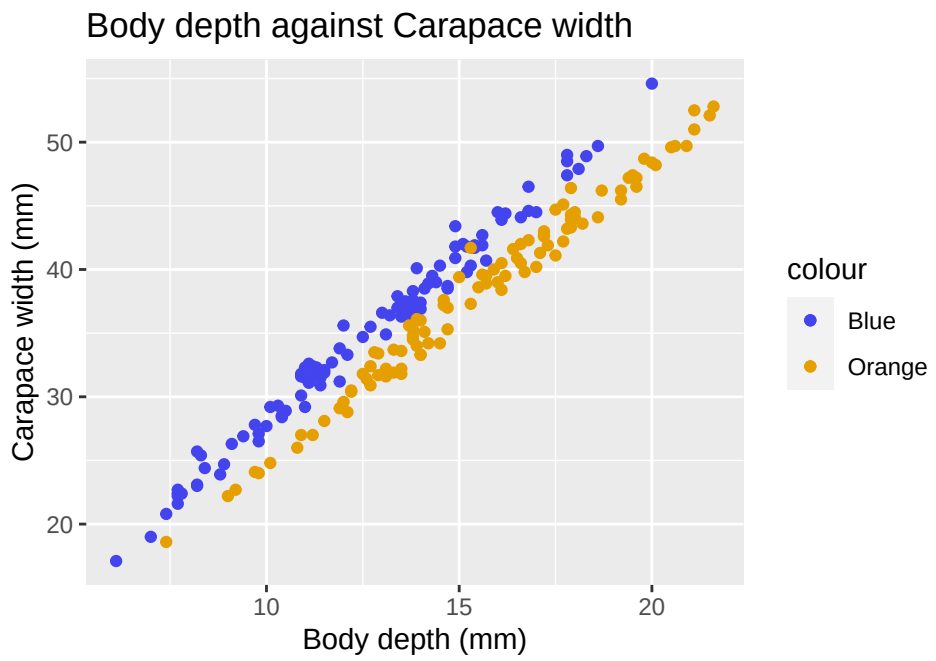


Figure 3: Spread of Body depth against Carapace width based on color

This plot again shows that the blue crabs on average seem to have a wider carapaces, but the orange crabs tend to have deeper bodies. The attributes seem to very correlated with each other and this again supports the theory that there may be multiple morphological attributes that will help determine the color of the crab. However the attributes above seem to be less clear in this then the attributes within figure one.

As shown inside the figures one and three, the metrics of the orange and the blue crabs appear to be in two different groups. This could indicate that the two species could be identifiable by different elements, including their carapace width. However, figure two indicates that the two groups of crabs appear to have some sort of difference even within their own respective group, this could indicate that gender plays a role in the sizes of the morphological measurements.

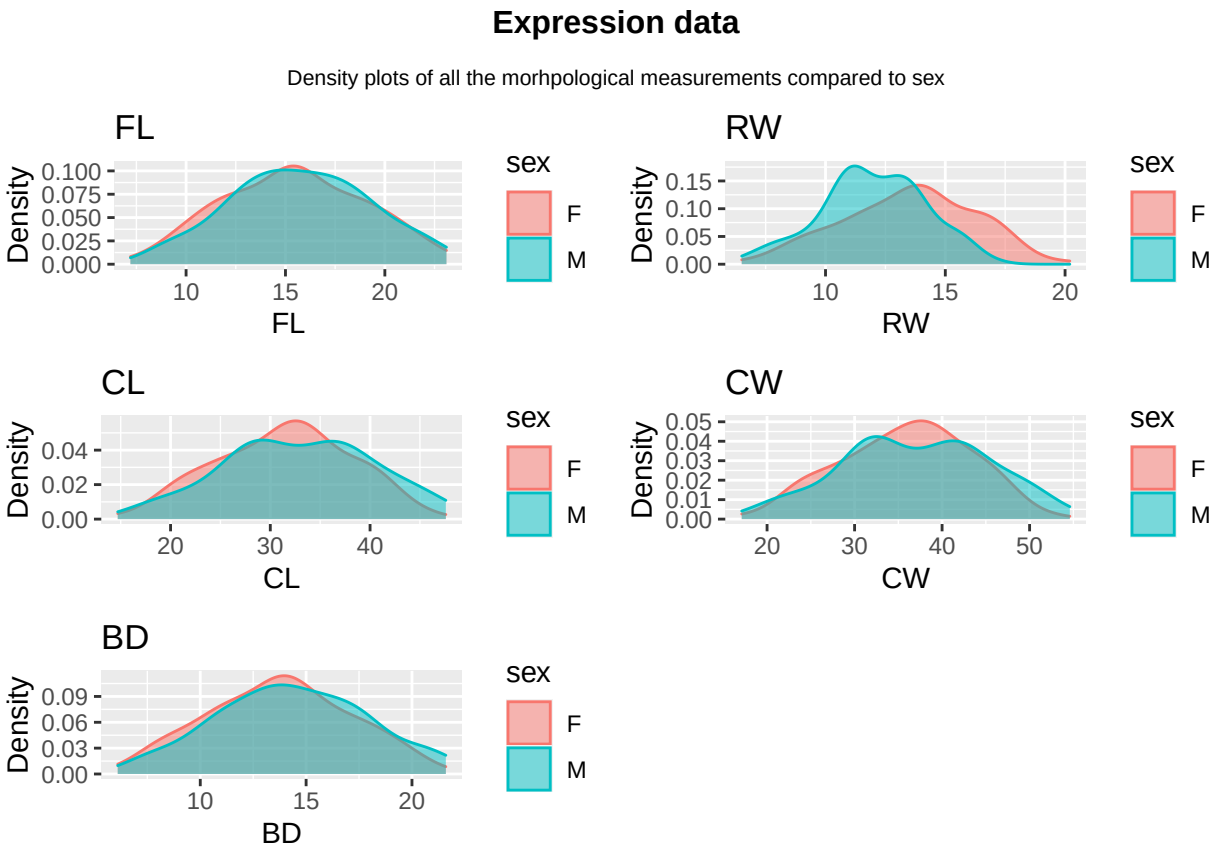


Figure 4: Checking if the gender makes a difference

As seen in the density distribution of figure four, there indeed seems to be a difference within the groups if they are divided by the male and females of the group.

In figure 4.1 the frontal lobe size in millimeters is plotted against the gender of the crabs, here it is clearly visible that there doesn't seem to be much difference in the frontal lobe size between the genders, the only difference that can be seen here appears to be in the the smaller measurements of the frontal lobe size, slightly indicating that the females might have a smaller frontal lobe size in general then males, this however cannot be confirmed from this density plot.

In figure 4.2 the Rear width in millimeters is plotted against the gender of the crabs, here it can be seen that the female crabs appear to have a bigger rear width then the male crabs, it could be that one colour species has a larger rear width, but the most plausible reason for this is that genetically speaking the female crab has a larger rear width.

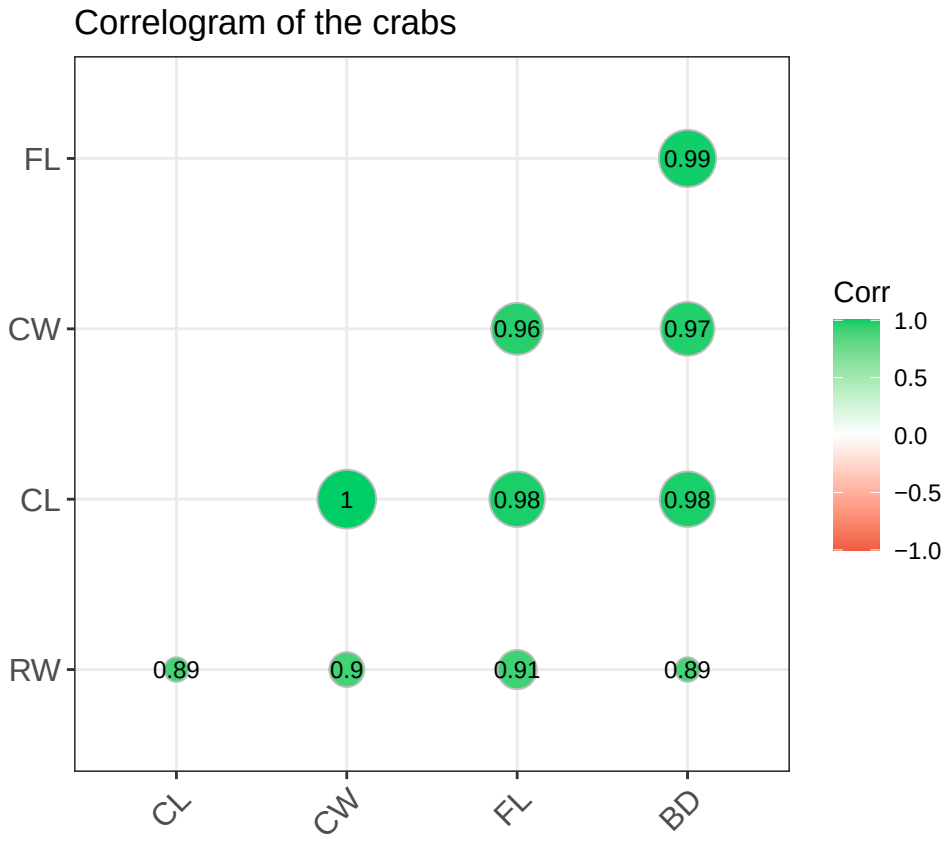
In figure 4.3 the Carapace length in millimeters is shown here against the gender of the crabs, here similar to figure 4.1 there do not seem to be many differences apart from the fact that more females appear to have a carapace length of around 30 á 35 millimeters in length, this however does not seem significant, and that a small portion of the male crabs seem to have a larger carapace length then the females, but since this doesn't seem like a mayority this could just be an anomaly.

In figure 4.4 the Carapace width is shown, measured in millimeters, plotted against the gender of the crabs similar to the figure 4.3 there do not seem to be a lot of differences, apart from the males where a small portion seem to have a larger carapace width then the females, this is however again a small portion so it may be insignificant.

In figure 4.5 the Body depth in millimeters is shown, this is plotted against the gender of the crabs, here the

distributions appear similar, making is seem like gender does not affect body depth what so ever.

One thing that is noticeable here is the distributions of the female crabs, these appear to be more of a normal distribution then the male crabs, which can be due to coincidence since it is a small data set, but it can also be that the body of a female crab is more normally distributed then the body of its male counterpart.



In the plot above the correlation between all the numerical elements of the data set are shown, this has been done to be able to answer the previously arisen question whether or not the data is correlated.

As seen in the plot above, the data inside the data set does indeed seem to correlate which supports the earlier findings, this could mean that there is more then one good attribute to be able to confidently support the investigation of the research question

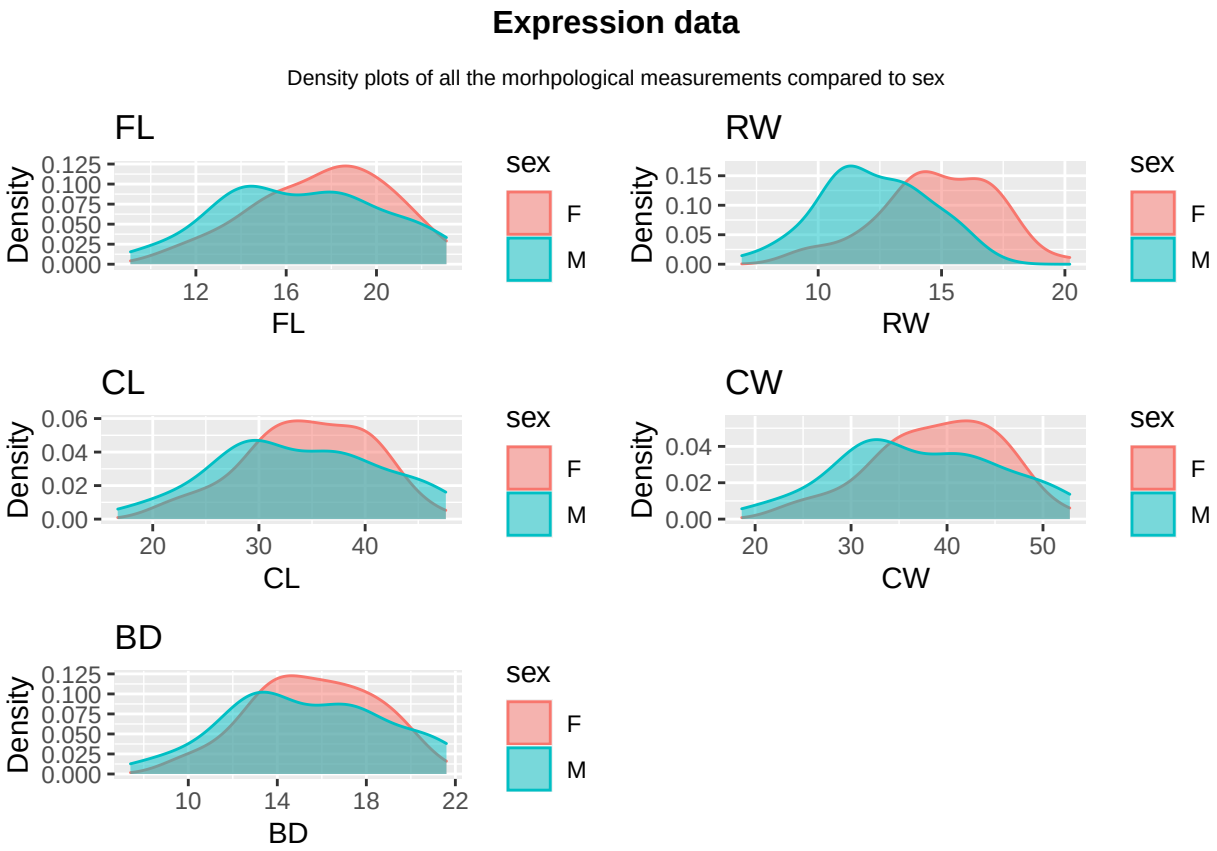
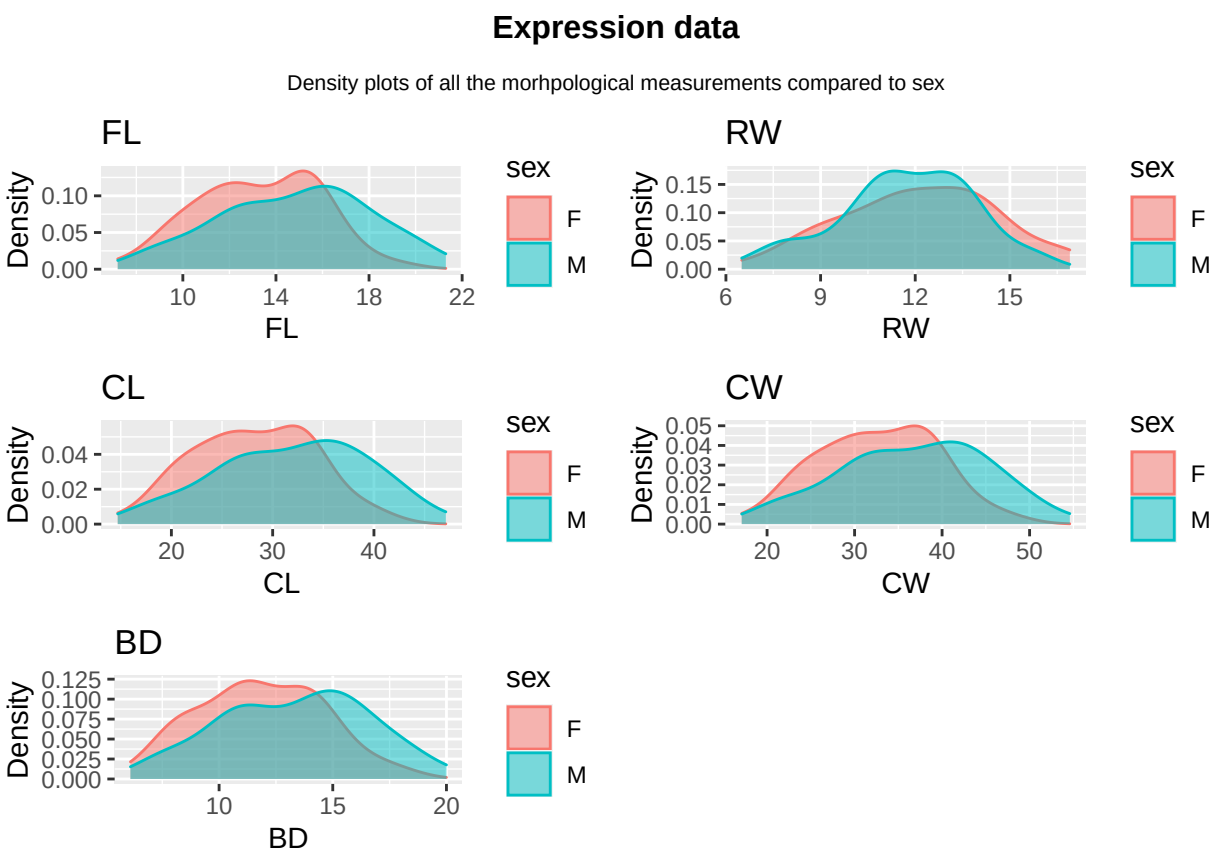


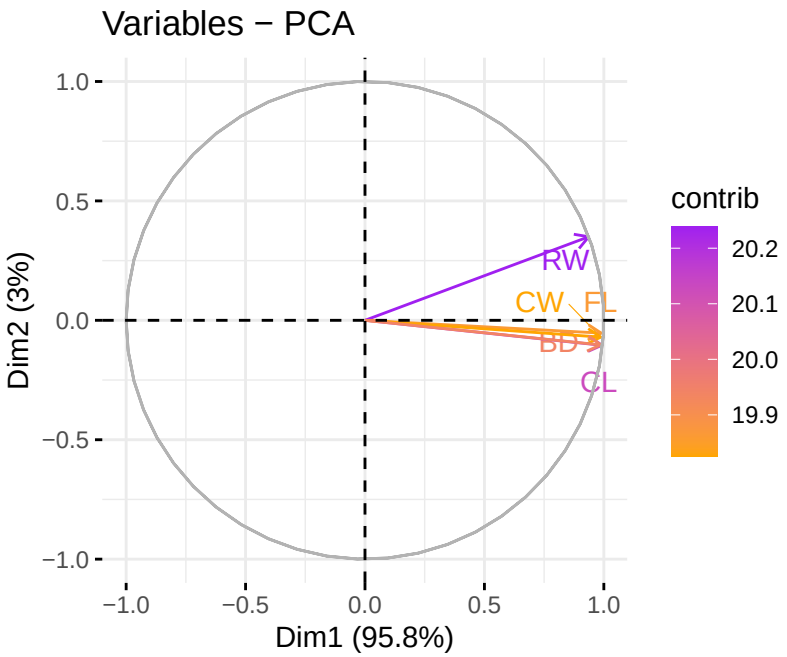
Figure 5: Checking if the gender makes a difference

As seen in figure six, the orange crabs their morphological measurements have been taken to be able to see if the sex matters when it comes to the colours of the crabs,



In figure six and seven the morphological variables of both color crabs have been put into a density plot to see if there is any correlation there, here it is however clearly visible these two separate colors differ in their measurements. As can be seen in figure six: There are no real differences with the morphological measurements for the orange crabs, apart from the rear width, and here the females are noted to have a larger rear width, apart from that some males appear to be on the lager side of the spectrum in the density plot, but this does not seem significant enough that it is to be mentioned. The orange crabs also seem to have somewhat of a normal distribution or bell curve when it comes to their measurements, this could indicate that they are either smaller or their morphological components are more formed to each other.

In figure seven the blue crabs are depicted, here differing from figure six, the male crabs of the blue species seem to be bigger in almost all aspects of the morphological components apart from rear width, this would indicate that for this species the male variant is bigger then the female variant, how ever, this is a data set with only 100 crabs per species and for a correct distribution a much larger sample size should be taken. For the carapace length, it seems that for this species alone it would be very good to be determined if the crab was of significant size, which species it would belong to. To see if this would work with a machine learning algorithm, further testing will need to be done.



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.

4.3.1 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 data set 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 bar plot below.

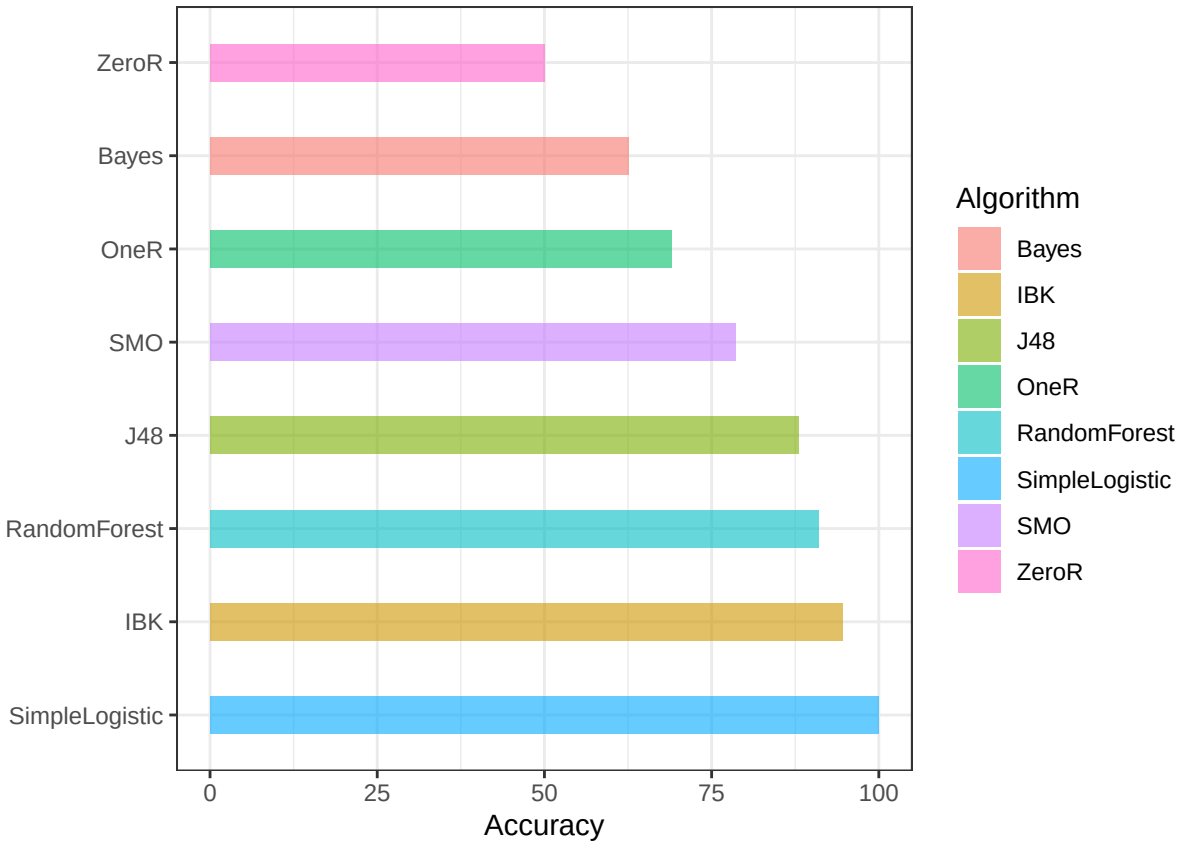


Figure 6: The accuracy of the machine learning algorithms ordered from low to high.

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.

4.3.2 Metalearners

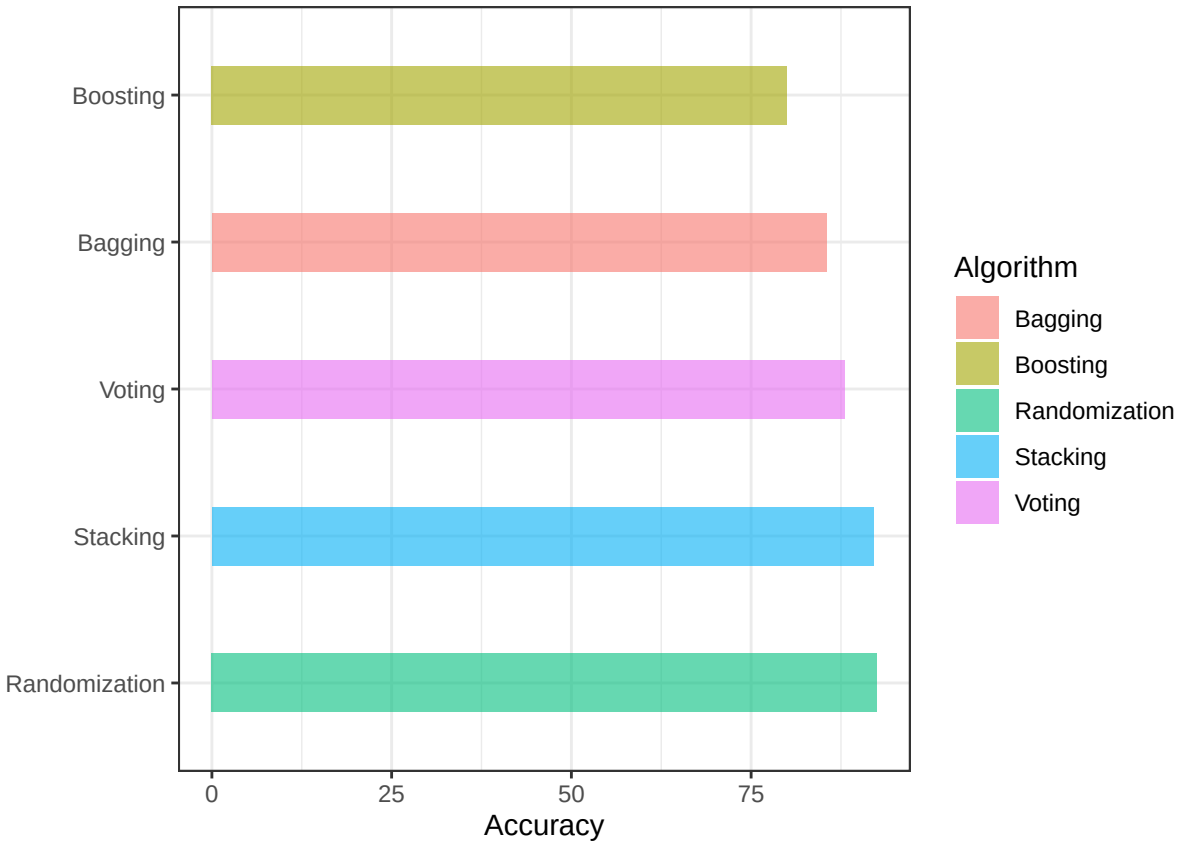


Figure 7: The accuracy of the meta machine learning algorithms ordered from low to high.

To get the best model for predicting the color of the crab, multiple metalearners using this SimpleLogistic method were tested. The tested meta learners are boosting, bagging, voting, stacking and randomization. These meta learners all use SimpleLogistics as the learners. As shown in figure 10, randomization got the best score out of the meta learners. This meta learner has a lower accuracy than the base SimpleLogistic model. It uses multiple SimpleLogistic learners with different settings. This means that the settings used in the base SimpleLogistics model are the best.

4.3.3 ROC-curves

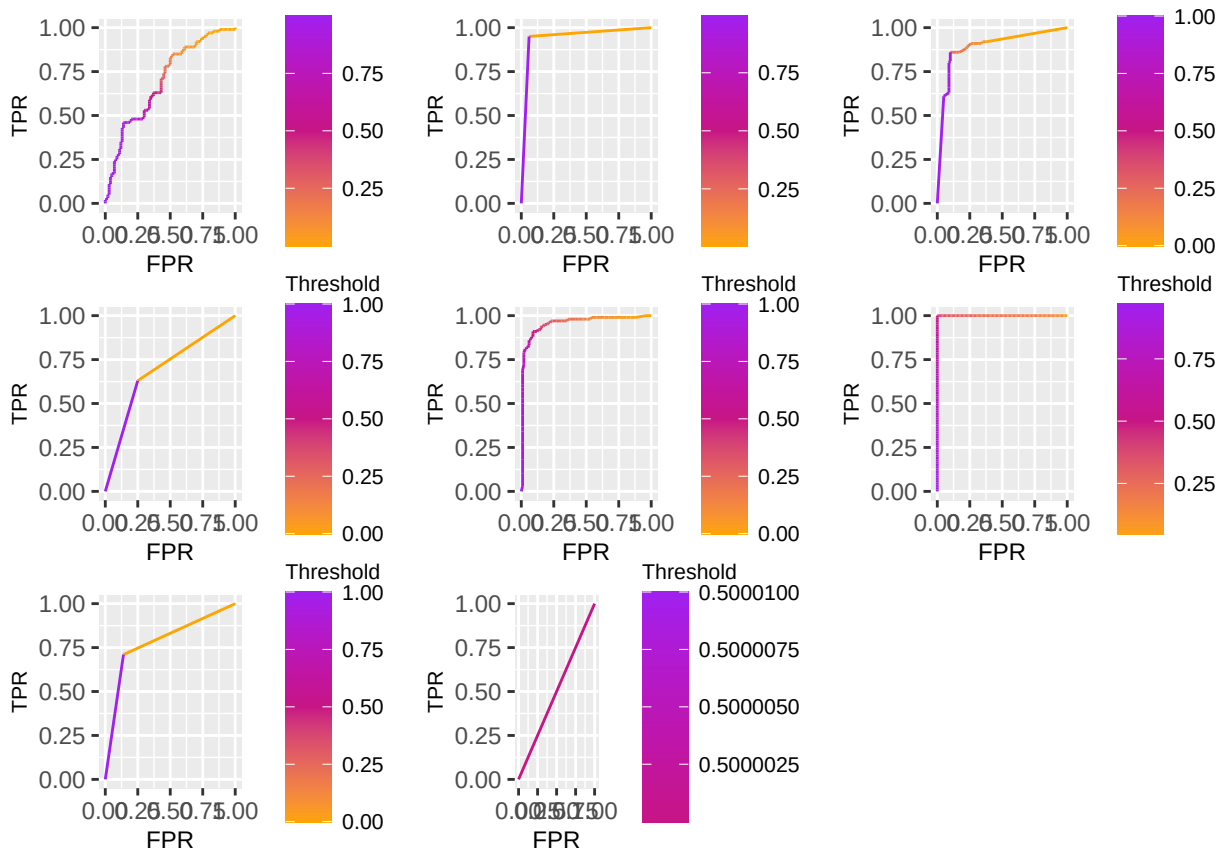


Figure 8: The false positive rate against the true positive rate of the machine learning algorithms

To get a better view of how the machine learning algorithms did, the ROC-curves of all of the algorithms were generated and shown in figure 11. The order of the plots is as follows: figure 11.1 Naive Bayes, figure 11.2 K-nearest neighbour, figure 11.3 J48 tree, figure 11.4 OneR, figure 11.5 Randomforrest, figure 11.6 SimpleLogistics, figure 11.7 SMO, figure 11.8 ZeroR. These plots also color the curve to match the threshold. As shown in figure 11.6, the ROC-curve of the SimpleLogistics model, it has an area under the curve of exactly 1. This means that this algorithm has a 100% accuracy.

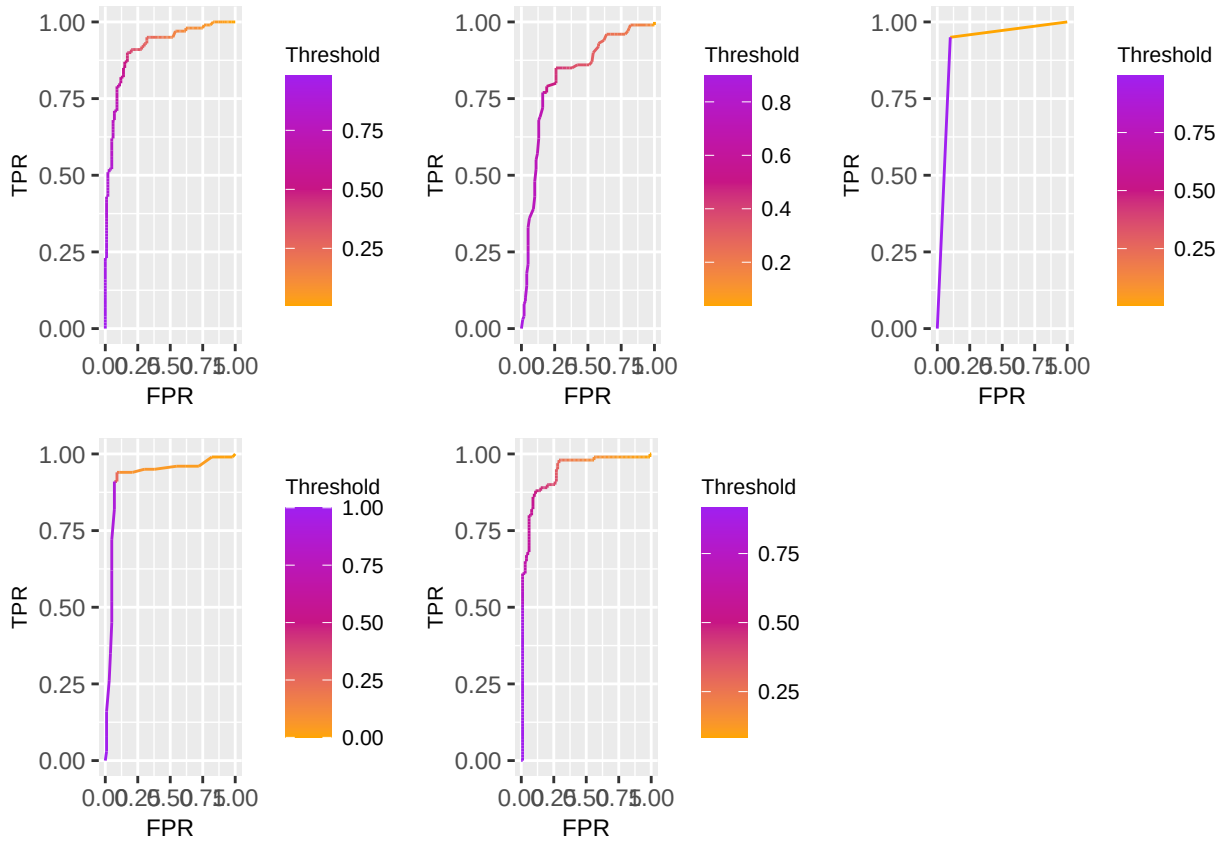


Figure 9: The metalearners: Stacking, Boosting, Bagging, Voting and Randomization

To give a better look at the meta learners, the ROC-curves of the meta learners were also generated, see figure 12. The order of these graphs is bagging, boosting, randomization, stacking and voting. These graphs show high areas under the curves, but not as high as the base SimpleLogistic model. This means that the base SimpleLogistics model can be better used to predict the color of the crab.

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 Future work proposal

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.

6.1 Project proposal for minor BIN

This project could be continued in a minor. The machine learning could be upgraded, or the wrapper could be improved. For the minor Application Design, a web app could be made to be able to classify new instances online. This would improve the project by making it more accessible for the user, and easier to understand. For the minor HighThroughput High Preformance Biocomputing, the machine learning part could be improved with new methods we will be learning there.

7 Sources

7.1 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.