Results_and_Discussion

Margriet van der Molen

2023-09-29

Results

From the Exploratory Data Analysis (EDA) came a few interesting findings. For the research question it is necessary that the data has about the same number of female and male penguins. Figure one shows that this was indeed the case. The difference to the number of Chinstrap penguins to the other species is stark. This means that there is fewer instances of Chinstrap penguins compared to the other species.

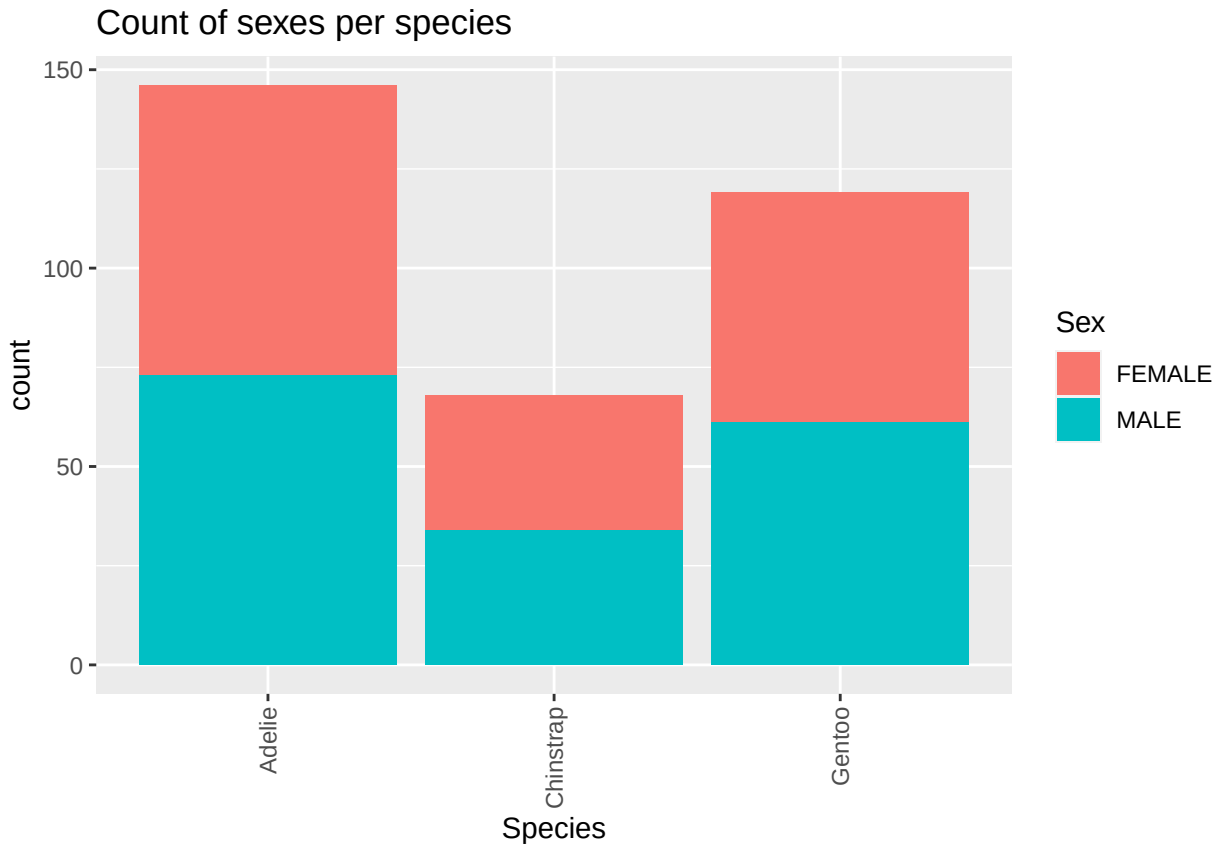

Figure 1: Number of penguins in the study per species. The bars are divided per sex where the number of female penguins are given a pink colour and the number of male penguins are shown in blue. Within the species, there is an equal number of male to female penguins.

To start, a few boxplots were created to get a sense of the distribution. These boxplots were grouped into the species and the sex.

When looking at figure 2, it is shown that Chinstrap and Gentoo penguins are more prone to sexual dimorphism in terms of the flipper length attribute than the Adelie penguins, though even the Adelie penguins appear to

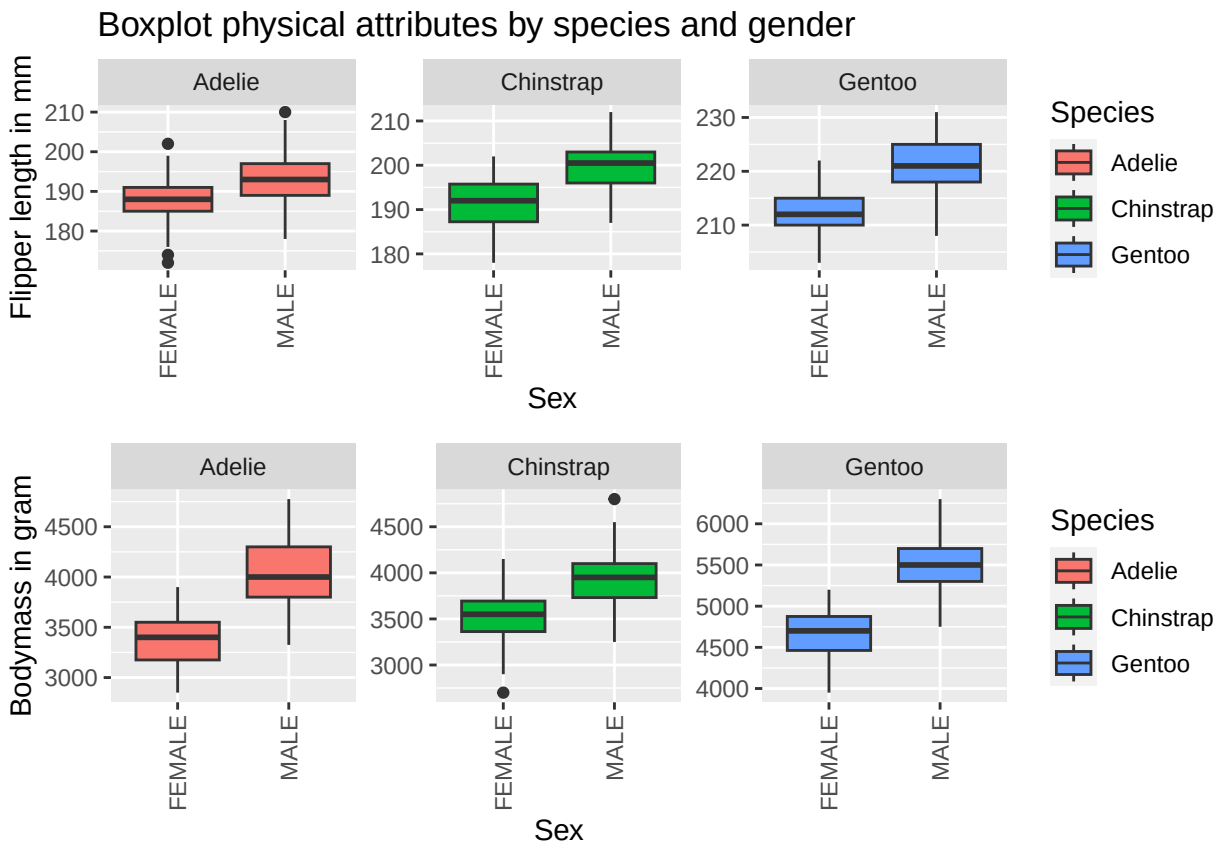

Figure 2: Boxplots of the flipper length attribute and body mass attribute. For flipper length Adelie penguins have a few outliers. For body mass Chinstraps have small outliers.

have a difference in body mass. With the information of the boxplot shown in figure 2, a two-way Anova was done.

Table 1: Table 2: A table showing the P-values from the two way ANOVA

	Flipper.Length..mm.	Body.Mass..g.
P value two-paired Anova Sex	6.278e-126	1.54e-123
P value two-paired Anova Species	2.461e-24	1.902e-57
P value Species * Sex	0.006314	0.0001973

From the test, it is shown that both flipper length and body mass, the sexes all significantly differ as do the species. In the bottom row, where the p value from the interaction between species and sex is, indicates this interaction has a significant effect on both flipper length and body mass.

Correlation was explored by using scatter plots and correlation scores, while distribution of the physical characteristics is shown on the diagonal.

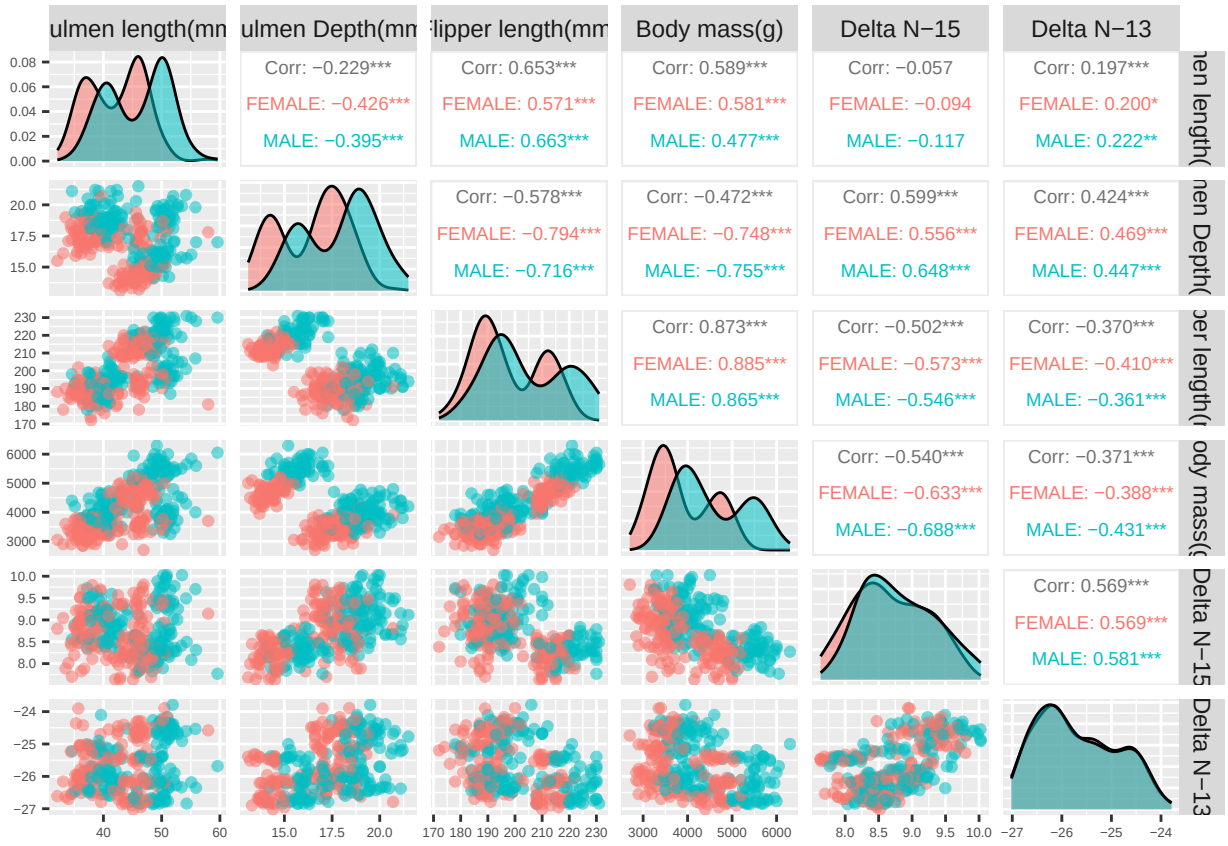

Figure 3: Scatter plots and correlation scores given of several attributes. Body mass and flipper length are said to have the strongest positive correlation while flipper length and culmen depth have the strongest negative correlation.The starts indicate the strength of the p value of the correlation test. 1 star is a p-value of 0.5, 2 stars is a value of 0.01, 3 stars is a p-value of 0.001.

Some scatter plots in figure 3 seem better divided than others. Body mass and flipper length in particular shows a strong positive correlation as the correlation score proves. In the plot that shows culmen depth and body mass, there's a notable division between two groups. The same occurs with flipper length and culmen depth.

In figure 3, the culmen length histogram appears to have two very clear peaks. This can also be applied to the culmen depth histogram. For flipper length, there's a peak at, approximately, 190 mm and another at 210 and 220 mm for both male and female penguins. When looking at body mass, female penguins tend to range between 2500 gram to around 5000 gram where as the male penguins range between 3500 gram and 6500 gram. The delta N-15 attribute peaks slightly at the values between 8 and 8.5 mille, and gradually declines in height. For the delta C-13 attribute, there is one peak at approximately -26 per mille of the plot, and after the peak it becomes more uniform in shape.

After determining what attributes may be useful, all previously mentioned algorithms were ran, taking into account the simplicity, and accuracy of these algorithms.

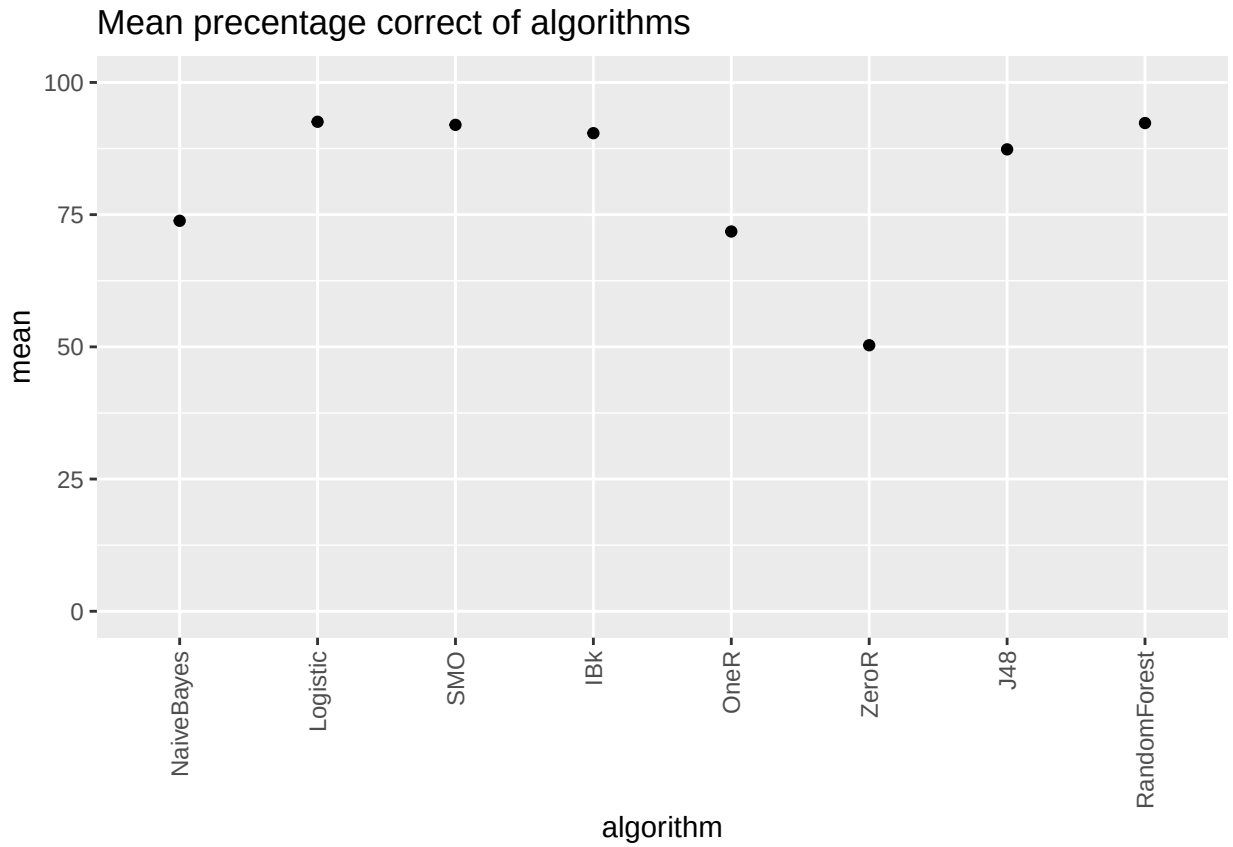

Figure 4: A figure showing the mean percentage correct of algorithms. Here the y-axis goes from 0 to a hundred 100%. The Logistic algorithm algorithm along with SMO and RandomForest are amongst the top performing algorithm while Zero-R, One-R, and NaiveBayes are amongst the lowest.

Figure 4 shows the mean of the accuracy value of the algorithms. All algorithms are out-performing Zero-R but the Logistic algorithm and the RandomForest make up the top two. Therefore, these two algorithms were chosen to focus and from the other algorithms, Zero-R, One-R, J48, and IBk were used for stacking with another J48 tree as the meta-learner.

The CVParameterSelection tool was used to determine that the iteration depth of the RandomForest algorithm should be sat at 45 and the iteration depth of the Logistic algorithm should be sat at 12. For the algorithms in the Stacking ensemble learner, the neighbours of the IBk tree was put to 11, and the J48 tree that is not the meta learner the variable 'minObj' was put to 10.

In figure 5, the finer details of the plot are hard to see as the values are much too close to each other. However, again, all algorithms are out-performing Zero-R. Stacking the simpler algorithms together has boosted their performance as they are now able to catch up to RandomForest and Logistic.

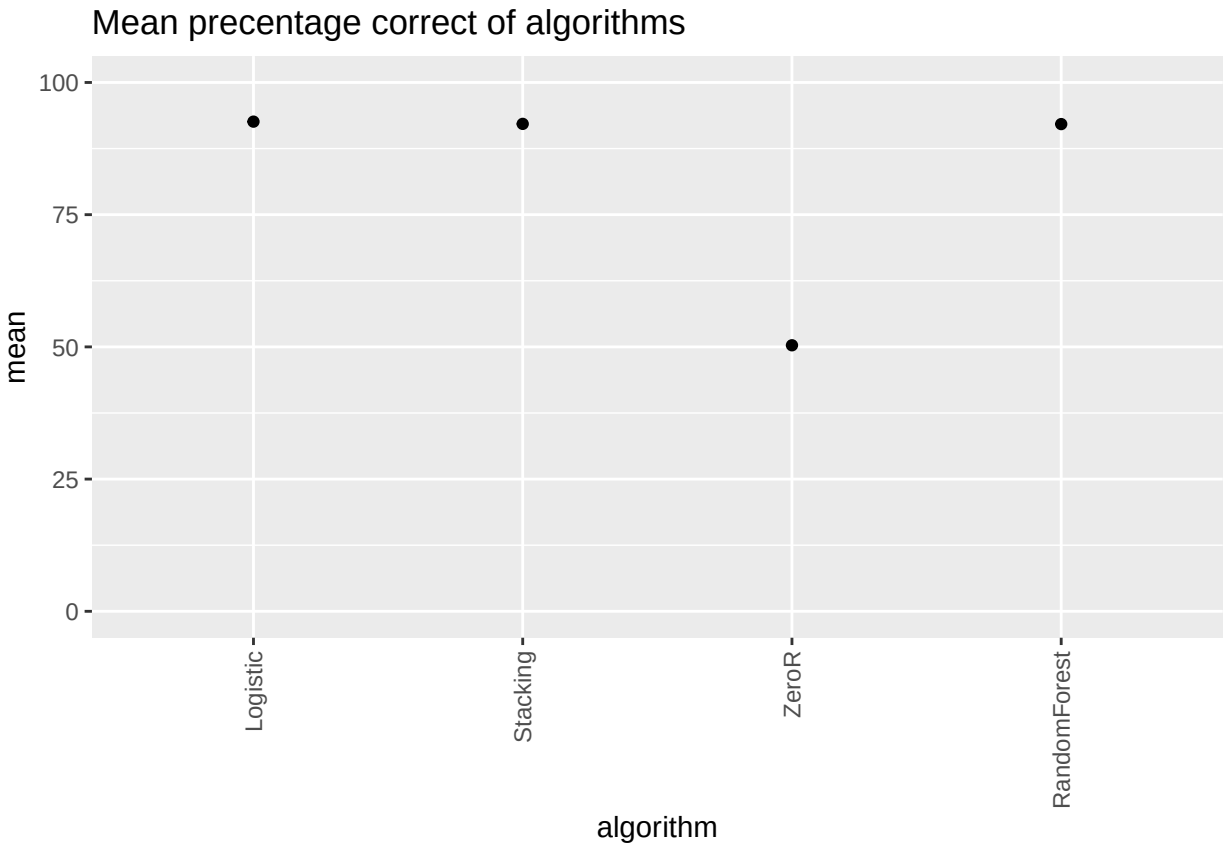

Figure 5: The mean percentage correct of the algorithms where the parameters were adjusted except for Zero-R which was ran with the default settings. Here, Logistic is the best performing though RandomForest and Stacking are not far behind.

For comparing RandomForest with the Logistic, both with their adjusted and before adjusted scores, a table was made.

	RandomForest	Logistic
Non adjusted	92.32	92.57
Adjusted	92.13	92.59

Discussion

In figure 1 the amount of Chinstraps was significantly lower than the other species. This means that the data set is slightly imbalanced, however it should not have a big effect as the research question's main concern is about the sex of the penguins where there are no imbalances within species.

While looking at the histograms of figure 3, it was noted that some of the figures have more than one peak which can be explained by the fact that there are several species in this dataset. The scatterplots in figure 2 also showed several 'groups' forming in the culmen depth, culmen length, flipper length, and body mass characteristic. It is likely that these peaks and groups are caused by the fact that different species have a different distributions of the physical characteristics. Something that is visible in the boxplot of figure 2 with the flipper length and body mass attributes. Here it is already visible that between species there is some difference and within species, there also seems to be a difference between the sexes. A two way anova proves there is a significant difference between the penguin sexes and species, along with that fact that the interaction between sex and species also has an effect on the two attributes. Looking back to the histograms of figure 3, there's now a possible explanation for the different peaks that are shown within the graph.

The scatter plots in figure 3 also showed correlation. In the model, correlation can be used to predict certain attributes. Flipper length and body mass have a strong positive correlation, if a penguin has more mass it will likely need broader flippers to move forward. Correlation combined with the findings of the anova could mean that flipper length and body mass could be attributes of interest for our model.

Conclusion

The EDA showed that different species have different averages across the attributes which causes different peaks when looked at in a histogram. Within the species, male penguins differ significantly from their female counterparts.

Flipper length, body mass, and culmen depth due to their correlation will be useful attributes for the model. Flipper length and body mass in particular as these attributes were shown to be significantly different within species.