A machine learning approach to predicting penguin sex

Margriet van der Molen

2023-09-29

Contents

Introduction

Most species of penguins appear not to prone to sexual dimporhism. This is to say, the male penguins don't look very different from the female penguins and even their behavior can be quite similar. This causes an issue when researching penguins as, sometimes, male penguins are identified to be female penguins or the other way around. These issues can be especially troubling when breeding programs might need to be put up. A problem that will become more prevalent as temperatures rise and certain penguin species' habitats disappear.

In an attempt to aid the researchers in distinguishing male and female penguins, machine learning could well be an option to predict the sex of a penguin. The predictions will be based on physical characteristics of the penguins, and these penguins will come from 3 different species. As such, this paper will focus on answering the following question:

"How accurately can a machine learning model predict the sex of a penguin when given the measurements of some physical attributes"

Materials and Methods

In this section, the materials and methods used to set up the tool are discussed.

Materials

This paper used Weka[1] (version 3.8) for the machine learning algorithms, it also used R (version 4.0.4) for graphing, logging, and reporting the findings. The data that was used to train the model with is, predictably, about penguins, and was originally used in a paper about ecological sexual dimorphism where they examined if environmental variability is associated with differences in the foraging niches of the male and female penguins[3]. The data contains information about the location, physical characteristics, sample data, and data about the egg, of 345 penguins on Antartica. While all penguins were from the Anvers region, they were not all from the same island. All the penguins were nesting and blood samples, measurements, and molecular sexing, was all done when they were at the one-egg stage.

Table 1: A table with the name of the attribute and the description next to it. When relevant, the unit is also givin in the description.

Attribute	Description
studyName	Abbreviation of the studyname
Sample Number	ID for the blood sample that was taken
Species	Species of penguin (Common name followed by latin name)
Region	Which region the penguin was found
Island	Which island the penguin was found
Stage	Stage in which the penguin lives
Individual ID	ID for the penguin pairs
Clutch Completion	Whether or not the nest was completed
Date Egg	Date at which the egg is observed
Culmen Length (mm)	Length of the culmen in millimeters
Culmen Depth (mm)	Depth of the culmen in millimeters
Flipper Length (mm)	Length of the flipper in millimeters
Body Mass (g)	Weight of the penguin in grams
Sex	Sex of the penguin
Delta 15 N (0/00)	isotopic nitrogen found in blood in mille
Delta 13 C (0/00)	isotopic carbon found in blood in mille
Comments	additional comments on the penguins

The delta C-13 and N-15 values were found using an elemental analyzer interfaced with an isotope ratio mass spectrometer at the Stable Isotope Facility, University of California, and calculated using the following equation:

$$\delta N15 or C13 = (\frac{\delta R_{sample}}{\delta R_{standard}} - 1) * 1000$$

Here the R_sample is the ratio of found C-13 to C - 12 or N-15 to N-14 in the sample. The R_standard is the standard ratio for international standards (Vienna PeeDee Belemnite for carbon, and atmospheric N2 (Air) for nitrogen).

Furthermore, the wrapper was made with Java (version 21), using the WEKA package(version 3.8). All code and information such as the logbook can be found within the github repository for this paper.

Methods

First an exploratory data analysis (EDA) was done to determine what attributes may be of value for the algorithm. This was done through the creation of box plots, bar plots, scatter plots, and two-way ANOVAs.

Accuracy and simplicity of algorithm were important qualities for the algorithm. Accuracy due to the nature of the research question, and simplicity to make sure the process of assigning the label is well understood. From these two, accuracy would be more important and trading in some simplicity for accuracy was acceptable. All algorithms were 10-fold cross validated.

In the first step the following algorithms were used to determine how well the algorithms ran with their default values: ZeroR, OneR, NaiveBayes, J48, IBk, SMO, Logistics, and RandomForest.

After determining the two top performing algorithms in the first step, the parameters of these algorithms were optimised using the CVParameterSelection tool. The iteration depth was adjusted. Stacking was used to allow the weaker performing algorithms to perform better, this algorithm was made up from ZeroR, OneR, J48, and IBk, where J48 was also used as the metalearner. The parameters of the J48 and IBk algorithm were also optimized using the previously mentioned tool.

The tool WrapperSubsetEval with the Logistic algorithm was used for attribute selection. By looking at what attributes were used most often when classifying a penguin into the categories, all attributes from 80% and above were chosen. With this a final assessment was done, an algorithm was chosen. A ROC plot was made to further analyze the results by first running the explorer with the Logistic algorithm and its mentioned parameters with the adjusted data set, the female penguin was set to the positive value, and finally the plot was made using R.

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.

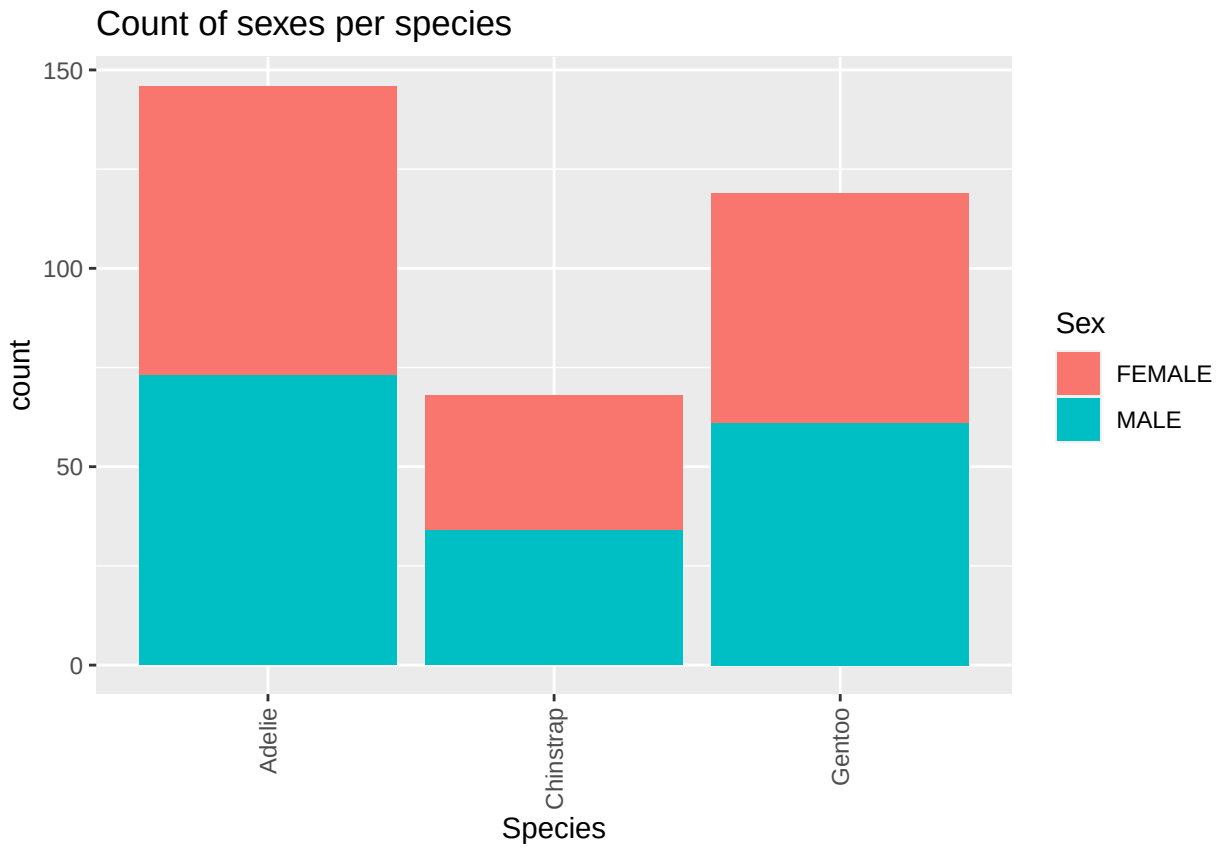


Figure 1: Number of penguins in the study per species. The bars are divided per sex where the number of female penguins are given a pink colour and the number of male penguins are shown in blue. Within the species, there is an equal number of male to female penguins.

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

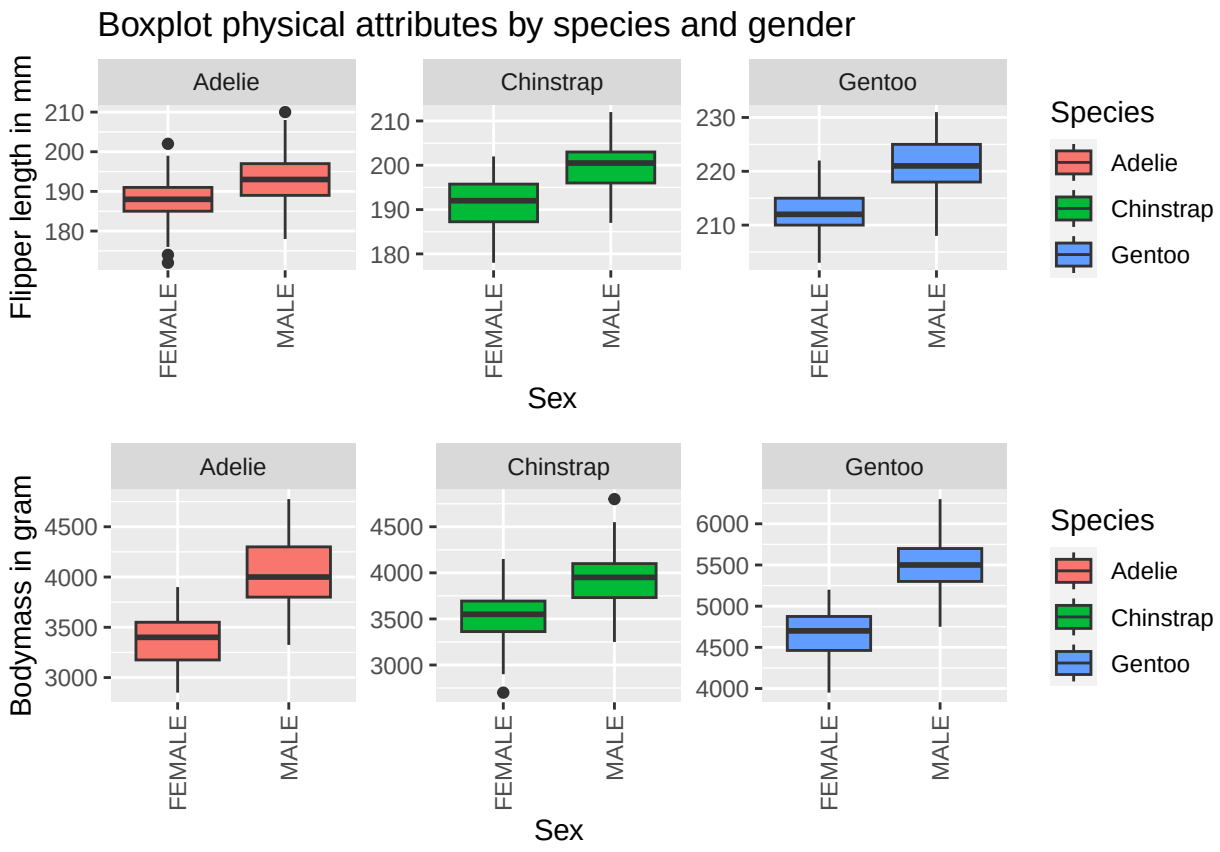


Figure 2: Boxplots of the flipper length attribute and body mass attribute. For flipper length Adelie penguins have a few outliers. For body mass Chinstraps have small outliers.

Table 2: A table with the p values for the two-way ANOVA. First with just flipper length and body mass, and lastly showing the interaction of the two on determining the sex.

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

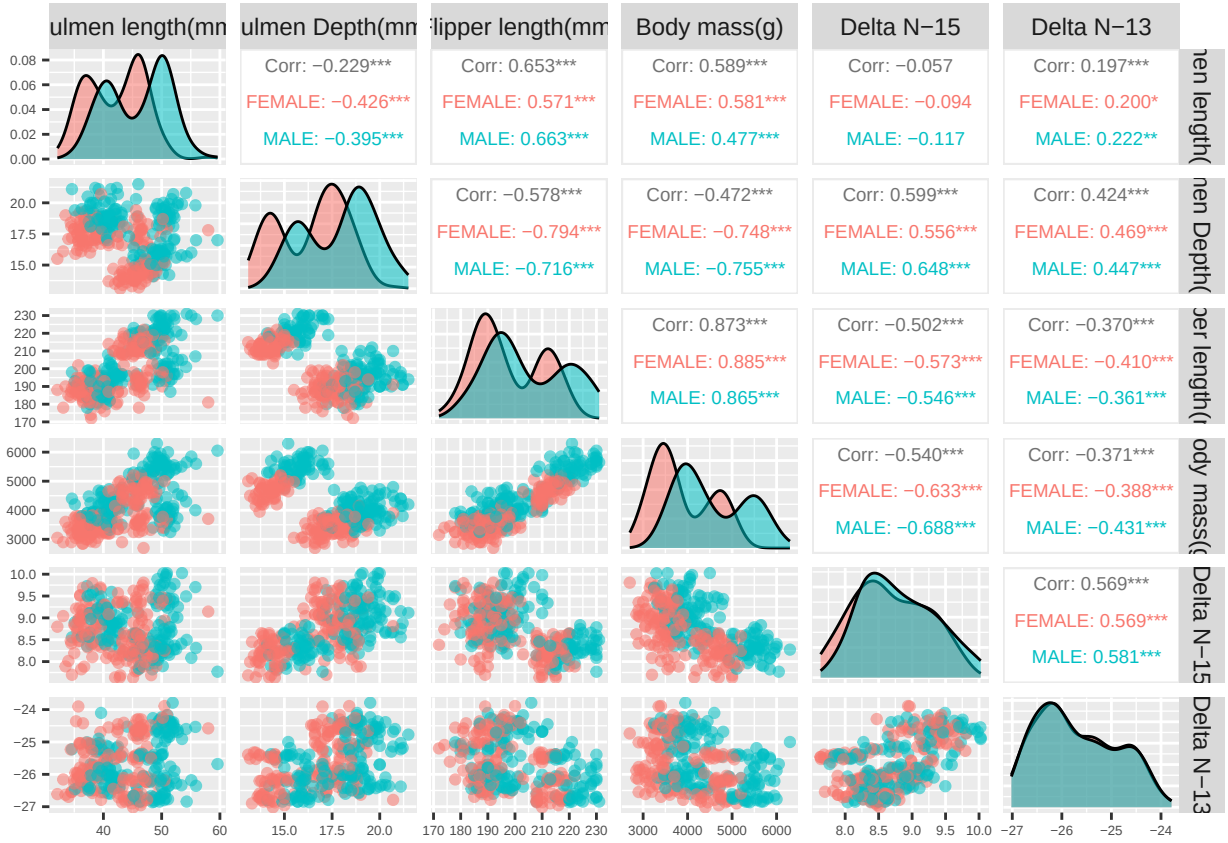


Figure 3: Scatter plots and correlation scores given of several attributes. Body mass and flipper length are said to have the strongest positive correlation while flipper length and culmen depth have the strongest negative correlation.The starts indicate the strength of the p value of the correlation 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

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

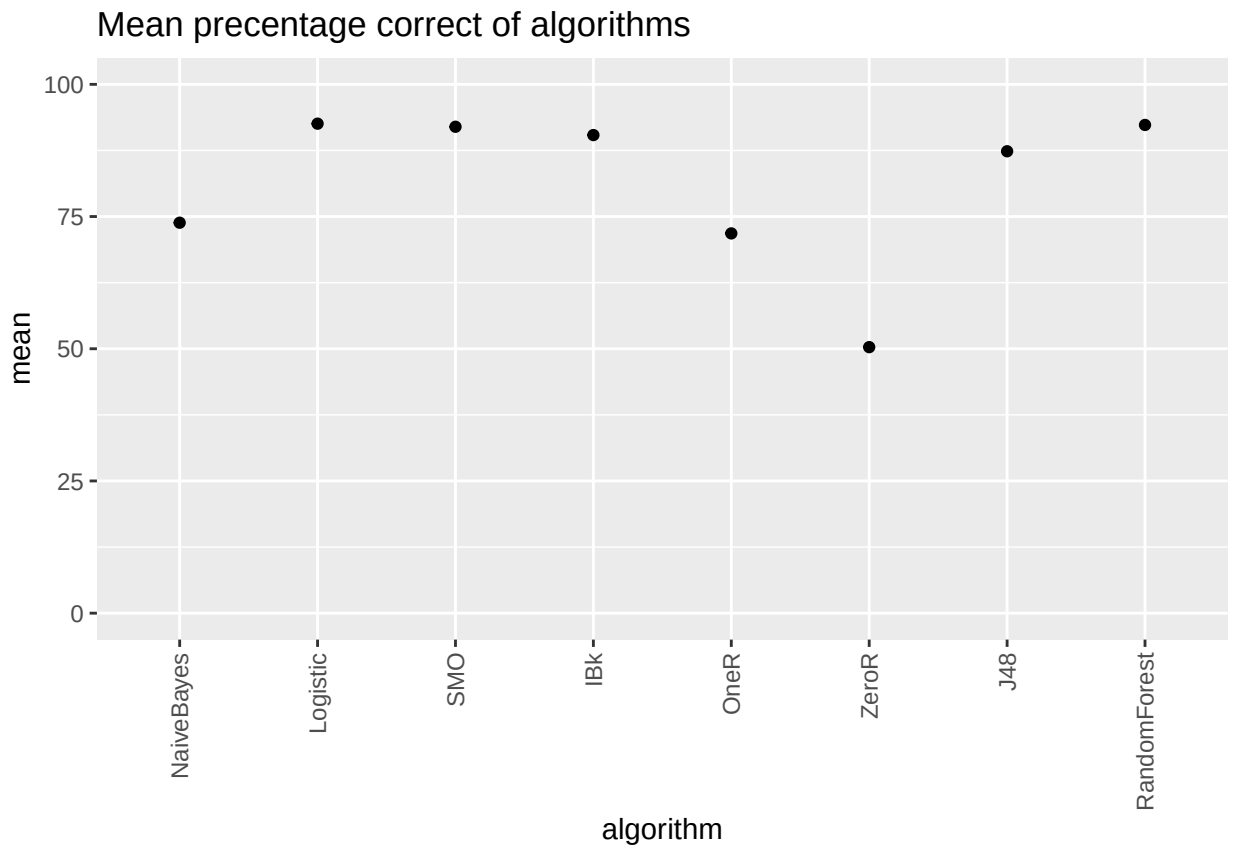


Figure 4: A figure showing the mean percentage correct of algorithms. Here the y-axis goes from 0 to a hundred 100%. The Logistic algorithm algorithm along with SMO and RandomForest are amongst the top performing algorithm while Zero-R, One-R, and NaiveBayes are amongst the lowest.

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.

To get a clearer view on the scores a table was made that compared the top two performing algorithms, RandomForest and Logistic, with the scores before and after adjusting the parameters.

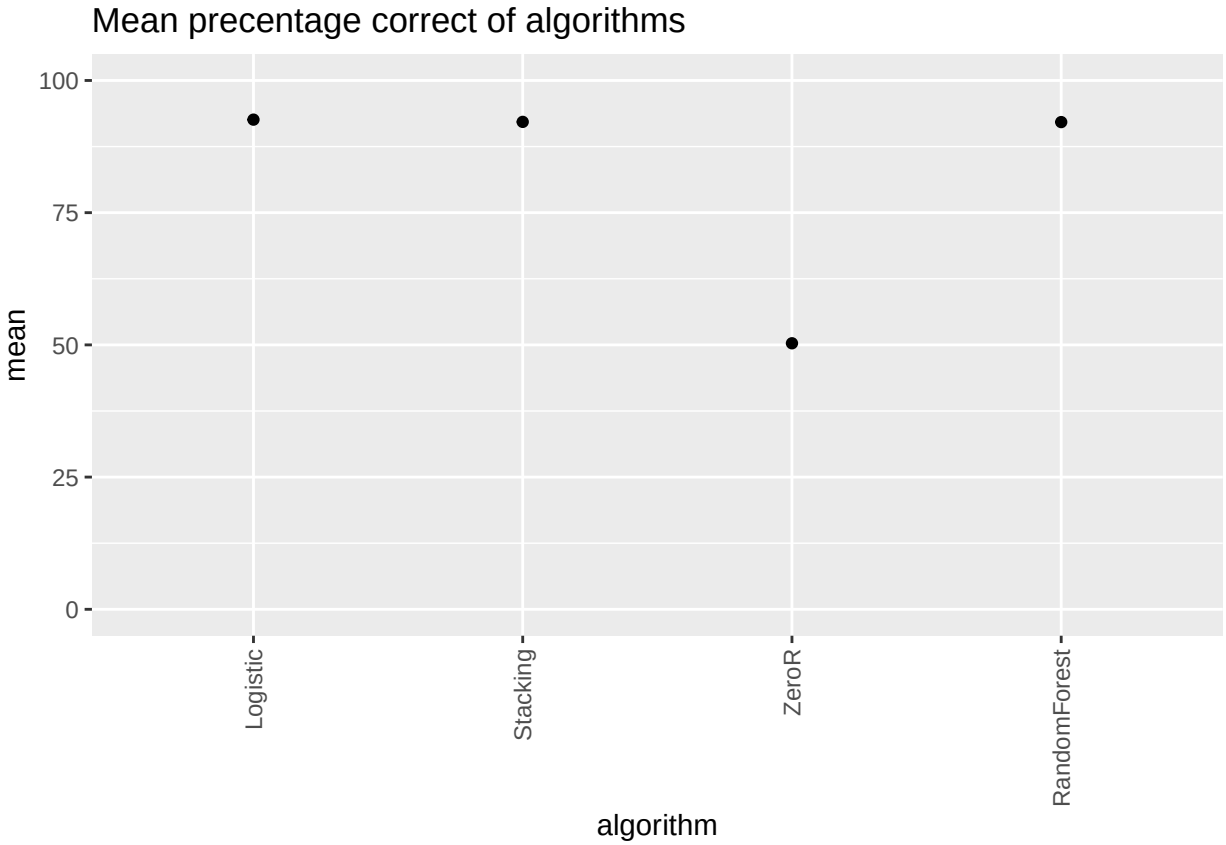


Figure 5: The mean percentage correct of the algorithms where the parameters were adjusted except for Zero-R which was ran with the default settings. Here, Logistic is the best performing though RandomForest and Stacking are not far behind.

	RandomForest	Logistic
Non adjusted	92.32	92.57
Adjusted	92.13	92.59

In the table above it is determined that the Logistic algorithm does perform better with the iterations set at 12, though the improvement is rather minor. The RandomForest setting does not perform better with iterations set at 45.

For attribute selection the WrapperSubsetEval tool and the BestSearch for the search method, 10-fold cross validation was used. The tool was ran on the Logistic algorithm. All attributes that were in 80% of the runs were kept while others were deleted. This resulted in the removal of the flipper length attribute and the island attribute.

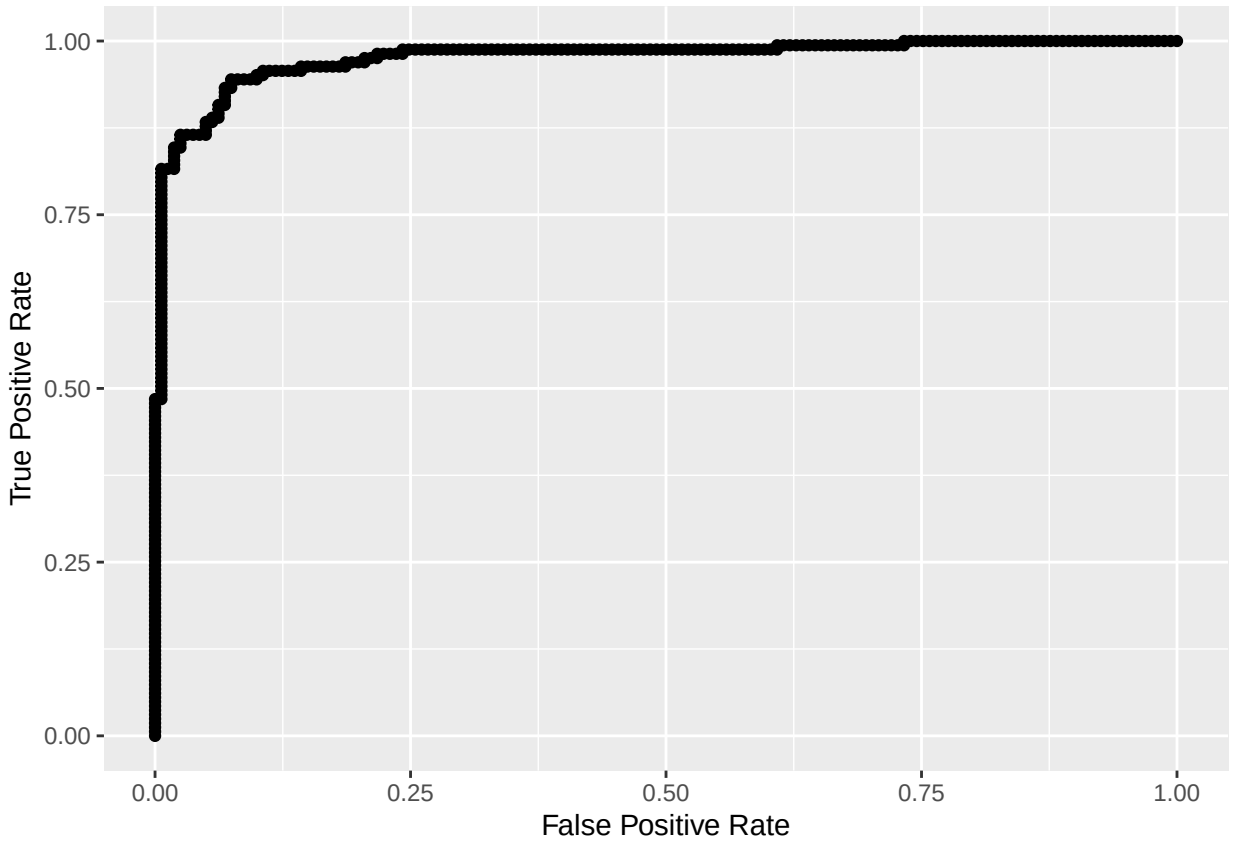
Again, WEKA was ran with the four algorithms that were used last time. Zero-R, Stacking, Logistic, and RandomForest, all with the optimized parameters and with 10-fold cross validation. This resulted in a table with the following mean correct percentage scores:

Table 4: Table with the different algorithms and data sets. The cleaned data set refers to the dataset before the WrapperSubsetEval tool was ran and the adjusted data set refers to the dataset after the WrapperSubsetEval was ran. The mean is a mean percentage correct scores.

algorithm	mean
adjusted_penguin_data_Logistic	92.84
adjusted_penguin_data_RandomForest	91.86
adjusted_penguin_data_Stacking	92.33
adjusted_penguin_data_ZeroR	50.3
cleaned_penguin_data_Logistic	92.59
cleaned_penguin_data_RandomForest	92.13
cleaned_penguin_data_Stacking	92.17
cleaned_penguin_data_ZeroR	50.3

Logistic and Stacking both show an improvement, whereas RandomForest shows a decrease in correct predictions. ZeroR, as predicted, shows neither improvement nor a decrease.

The results of the Logistic algorithm with adjusted data set and adjusted parameters were looked at further with a ROC curve where the positive value is the female penguin value.



The ROC plot above shows the performance of the Logistic algorithm. The area under the curve appears to be above 0.8 which is an excellent score as the perfect score would be 1. As a result of the different mean correct percentage scores, and the above ROC curve, the Logistic algorithm is chosen as the algorithm to use in the Java wrapper.

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. It is likely that due to the correlation between flipper length and body mass, flipper length was able to be removed later in the process, which led to an increase in the accuracy of the algorithm which can be seen in table 4 and the removal of the flipper length attribute for predicting penguin sex.

In figure 4 all algorithms were first ran with the default values, as can be seen all algorithms outperformed ZeroR, however when taking into account the initial goal of the research, only the top two performing algorithm were chosen for the next step. From the remaining algorithms, Zero-R, One-R, J48, and IBk were used for stacking, which led to creation of figure 5, where certain parameters were chosen to 'optimize'. Once more, Logistic outperformed the rest which is why this algorithm was used to run the attribute selection tool with. The other algorithms were not used for attribute selection which did result in a missed opportunity to see if these other algorithms could benefit from better attribute selection. In the end, taking into account the scores in table 4, the ROC curve made in figure 6, and the fact that it is a relatively simple algorithm, Logistic was chosen as the algorithm for the final product.

For parameter optimization, there was a focus on making sure there was no overfitting going on. This meant iteration depth was altered for the RandomForest and Logistic algorithm for stopping the algorithm early. More parameters could have been optimized but due to the relatively short time span, it was decided just the iteration depth would be optimized.

It is worth keeping in mind that this data only carried three species of penguins, and thus any other species are not supported in this algorithm. For future research, it may be worth looking into a way to classify the penguins without the species attribute as well as the delta C13 and delta N15 attributes as these two attributes were harder to accurately measure than the other numerical attributes.

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. Likely due to the high correlation between body mass and flipper length, the flipper length attribute was left out.

The Logistic model with species, culmen length, culmen depth, bodymass, delta C13, and delta N15, can predict the sex of the penguin with around 92% accuracy.

References

[1] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.*

[2] Daróczi G, Tsegelskyi R (2022).: *pander: An R 'Pandoc' Writer. R package version 0.6.5*,https://CRAN.R-project.org/package=pander

[3] Gorman KB, Williams TD, Fraser WR (2014)*Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081.*, doi:10.1371/journal.pone.0090081