Introduction

Margriet van der Molen

2023-10-10

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

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.

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, and a ROC plot was made to further analyze the results.

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