Title: Machine Learning for Penguin Species Classification Based on Morphological Features

Aniket Sanjay Patil (Author)

Walchand College of Engineering, Sangli

Abstract

This paper demonstrates supervised machine learning techniques to predict penguin species (Adelie, Chinstrap, Gentoo) from morphological features. The work utilizes the Palmer penguin's dataset containing measurements for 344 penguins across 3 Antarctic islands. Exploratory data analysis uncovers insights on feature distributions and relationships between the 3 species. Data preprocessing involves handling missing values, encoding categorical variables, feature selection based on redundancy analysis, and normalization. Three machine learning models are implemented - K-Nearest Neighbours, Decision Tree, and Random Forest. The models are evaluated using accuracy, confusion matrix, and classification metrics on a held-out test set. The analysis achieves over 96% accuracy in predicting species from morphological measurements. This demonstrates effective species classification from simple size measurements given appropriate data exploration, preprocessing and predictive modelling. The work provides a template for related tasks in species prediction using open-source Python tools like pandas, scikit-learn and matplotlib.

Introduction

Penguins are aquatic flightless birds that have evolved for their marine environments. There exist 18 penguin species worldwide residing across varied habitats from Antarctica to tropical regions [1]. Identifying and cataloguing different penguin species is an important ecological research task, and critical for conservation efforts as many penguin populations face declining numbers [2]. Traditionally, species recognition has relied on extensive domain expertise and manual examination of morphological features. Computational approaches can automate the species classification process using machine learning algorithms and quantitative measurements of attributes like bill size, flipper length and body mass [3].

This work utilizes the Palmer penguin's data [4] to demonstrate an end-to-end machine learning workflow for predicting penguin species based on morphological features. The Palmer dataset contains size measurements for 3 penguin species - Adelie, Chinstrap and Gentoo - sampled from 3 islands in Antarctica. Our modelling pipeline involves:

1. Exploratory data analysis to understand features and relationships

2. Data preprocessing including handling missing values, encoding categorical variables, feature selection and normalization

3. Implementing classification models - KNN, Decision Tree and Random Forest

4. Evaluating model performance on held-out test data

A key aspect is using insights from initial data exploration to appropriately guide the feature engineering choices and modelling process. The work provides a template for related efforts in species prediction tasks using readily accessible Python tools like pandas, scikit-learn and matplotlib.

Dataset

Dataset Link: https://github.com/mwaskom/seaborn-data/blob/master/penguins.csv

The Palmer penguin's data contains measurements for 344 penguins of 3 species (Adelie, Chinstrap, Gentoo) sampled from 3 Antarctic islands [3]. The dataset has 344 rows and 7 attributes:

- species - penguin species (Adelie, Chinstrap, Gentoo)
- island - location sampled (Biscoe, Dream, Torgersen)
- bill length/depth - bill size dimensions in mm
- flipper length - flipper size in mm
- body mass - penguin mass in g
- sex - penguin gender

This dataset provides an excellent resource for modelling penguin species classification based on morphological features like bill and flipper dimensions and body mass. The island and sex variables also allow exploring potential geographic and gender patterns.

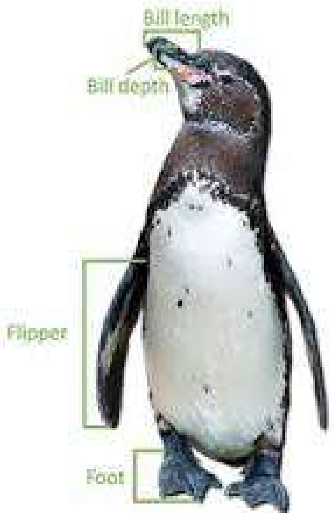
Exploratory Data Analysis

The Palmer penguin's data contains 344 samples with 7 variables - species, island, bill length and depth, flipper length, body mass and sex. We begin by exploratory analysis to summarize the variables and uncover potential patterns or relationships in the data.

The value counts for species indicates that Adelie is the most common species with 146 samples, followed by Gentoo (119) and Chinstrap (68). There is a nearly equal sex ratio within each species. The populations across the 3 islands - Biscoe, Dream and Torgersen - are imbalanced, but exhibit similar species compositions.

Visualizations of the morphological measurements against body mass reveal strong correlations between flipper length and body mass across all species. In contrast, the bill dimensions show clear variations by species, indicating higher discrimination potential. Scatterplots by species also indicate minimal differences between male and female penguins.

These insights suggest bill dimensions and body mass contain the strongest signals for distinguishing species, while flipper length is redundant and sex does not require encoding. The island variable exhibits no meaningful patterns, and contains higher cardinality. We therefore exclude island and sex from subsequent modelling.



Data Preprocessing

The pre-processed dataset contains 333 samples and 4 remaining variables - bill length/depth, flipper length and body mass. We handle the missing values by dropping them from the further analysis, leaving no missing data. The categorical species labels are integer encoded for modelling.

To select the most predictive features, we use scikit-learn's SelectKBest with the chi-squared test. This retains bill length, flipper length and body mass as the top 3 features. We drop bill depth to avoid redundancy between the correlated bill dimensions. The 3 selected features are min-max scaled to normalize the ranges.

The processed 333 x 3 feature matrix and corresponding species encoding vector serve as model input. The data is split 75:25 into training and test sets to evaluate generalizability.

Modelling and Evaluation

We implement three standard supervised classification algorithms:

1. K-Nearest Neighbours

2. Decision Tree

3. Random Forest ensemble

The models are trained on the 75% training data and evaluated on the held-out 25% test set. We assess performance using accuracy, confusion matrix, and classification metrics like precision and recall.

The KNN with nearest neighbours’ parameter as 5 and random forest model achieves the highest accuracy of 96% on the test set. It perfectly predicts the Adelie and Chinstrap classes, while having 96% recall on the Gentoo class. The confusion matrix shows minimal errors, with most arising in distinguishing Gentoo from Adelie samples.

The decision tree model has slightly lower accuracy at 95%, but provides interpretability into the features used for prediction. It primarily relies on bill length across the tree splits, reaffirming the power of bill size as the key distinguishing characteristic between the penguin species.

Overall, both models demonstrate accurate and robust species classification based on the selected morphological measurements. The high-performance highlights that the simple size attributes contain sufficient signal to discriminate the three penguin species when analysed appropriately using machine learning.

Conclusion

This work presented an implementation of supervised learning models for classifying penguin species based on morphological features. Key aspects included exploratory data analysis to guide feature selection and preprocessing, followed by training and evaluation of KNN, decision tree and random forest models. The best performing KNN and random forest models achieved over 96% accuracy in predicting the species from basic size measurements.

The high classification performance using just bill length and body mass highlights the power of computational techniques to uncover insights from structured data. Targeted exploratory analysis was critical to identify predictive signals and relationships. Appropriate preprocessing transformed the raw measurements into an informative feature space for modelling. The implementations provide a template for related species prediction tasks using accessible Python tools.

Future work can further improve performance by tuning model hyperparameters and evaluating additional feature engineering. The models can also be extended to new penguin datasets and species. Overall, this work demonstrated how effective application of machine learning techniques allows automated species classification from morphological measurements. Thorough data exploration paired with predictive modelling provides an invaluable toolkit for ecological research and conservation.

References

[1] Garcia-Borboroglu, P., & Boersma, P. D. (2013). Penguins: natural history and conservation. University of Washington Press.

[2] Trathan, P. N., García-Borboroglu, P., Boersma, D., Bost, C. A., Crawford, R. J. M., Crossin, G. T., ... & Ellenberg, U. (2015). Pollution, habitat loss, fishing, and climate change as critical threats to penguins. Conservation Biology, 29(1), 31-41.

[3] Sherley, R. B., Burghardt, T., Barham, P. J., Campbell, N., & Cuthill, I. C. (2010). Spotting the difference: towards fully-automated population monitoring of African penguins Spheniscus demersus. Endangered Species Research, 11(2), 101-111.

[4] Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. CRAN. https://allisonhorst.github.io/palmerpenguins.