# Q1 - Visualization and analysis of the Palmer dataset

The Palmer penguin dataset consists of 344 records of the physical attributes of three species of penguin living on three islands in Antarctica (Table 1) [1]. In this report, consideration is given to data cleaning and preparation, the dataset is explored through visualization and analysis is carried out to compare the accuracy performances of a small number of AI approaches.

| Attribute | Type | Values in the dataset |
|---|---|---|
| species | categorial | Adelie, Chinstrap, Gentoo |
| island | categorial | Torgersen, Biscoe, Dream |
| bill length | numerical | 32.1mm - 59.6mm |
| bill depth | numerical | 13.1mm - 21.5mm |
| flipper length | numerical | 172mm - 231mm |
| body mass | numerical | 2700g - 6300g |
| sex | categorial | Male, Female |

**Table 1:** Attributes of the Palmer penguin dataset

## Data cleaning - missing values, standardization and data imbalance

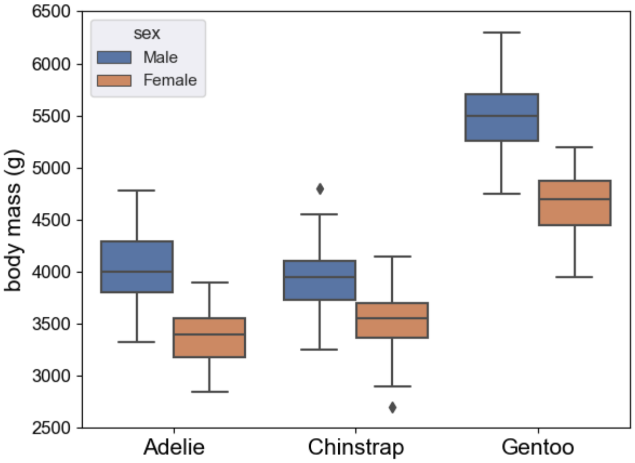

**Figure 1:** All numerical features show a significant statistical difference between the male and female measurements, as seen in the body mass boxplot above. Shown are median values, Q1 and Q3 quartiles, as well as outliers that are outside the range Q1-1.5IQR to Q3+1.5IQR, where IQR=Q3-Q1.

Two of these records can be deleted immediately as they are missing values for all of the numerical attributes and the sex feature and any imputation is unlikely to be reliable. The remaining nine records have no value only for the sex attribute. As can be seen in Figure 1, the physical attributes of the male and female of each species are statistically different and so it is reasonable to consider assigning a sex to those records missing this attribute. Following standardization, a Shapiro-Wilk test was performed to confirm each numerical attribute exhibits a normal distribution [2] and Z-tests were performed to assess separately both the hypothesis that the missing sex value is male and that it is female [3]. It was found that two of the records could be imputed as male and three as female and these were then retained in the dataset. The remaining four records were removed from the dataset. The cleaned dataset consisted of 338 records made up of 147 Adelie penguins (74 male, 73 female), 68 Chinstrap penguins (34 male, 34 female) and 123 Gentoo penguins (62 male, 61 female).

A number of the methods applied in this work involve distance measures and so may be biased in favour of features with smaller standard deviations [4]. This bias can be removed by standardizing the four numerical attributes independently (to give zero mean and unity standard deviation). Standardization uses only the statistics of training sets, but standardization is also applied to test sets. If a dataset is imbalanced, AI approaches may be biased in predicting classes that are more commonly found in the training data. The Palmer penguin dataset is somewhat imbalanced, with the number of Chinstrap records being around half of that of either Adelie or Gentoo, which are present in similar numbers. The importance of imbalance depends on the analysis method applied. It is known that all the methods adopted in the current work are generally little affected by imbalanced data [5] and so no modifications were made to reduce imbalance.

## Visualization of the dataset

Figure 2 shows the species distribution across the three islands in the study. Chinstrap and Gentoo penguins are found only on one island, so island is a potential confounding factor, possibly affecting physical characteristics due to environmental factors (such as predators or food supply). A Shapiro-Wilk test was used to confirm that the numerical features of the Adelie penguins (that are found on all the islands) are normally distributed and an ANOVA test confirmed that their physical characteristics are not significantly influenced by the island inhabited. Consequently, it was considered unlikely that the island is a confounding factor in the dataset.

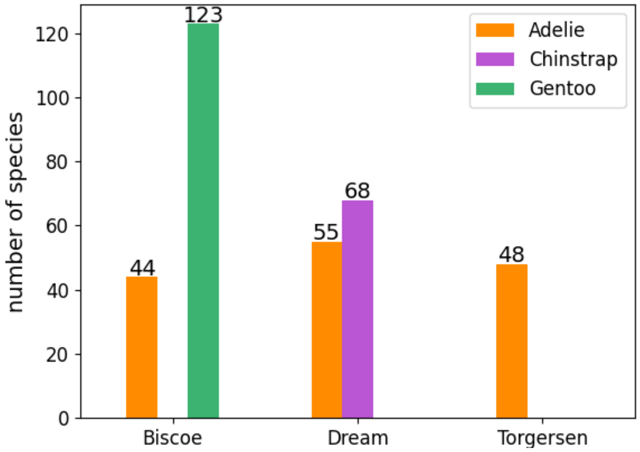

**Figure 2:** All numerical features show a significant statistical difference between the male and female measurements, as seen in the body mass boxplot above. Shown are median values, Q1 and Q3 quartiles, as well as outliers that are outside the range Q1-1.5IQR to Q3+1.5IQR, where IQR=Q3-Q1.

Pairwise scatterplots for the numerical features are shown in Figure 3. It can be seen that bill depth in combination with either flipper length or body mass yields a separable cluster of Gentoo penguins (shown in green) allowing them to be identified. No pairwise combination of numerical features completely separates Adelie (orange) from Chinstrap (purple) clusters, but good separation is provided in the distributions involving bill length, making this a candidate feature for distinguishing between these species.

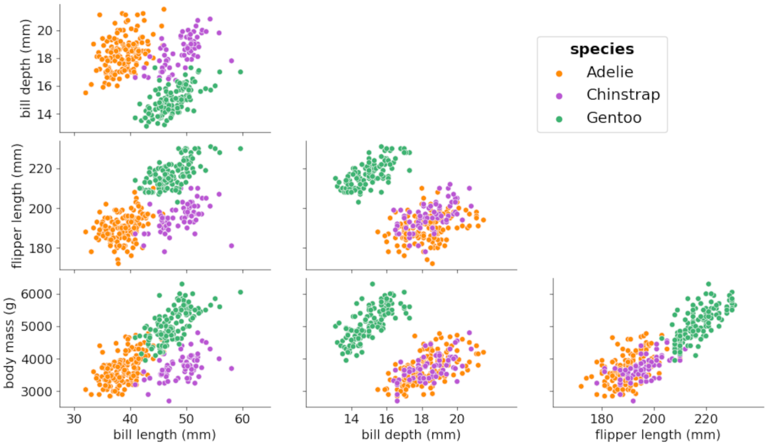

**Figure 3:** All numerical features show a significant statistical difference between the male and female measurements, as seen in the body mass boxplot above. Shown are median values, Q1 and Q3 quartiles, as well as outliers that are outside the range Q1-1.5IQR to Q3+1.5IQR, where IQR=Q3-Q1.

Figure 1 above shows there is a difference in the body masses of the male and female samples for each of the three species. Differences between the sexes for the other three numerical physical characteristics in the dataset were also apparent. Since narrower distributions are apparent if the sex of the species is considered rather than just the species itself, including sex is likely to provide a finer grained distinction for species classification and this knowledge can be used to improve performance, as discussed in the analysis section below.

## Implementation

All code was written in Python 3.11 [6] using ‘Scikit-Learn’ libraries [7] running under Ubuntu Linux [8]. The code is available in a Gitub repository [9]. Predicting the penguin species from the given features is a classification problem. Results are obtained from a baseline method, two convention classification approaches, namely k-Nearest Neighbour (knn) [10] and random forest [11], unsupervised k-means (following cluster labelling) [12] and a novel combined visualization and analysis (CVA) approach introduced here that uses insights from visualizations combined with two-dimensional linear Support Vector Machines (SVMs) classification.

For all the methods implemented, 20% of the dataset was kept for a test set. To reduce the potential for overfitting, the classification methods (all but k-means) were trained using 'holdout validation', where the remaining 80% of the dataset was used in a five-fold cross-validation configuration [13]. For all methods, the Scikit-Learn function GridSearchCV was employed to tune metaparameters [7]. Table 2 shows the values selected for the metaparameter grid for each of the AI methods and those values that gave best performance were selected to obtain the accuracy results from the test set. Scikit-Learn includes pseudo-random procedures for selecting validation and test set values and 100 of these were used both when selecting metaparameters and when deriving accuracy results.

| Method | Metaparameters | Value |
|---|---|---|
| knn | number of nearest neighbours | k 1, 2 |
| | weight function for prediction | unifo |
| | distance metric for computing neighbours | Manh |
| random forest | number of trees in the forest | 5, 10, |
| | maximum depth of trees | no ma |
| | minimum number of samples to split node | 2, 5, |
| | minimum number of samples at leaf node | 1, 2, |
| | function to measure quality of split | gini, |
| k-means | number of clusters k | 2, 3, |
| | centroid initialization method | k-mea |
| | number of runs with different centroid seeds | 2, 5, |
| | maximum number of iterations | 5, 10, |
| CVA | regularization parameter (C) | 0.1, 1 |
| | kernel coefficient (gamma) | 1, 0.1 |
| | kernel type | rbf, li |

**Table 2:** Metaparameters considered in training the methods. The values shown in bold are those that most consistently produced results of best accuracy during validation and so were selected for generating results

Table 3. Mean classification accuracy from 100 test sets each generated by a pseudo random approach and using the parameters identified in Table 2 Method Accuracy baseline, most numerous species 43.49kNN, all features 99.24kNN, no island 99.46random forest, all features 98.57random forest, no island, flipper length or body mass 98.59k-means, all numerical features 97.06k-means, separate clusters for each sex 98.23CVA using bill depth, flipper length, bill length 98.56CVA using bill depth, flipper length, bill length, sex 98.78

You can write in **bold**, or *italics* or `true type`, often the latter is used for specific commands or libraries in a programming language, as in 'I used `numpy` v1.23.4 to…'. Notice the use of the left quote symbol found in the top left of the keyboard to get the left quote. There is also blackboard bold often used for things like $\mathbb{R}$ for real numbers and there is calligraphic for fancy things like $\mathcal{L}$ but this is becoming increasing irrelevant to what you are likely to need!

# References

[1] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data.* R package version 0.1.0. 2020. DOI: 10.5281/zenodo.3960218. URL: https://allisonhorst.github.io/palmerpenguins/.

[2] Samuel S Shapiro and Maurice B Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[3] David Freedman, Robert Pisani, and Roger Purves. *Statistics.* W. W. Norton & Company, 2007. ISBN: 978-0393929720.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. ISBN: 978-0387848570.

[5] Haibo He and Yunqian Ma. “Learning from imbalanced data”. In: *Knowledge and Data Engineering, IEEE Transactions on* 21.9 (2009), pp. 1263–1284.

[6] Python Software Foundation. *Python 3.11 Documentation*. https://docs.python.org/3.11/. 2022.

[7] scikit-learn contributors. *scikit-learn: Machine Learning in Python*. Version 1.2.2. 2023. URL: https://scikit-learn.org.

[8] Canonical Ltd. *Ubuntu 20.04.1 LTS*. Canonical Ltd. London, UK, 2020. URL: https://releases.ubuntu.com/20.04/.

[9] Tim Mulvaney. *AI coursework repository*. 2024. URL: https://github.com/timmulvaney/AI.

[10] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] Leo Breiman and Adele Cutler. “Random Forests”. In: *Machine Learning*. Vol. 45. 1. Springer. 2001, pp. 5–32.

[12] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, 2005.

[13] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.