

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.

Feature	Type	Values in the dataset	Importance
island	categorical	Torgersen, Biscoe, Dream	0.13
bill length	numerical	32.1mm - 59.6mm	0.16
bill depth	numerical	13.1mm - 21.5mm	0.02
flipper length	numerical	172mm - 231mm	0.65
body mass	numerical	2700g - 6300g	0.01
sex	categorical	Male, Female	0.04
species	categorical	Adelie, Chinstrap, Gentoo	class

Table 1: Attributes of the Palmer penguin dataset.

Importance is calculated using random forest.

Data cleaning - missing values, encoding, standardization and imbalance

The two records missing the sex and all numerical features were removed as imputation is unlikely to be reliable. The remaining nine records are missing only the sex attribute. Figure 1 shows the physical attributes of the male and female of each species differ statistically and so it is reasonable to consider imputing sex for those records. Following standardization, a Shapiro-Wilk test confirmed each numerical attribute has a normal distribution [2] and separate Z-tests were applied to assess the hypotheses that the missing sex value is male or female [3]. Two of the records could be imputed as male and three as female and these were retained. with the remaining four records being removed. The cleaned dataset has 338 records, 147 Adelie (74 male, 73 female), 68 Chinstrap (34 male, 34 female) and 123 Gentoo (62 male, 61 female).

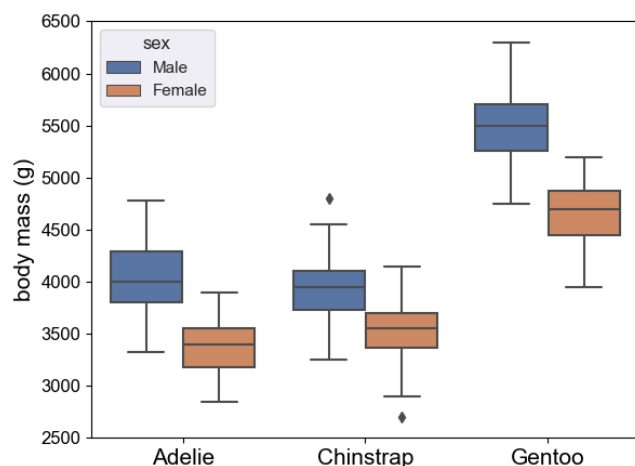


Figure 1: All numerical features show a significant statistical difference between male and female, as in the body mass above. Shown are median values, upper and lower quartiles, and outliers.

The categorical features in the dataset were encoded to numerical values. ‘One-shot’ encoding of categorical features was not found to improve performance. A number of AI methods are known to be biased in favour of numerical features with smaller standard deviations [4], but this can be reduced by standardization of features to zero mean and unity standard deviation. Standardization statistics are calculated only from the training set, but are also applied to the test set. If a dataset is imbalanced, AI predictions may be biased towards classes more frequently found in the training data. In the Palmer penguin dataset, the number of Chinstrap records is around half of that of both Adelie and Gentoo, but, as all the methods adopted in the current work are known to be little affected by imbalanced data [5], no modifications were made.

Visualization of the dataset

Figure 2 shows the species distribution for the islands. Chinstrap and Gentoo penguins are found only on one island, making island a potential confounding factor whose individual environmental factors may influence physical characteristics. A Shapiro-Wilk test confirmed the normal distribution of the numerical features of the Adelie penguins (found on all islands) and an ANOVA test

confirmed the features are not significantly influenced by the island inhabited. Consequently, the island is not a confounding factor in the dataset.

The feature importance scores in Table 1 represent relative contributions in predicting the species and were calculated using a random forest approach. Results reported later show that performance improvements can be achieved by concentrating classification on the more important features.

The pairwise scatterplots for the numerical features are shown in Figure 3. 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 completely separates Adelie (orange) from Chinstrap (purple) clusters, but the best candidate feature for doing so is in the distributions involving bill length.

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.

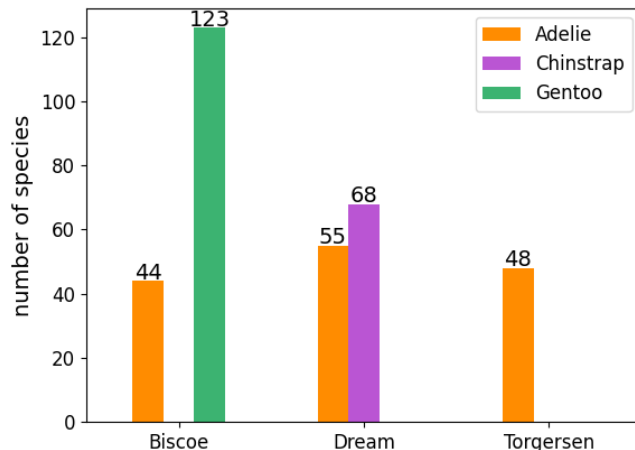


Figure 2: Adelie samples are from all three islands, but Gentoo and Chinstrap from only one

Methodology

All code is in Python 3.11 [6] using ‘Scikit-Learn’ libraries [7] 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 that uses practical visualizations and Support Vector Machine (SVM) classification.

To reduce the potential for overfitting, the classification methods (all but k -means) were trained using ‘holdout validation’, where 80% of the dataset was used in a five-fold cross-validation configuration [13]. The remaining 20% was kept for a test set. For all methods, the Scikit-Learn function GridSearchCV was employed to tune metaparameters [7]. Table 2 shows the values selected for the metaparameter grid. Those giving the best performance were selected to generate accuracy results (the percentage of correctly predicted species) from the test set. Calculation of the metrics

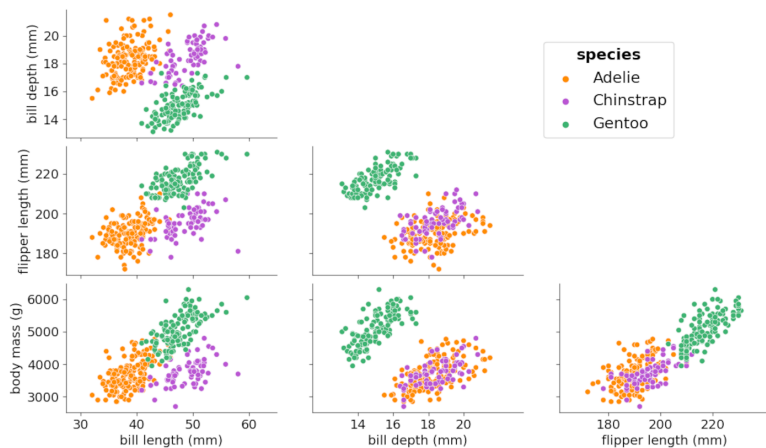


Figure 3: Pairwise distributions of numerical features. Gentoo can be distinguished, but Adelie and Chinstrap may not be completely separable from one another

‘precision’ and ‘recall’ were also considered, but these are only relevant if false positives or false negatives (respectively) are to be avoided.

Scikit-Learn provides 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.

Results and analysis

The results in Table 3 include a baseline that is used to demonstrate performance improvements achieved by the AI methods being considered. In classification, the baseline method is often simply to select the most frequent class in the observations and, in this work, this is the Adelie penguins, giving an accuracy of 43.49% (147/338).

Classification method 1 - *knn*

The performance of *knn* was found to be improved by omitting features from training. An exhaustive search involving omitting all combinations of features in turn determined that the best accuracy was obtained when island was omitted and this occurred when $k=3$. It appears that island was not providing any additional information and the higher value of k implies better generalization may have been achieved.

Classification method 2 - Random forest

Including all of the features in the analysis provided an accuracy marginally better than could be achieved using *knn* when its features were carefully selected. No performance improvement was found by using fewer features, indicating that considerably less implementation effort is needed to achieve good performance using random forest. A marginal improvement in performance was apparent when island, flipper length or body mass were not included in training.

Unsupervised method - *k*-means Although an unsupervised clustering method, *k*-means can be used for classification by matching clusters to classes. The *k*-means method is

Method	Metaparameters	Values considered
<i>knn</i>	number of nearest neighbours k weight function for prediction distance metric for neighbours	<i>1</i> , 2, 3, 4, 5, 6, 8, 10 <i>uniform</i> , distance <i>Manhattan</i> , Euclidean
random forest	number of trees in the forest maximum depth of trees minimum samples to split node minimum samples at leaf node function for quality of split	5, <i>textbf{10}</i> , 15, 20, 25 <i>no maximum</i> , 10, 20 2, 5, 10 <i>1</i> , 2, 4 <i>gini</i> , entropy
<i>k</i> -means	number of clusters k centroid initialization method number of runs for centroid seeds maximum number of iterations	2, 3, 4, 5, 6, 7, 8, 9, 10 <i>k-means++</i> , random 2, 5, 10, 20 5, <i>10</i> , 20, 50
CVA	regularization parameter (C) kernel coefficient (gamma) kernel type	0.1, 1, <i>10</i> , 100 <i>1</i> , 0.1, 0.01, 0.001 rbf, <i>linear</i> , polynomial

Table 2: Metaparameters values shown in italics most consistently produced training results of best accuracy during validation and were selected for generating results

Method	Accuracy
baseline, most numerous species	43.49%
<i>k</i> NN, all features	99.24%
<i>k</i> NN, no island	99.46%
random forest, all features	98.57%
random forest, no island, flipper length or body mass	98.57%
<i>k</i> -means, all numerical features	97.06%
<i>k</i> -means, separate clusters for each sex	98.23%
CVA using bill depth, flipper length, bill length	98.56%
CVA using bill depth, flipper length, bill length, sex	98.78%

Table 3: Mean classification accuracy from 100 test sets each generated by a pseudo random approach and using the parameters identified in Table 2

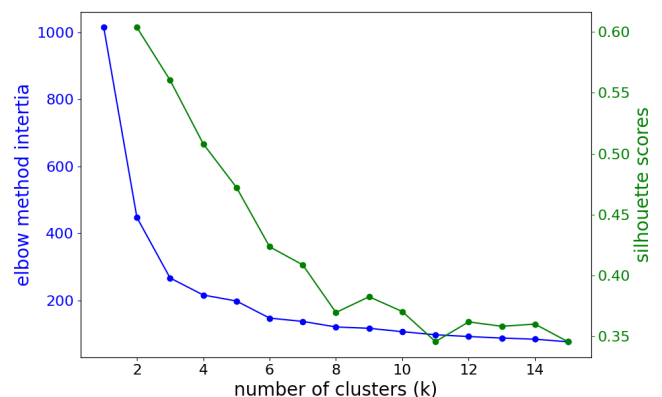


Figure 4: For *k*-means, the elbow method is the change in slope of ‘inertia’ ($k=3$) and the silhouette method is the score closest to 1 ($k=2$)

normally applied only to numerical features so only they were included in this work. The number of clusters (k) can be selected using elbow and silhouette methods, as shown in Figure 4. Empirically, accuracy improved significantly when $k \geq 4$ as clusters were not reliably formed for all three species for smaller values of k , Figure 5 illustrates the mapping of classes to clusters for two feature dimensions. No improvement in accuracy was obtained by reducing the number of features, but, in an additional experiment, a set of k -means clusters was created for each sex and this led to a small improvement in accuracy.

A novel combined visualization and analysis (CVA) approach The CVA approach involves visualizing pairwise plots of features to identify a short sequence of two-dimensional SVM classifiers. CVA requires manual effort to understand the nature of the dataset, in contrast with ‘black box’ classification approaches that are applied with limited knowledge of the method and little insight into the nature of the data. The drawbacks of the CVA approach are that it may not always be feasible to extract the necessary visualizations insights particularly for large dimensional datasets. For the Penguin data, CVA was able to produce results of accuracy almost as good as conventional approaches.

An application of CVA to the Penguin dataset is illustrated in Figure 6. Figure 6a shows the relationship between bill depth and flipper length and SVM is used to find a suitable ‘decision boundary’ that separates Gentoo from the other two species. Figure 6b then shows a second SVM line that best separates Adelie and Chinstrap using bill length and bill depth. A small improvement in accuracy was achieved when separate SVM models were developed for each penguin sex.

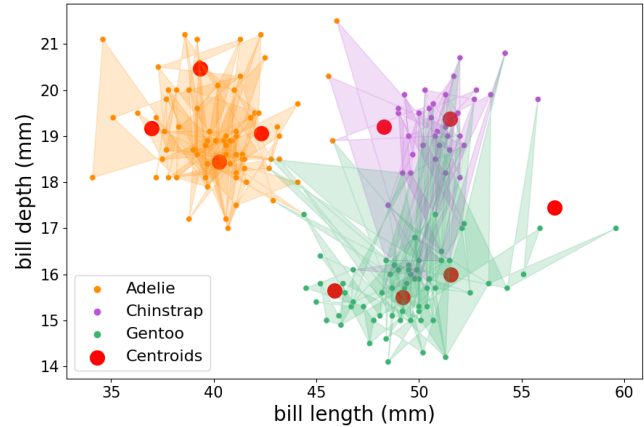
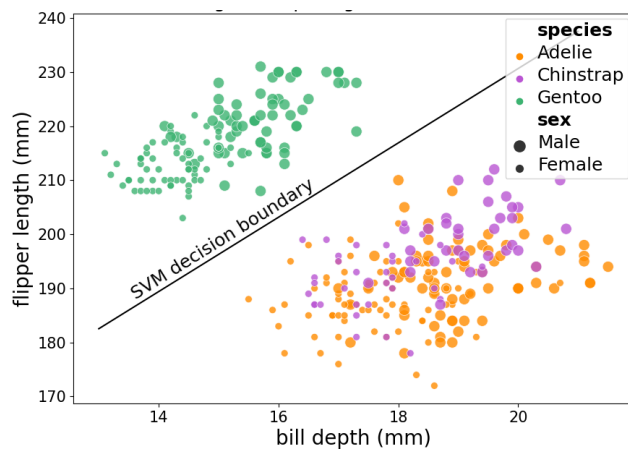
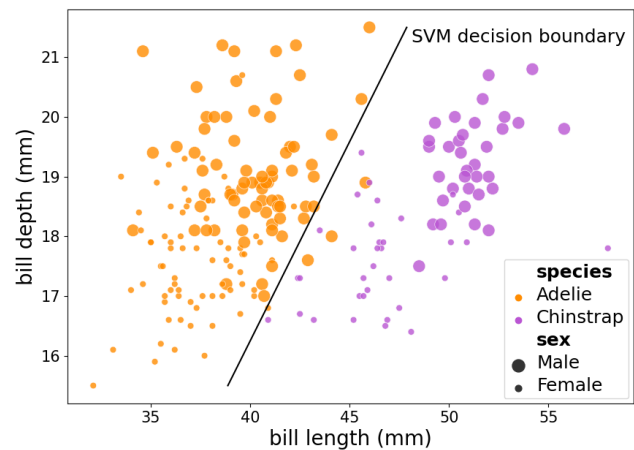


Figure 5: k -means clusters mapped to species using majority voting. Class assignments are shown by polygon colour ($k=10$, 50 samples coloured).



(a) Gentoo can be distinguished from other species



(b) Adelie and Chinstrap can be partially separated

Figure 6: Two-stage CVA approach with boundaries fitted using SVM to training data of feature pairs

Conclusions

With careful data preparation, optimization of metaparameters and robust application of training and testing methods, the k nn and random forest classification methods produced high-quality results. The k -means classification accuracy results were somewhat worse, but this is to be expected

as the approach does not take advantage of target data information that is known to the supervised approaches.

A classifier that is able to achieve 100% accuracy for the given data is possible, but its performance when applied to new unseen data would likely exhibit poor generalization. The novel CVA approach is designed to use insights available in visualizations. Although needing to be tailored to each problem and not well-suited to high-dimensionality data, its internal operations are easy to visualize, an advantage not afforded to general-purpose classification methods. For the penguin data, it was able to produce accuracy results similar to those of other classification methods.

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.