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 then explored through visualization and analysis is carried out to compare the

Feature	Type	Values in the dataset	Importance
island	categorical	Torgersen, Biscoe, Dream	0.12 (4)
bill length	numerical	32.1mm - 59.6mm	0.37 (1)
bill depth	numerical	13.1mm - 21.5mm	0.17 (3)
flipper length	numerical	172mm - 231mm	0.23 (2)
body mass	numerical	2700g - 6300g	0.11 (5)
sex	categorical	Male, Female	0.01 (6)
species	categorical	Adelie, Chinstrap, Gentoo	class

Table 1: Palmer penguin dataset features. Importance was calculated using random forest and a ranking is shown.

accuracy of a small number of AI approaches in classifying penguin species.

Data cleaning - missing values, encoding, standardization and imbalance

The two records missing the sex amd 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
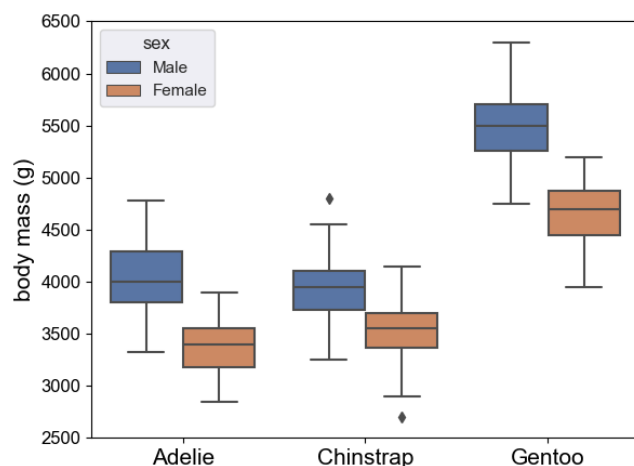

Figure 1: All numerical features show a significant statistical difference between male and female, as in the body mass example above. Shown are median values, upper and lower quartiles, and outliers.

male, 73 female), 68 Chinstrap (34 male, 34 female) and 123 Gentoo (62 male, 61 female).

The categorical features in the dataset were encoded to numerical values. 'One-shot' encoding of categorical features was considered but 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. In this work, standardization statistics were calculated only from training sets, but applied to all data. 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 for either Adelie or 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. It is clear the island is not a confounding factor in the dataset.

Table 1 shows the feature importance scores (and rank in parentheses) indicating relative contributions to predicting the penguin species. Results reported later show that performance improvements can be achieved by concentrating classification on the more 'important' features.

Pairwise scatterplots for the numerical features are shown in Figure 3. Bill depth, combined 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 com-

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

pletely separates Adelie (orange) from Chinstrap (purple) clusters, but the best candidate feature for doing so is in 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 seen 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 'results and analysis' section below.

Methodology

The code is available on Gitub [6] and was written in Python 3.11 [7] using 'Scikit-Learn' libraries [8]. Predicting the penguin species from the given features is a classification problem. Results are obtained from two conventional classification approaches, namely k-Nearest Neighbour ($k$nn) [9] and random forest [10], unsupervised k-means (following cluster labelling) [11] and a novel combined visualization and analysis (CVA) approach that is a mix of 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 [12]. The remaining 20% was kept for a test set. For all methods, the Scikit-Learn function GridSearchCV

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

was employed to tune metaparameters [8]. Table 2 shows the values selected for the metaparameter grid. Those giving the best performance were selected to generate accuracy results (the
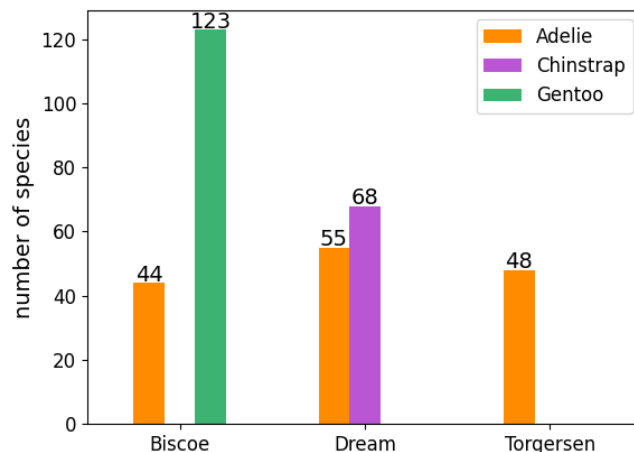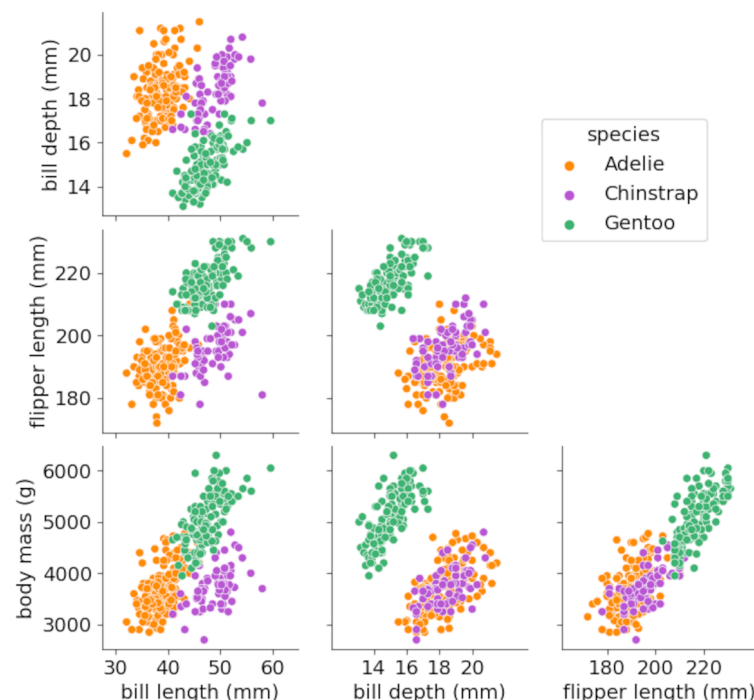
percentage of correctly predicted species) from the test set. The metrics 'precision' and 'recall' were also caculated, but as these are only relevant if false positives or false negatives (respectively) are specfically to be avoided, they are not included in this report.

Results and analysis

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. 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 - *k*nn An additional test was carried to investigate if, compared to the accuracy found for all features, the omission of features and combinations of features could improve accuracy. An improvement was found when island was omitted and when $k=3$. It appears that island did not provide 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 gave an accuracy marginally better than achieved using *k*nn. In contrast with *k*nn, no performance improvement was found 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 a clustering method, *k*-means can be used for classification by matching clusters to classes. The *k*-means method is normally applied only to numerical features and 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,

Method	Metaparameters	Values considered
*k*nn	nearest neighbours k	*1*, 2, 3, 4, 5, 6, 8, 10
	prediction weight function	*uniform*, distance
	neighbours distance metric	*Manhattan*, Euclidean
random forest	trees in the forest	5, *10*, 15, 20, 25
	maximum depth of trees	*no maximum*, 10, 20
	min samples to split node	*2*, 5, 10
	min samples at leaf node	*1*, 2, 4
	quality of split function	*gini*, entropy
k-means	number of clusters k	2, *3*, 4, 5, 6, 7, 8, 9, 10
	centroid initialization	*k-means++*, random
	runs for centroid seeds	2, *5*, 10, 20
	max number of iterations	5, *10*, 20, 50
CVA	regularization parameter	0.1, 1, *10*, 100
	kernel coefficient	*1*, 0.1, 0.01, 0.001
	kernel type	rbf, *linear*, 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 (range)
baseline, Adeleie species	43.49%
*k*NN, all features	99.24% (97.06%-100%)
*k*NN, no island	99.46% (97.06%-100%)
random forest, all features	98.49% (92.65%-100%)
random forest, no body mass	98.49% (92.65%-100%)
k-means, numerical features	97.06%
k-means, two sex clusters	98.23%
CVA, three main features	98.56%
CVA, as above but with sex	98.78%

Table 3: Classification accuracy mean value and range for 100 pseudo-random test sets and using the metaparameters identified in Table 2
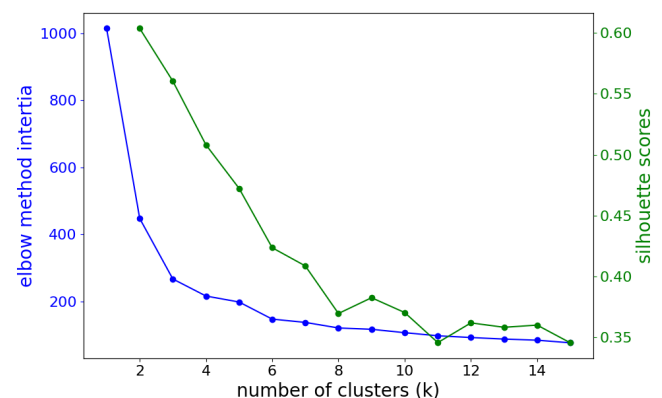

Figure 4: The *k*-means elbow is the change in slope of 'inertia' ($k=3$) and the silhouette is the 'score' closest to 1 ($k=2$)

Figure 5 illustrates the mapping of classes to clusters for two feature dimensions. No accuracy improvement was obtained by reducing the number of features, but, when a separate 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 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 often applied with limited knowledge of the method and little insight into the nature of the data. The main drawback of the CVA approach is that it may not always be feasible to extract the necessary visualization insights, particularly for large dimensional datasets.

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

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. CVA was able to produce results of accuracy almost as good as conventional approaches and a small improvement was apparent when separate SVM models were developed for each penguin sex.
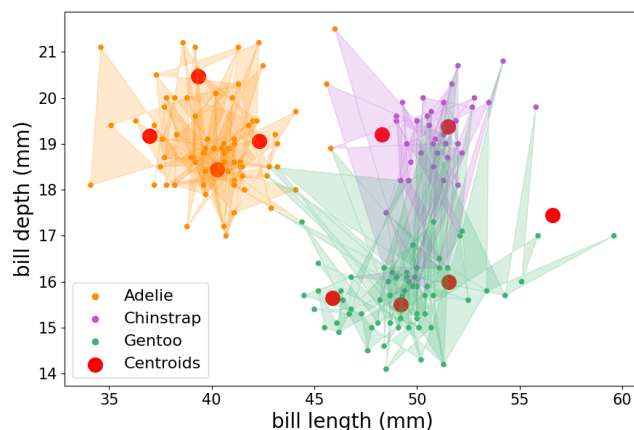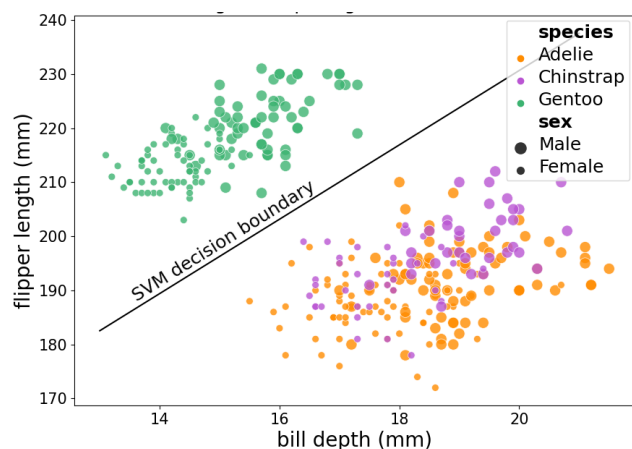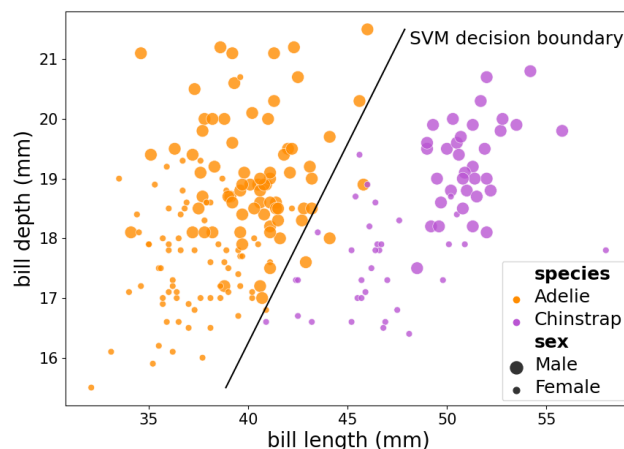

(a) Gentoo can be distingushed from other species

(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. As expected, the k-means accuracy results were comparitively worse as in training it does not take advantage of target data information known to the supervised approaches.

The novel CVA approach was able to produce accuracy results similar to those of other classification methods. Although needing to be tailored to each problem and not well-suited to high-

dimensionality data, its internal operations are transparent and this is in contrast with general-purpose classification methods.

A classifier that is able to achieve 100% accuracy for the given data is possible, but is likely exhibit poor generalization.

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

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

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

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

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

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

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