Q1 - Visualization and analysis of the Palmer penguin 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, the data are cleaned and carefully prepared for investigation, the dataset is explored through visualization and analysis is carried out to compare the accuracy of a small number of AI approaches in classifying penguin species.

Feature	Type	Values in the dataset	Importance
island	categorial	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	categorial	Male, Female	0.01 (6)
species	categorial	Adelie, Chinstrap, Gentoo	class

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

Data cleaning - missing values, encoding, standardization and imbalance

The two records missing 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, making it reasonable to consider imputing sex for those records would be reliable. Following standardization, a Shapiro-Wilk test confirmed all the numerical attributes have a normal distribution [2] and 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; the remaining four records were removed. The cleaned dataset has 338 records, 147 Adelie (74M, 73F), 68 Chinstrap (34M, 34F) and 123 Gentoo (62M, 61F).

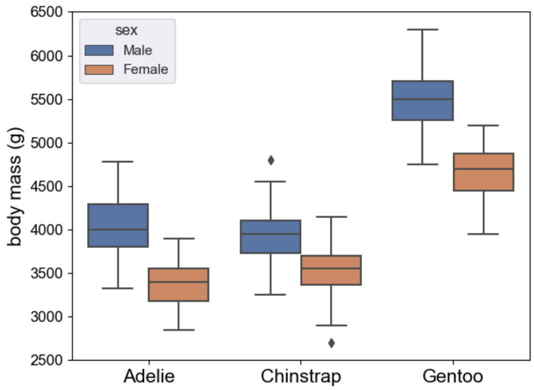


Figure 1: All numerical features show a significant statistical difference between male and female, as in the body mass example above. Shown are median values, upper and lower quartiles, and outliers.

The categorical features in the dataset were encoded to numerical values as this is required for a number of the AI methods. Some AI methods are known to be biased in favour of numerical features with smaller standard deviations [4], but this bias can be reduced by the standardization of features to zero mean and unity standard deviation. Standardization statistics were calculated only from training sets, but were applied to all data. If a dataset is imbalanced, AI predictions may be biased towards classes more frequently found in the training data. All the methods adopted in the current work are known to be little affected by imbalanced data [5], so no modifications were made.

Visualization of the dataset

Figure 2 shows Chinstrap and Gentoo penguins are found only on one island, making it a potential confounding factor since environmental factors may influence physical characteristics. A Shapiro-Wilk test showed the numerical features of the Adelie penguins have a normal distribution and an ANOVA test confirmed Adelie features are not significantly influenced by the island inhabited. Island is thus unlikely to be a confounding factor.

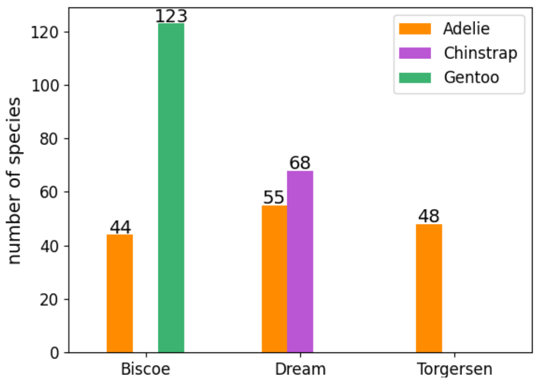


Figure 2: Adelie is on all three islands, but Gentoo and Chinstrap samples are from only one.

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 completely separates Adelie (orange) from Chinstrap (purple) clusters, but the best candidate feature for doing so is bill length.

Figure 1 above shows the normative body masses of the male and female differ for all the species. Differences between the sexes for the other three numerical physical characteristics were also apparent. Since narrower distributions are extracted if the sex of the species is considered rather than just the species itself, sex is likely to provide a finer grained distinction for species classification. This knowledge was able to improve performance in some cases, as discussed in ‘results and analysis’.

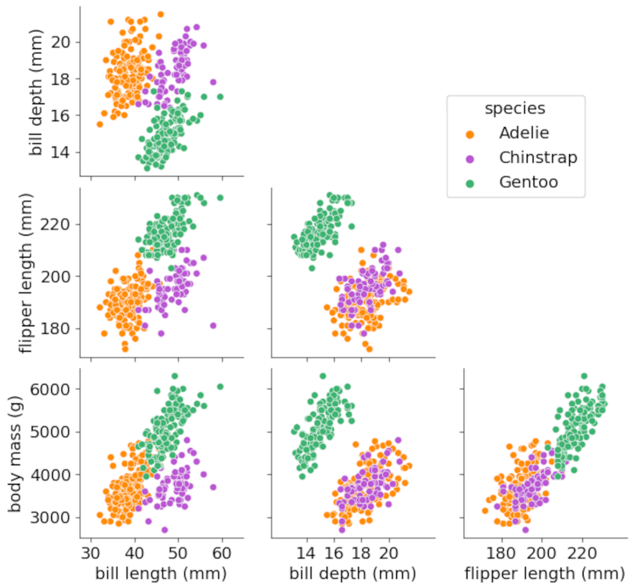


Figure 3: Pairwise distributions of numerical features. Gentoo can be distinguished, but Adelie and Chinstrap may not be completely separable from one another

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 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 calculated, but as these are only relevant if false positives or false negatives (respectively) are of interest, they are not included in this report.

Method	Metaparameters	Values considered
$k$nn	nearest neighbours k	***1***, 2, 3, 4, 5, 6, 8, 10
	prediction function	***uniform***,distance
	distance metric	***Manhattan***,Euclid
random forest	number of trees	5, ***10***, 15, 20, 25
	maximum depth	***no max***, 10, 20
	min samples to split	***2***, 5, 10
	min samples at leaf	***1***, 2, 4
	split function	***gini***,entropy
k-means	number of clusters k	2, 3, ***4***, 5, 6, 7, 8, 10
	centroid initialize	***k-means++***,random
	runs for centroid	2, ***5***, 10, 20
	maximum iterations	5, ***10***, 20, 50
CVA	regularization	0.1, 1, ***10***, 100
	kernel coefficient	***1***, 0.1, 0.01, 0.001
	kernel type	rbf, ***linear***,poly

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

Results and analysis

The Scikit-Learn pseudo-random procedure was used for selecting validation and test set values and 100 (indices 1 to 100) of these were used both when selecting metaparameters and when deriving accuracy results. The results in Table 3 include a baseline to demonstrate the performance improvements achieved by the AI methods. In classification, the baseline method is often simply to select the most frequent class in the observations. Here, this is the Adelie penguins, giving an accuracy of 43.49% (147/338).

Method	Accuracy (range)
baseline, Adeleie species	43.49%
kNN, all features	99.24% (97.06%-100%)
kNN, no island	99.46% (97.06%-100%)
random forest, all features	98.57% (95.59%-100%)
random forest, no body mass	98.49% (92.65%-100%)
k-means, numerical features	97.03% (94.12%-97.06%)
k-means, two sex clusters	99.18% (94.12%-100%)
CVA, three main features	98.98% (95.35%-100%)
CVA, separate sex models	99.25% (90.91%-100%)

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

***k*nn** The accuaracy of $k$nn obtained using the dataset could be improved by removing certain features (particularly those shown as less important in Table 1). The best improvement was found when island was omitted and when k=3. It appears that island did not provide any additional information and the larger value of k implies better generalization may have been achieved.

Random forest Including all of the features in the analysis, the accuracy of random forest was marginally worse than achieved using $k$nn. No performance improvement was found using fewer features, indicating that random forest may be less influenced by superfluous information in the training data.

***k*-means** Although a clustering method, k-means can be used for classification by matching clusters to classes. Figure 4 illustrates the mapping of classes to clusters for two feature dimensions and k=10. k-means is normally applied to numerical features and only they were included in this work. Using the elbow approach, the number of clusters (k) was found to be two, but was found to be three using the silhouette method. Empirically, accuracy improved when $k \geq 4$, as otherwise clusters were not reliably always formed for all three species. An accuracy improvement was achived by creating a separate set of clusters for each sex.

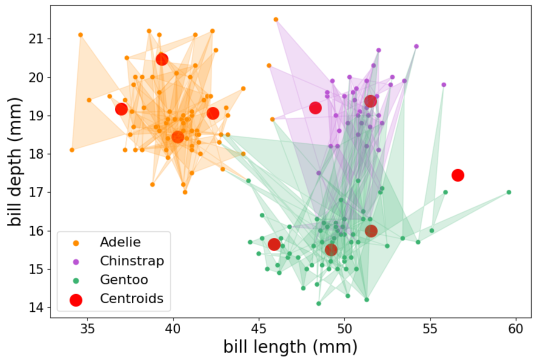
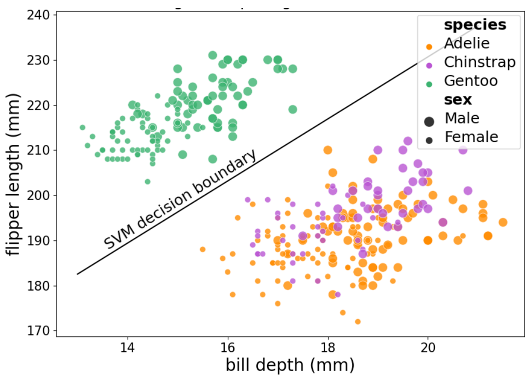
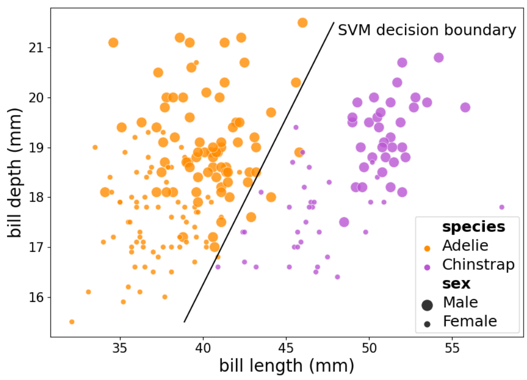


Figure 4: k-means clusters mapped to species using majority voting. Mappings are shown by polygon colour (k=10, 50 samples coloured).

CVA requires manual inspection of pairwise feature plots to identify suitable sets of two-dimensional SVM classifiers. An application of CVA to the Penguin dataset is given in Figure 5. Figure 5(a) shows the relationship between bill depth and flipper length, and SVM determines finds a suitable 'decision boundary' to separate Gentoo from the other two species. Figure 5(b) plots bill length against bill depth and shows the SVM boundary that best separates Adelie and Chinstrap. An improvement in accuracy was apparent when separate SVM models were developed for each sex.



(a) Gentoo can be distingushed from other species



(b) Adelie and Chinstrap can be partially separated

Figure 5: 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 its training 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, in contrast with many general-purpose classification methods.

Question 2 - Ethical challenges and threats in AI

Racial Bias in Medical Algorithms

In 2019, a widely used US healthcare algorithm was found to discriminate by prioritising hospital services based on historical spending records, resulting in the allocation of relatively less future funding and fewer referrals for black patients [13, 14]. Through the application of a series of test data sets, Obermeyer et al. [15] identified this inadvertent bias and the team was able to mitigate against it by adjusting the model's training labels.

The fact that this third-party assessment and adjustment were possible, demonstrates how exposing a model's internal operations can aid bias identification and removal [16, 17]. Ensuring greater transparency of AI models is becoming the subject of legislation, for example the 2023 EU AI Act aims to enforce transparency principles by requiring developers to disclose an algorithm's variables, data sources, and selection logic [18, 19]. While ensuring that organisations building AI systems are held accountable for the processes used in their development may lead to algorithmic changes that reduce bias [20, 21], care needs to be taken that the removal of bias doesn't significantly affect the performance of the model in its application domain [22].

AI system safety and existential risks in warfare

Recent developments in AI have led many researchers to believe that AI systems capable of directly acting in the real world based on decisions they have taken autonomously will become available later this century [23]. With such advancements comes the risk that AI systems whose decision-making does not prioritise human welfare may pose a threat to life [24].

A specific example of a military AI system posing an existential risk [25, 26], is one that decides maximising human casualties would be the best strategy to achieve a high-level battlefield objective [27]. Recent deployments of automated missiles that activate on target acquisition [28, 29], have raised ethical concerns over the use of AI in situations where human beings are potential targets [30]. If such advanced AI was given control of powerful military weapons and applied more widely, the ramifications for the human race's survival could be profound [31].

Addressing these existential threats requires international cooperation to guarantee the transparency of AI algorithms [32, 33]. A potential future safeguard is to include a human-controlled override in all military AI systems [34], although Russell [35] warns that super-intelligent AI may be capable of removing such safety measures. Ultimately, a global strategy that prioritises human wellbeing in all areas of AI usage will be essential.

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.

[13] S. Jemielity. *Health care prediction algorithm biased against black patients, study finds*. [Online]. Available: `https://news.uchicago.edu/story/health-care-prediction-algorithm-biased-against-black-patients-study-finds`. [Accessed 27 03 2024]. 2019.

[14] H. Ledford. "Millions Affected by Racial Bias in Health-care Algorithm". In: *Nature* 574 (Oct. 2019), pp. 608–609.

[15] Z. Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366 (Oct. 2019), pp. 447–453.

[16] B. Séroussi, K. F. Hollis, and L. F. Soualmia. "Transparency of Health Informatics Processes as the Condition of Healthcare Professionals' and Patients' Trust and Adoption: the Rise of Ethical Requirements". In: *Yearbook of Medical Informatics* 29.01 (2020), pp. 007–010.

[17] P. D. Winter and A. Carusi. “(De)troubling transparency: Artificial Intelligence (AI) for clinical applications”. In: *Medical Humanities* 49.1 (2023), pp. 17–26.

[18] European Parliament. *EU AI Act: first regulation on artificial intelligence.* [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. [Accessed 27 03 2024]. 2023.

[19] L. Edwards. “The EU AI Act: a summary of its significance and scope”. In: *Artificial Intelligence (The EU AI Act)* 1 (2021).

[20] J. Donovan et al. “Algorithmic Accountability: A Primer”. In: (2018).

[21] T. Lawry. *AI in Health.* 1st. Boca Raton: CRC Press, Taylor & Francis Group, 2020.

[22] M. Kearns and A. Roth. *The Ethical Algorithm.* 1st. New York: Oxford University Press, 2020.

[23] K. Grace et al. “When Will AI Exceed Human Performance? Evidence from AI Experts”. In: *Journal of Artificial Intelligence Research* 62 (July 2018), pp. 729–754.

[24] T. Ord. *The Precipice: Existential Risk and the Future of Humanity.* 1st. New York: Hachette Books, 2020.

[25] R.V. Yampolskiy. “Taxonomy of Pathways to Dangerous Artificial Intelligence”. In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.* Phoenix, Arizona, Mar. 2016.

[26] M. L. Cummings. *Artificial Intelligence and the Future of Warfare.* London: Chatham House for the Royal Institute of International Affairs, 2017.

[27] J. Barrat. *Our Final Invention.* 1st. New York: Thomas Dunne Books, 2013.

[28] K. Atherton. *Brookings: Loitering Munitions Preview the Autonomous Future of Warfare.* [Online]. Available: https://www.brookings.edu/articles/loitering-munitions-preview-the-autonomous-future-of-warfare. [Accessed 30 03 2024]. 2021.

[29] I. Bode and T. Watts. “Loitering Munitions and Unpredictability: Autonomy in Weapon Systems and Challenges to Human Control”. In: (2023).

[30] R. J. Emery. “Algorithms, AI, and Ethics of War”. In: *Peace Review: A Journal of Social Justice* 33 (2021), pp. 205–212.

[31] M. Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence.* 1st. Toronto: Alfred A. Knopf, 2017.

[32] P. Cihon. “Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development”. In: (2019).

[33] D. Leslie. “Understanding Artificial Intelligence Ethics and Safety”. In: *CoRR* abs/1906.05684 (2019).

[34] A. Critch and D. Krueger. *AI Research Considerations for Human Existential Safety*. 2020.

[35] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. 1st. New York: Viking.