1 Predicting Palmer Penguins

1.1 Introduction

The objective of this investigation was to explore the Palmer Penguins dataset and develop predictive models for the species of penguins (Adelie, Chinstrap, and Gentoo) based on various physical measurements (e.g. bill length, bill depth, flipper length, and body mass).

1.2 Data Exploration

The initial stage in my research was to check the data for missing values and do simple statistical tests to determine the distribution of the characteristics. I discovered that the following columns included missing values: `bill_length_mm`; `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`, and `sex`, which were addressed using mean imputation for numerical characteristics.

To focus on the most relevant features for species classification, I excluded the `rowid` and `year` columns from my initial pair plot visualisation. The pairplot in Figure 1 depicts the links between the major physical measures and the species. The pairplot demonstrated a strong distinction between the penguin species based on physical measures. This was unexpected because it does not always occur in real-world datasets. The obvious distinction indicated that the chosen characteristics were highly discriminative for this classification job, making it easier for models to differentiate between species.

1.3 Unsupervised Learning

I utilised K-Means clustering to investigate the data's inherent structure without using species labels. The data was standardised before applying K-Means to three clusters representing the three species. Figure 2 displays the generated clusters. One striking discovery in Figure 2 was the clear separation of species, demonstrating that even basic models have the ability to attain high accuracy.

The clusters were compared against the actual species labels using a contingency table (Table 1). Another interesting observation was that the Gentoo species was almost perfectly clustered by the K-Means algorithm, suggesting strong underlying patterns in the data in the absence of prior labelling.

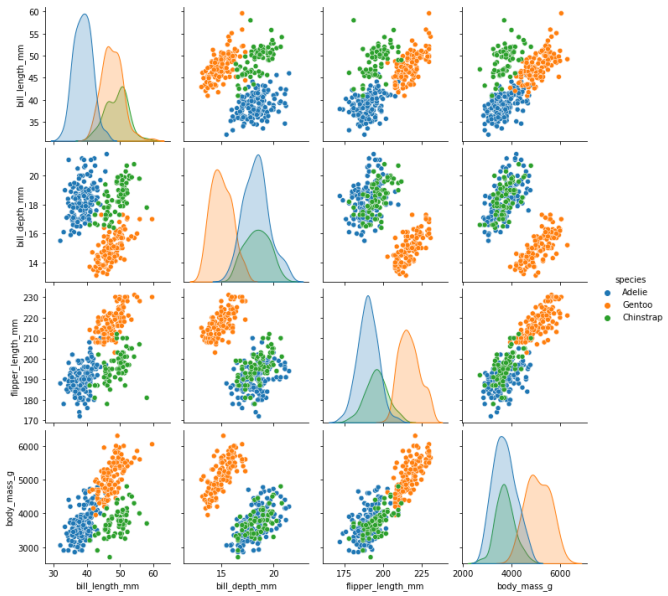

Figure 1: Pairplot of numerical features coloured by penguin species.

	Cluster 0	Cluster 1	Cluster 2
Adelie	127	0	25
Chinstrap	5	0	63
Gentoo	0	123	1

Table 1: Contingency table comparing K-Means clusters with actual species.

1.4 Classification Algorithms

I used two categorisation algorithms: k-NN and Random Forest. In the dataset, I noticed an imbalance in the number of cases for each species and employed upsampling to guarantee that all classes were represented evenly in the training dataset. This strategy was not immediately visible, but it was critical for training successful classifiers since it prevented the model from becoming biased towards the majority class. The data was then divided into two sets: training and testing (70% training, 30% testing).

1.4.1 k-Nearest Neighbours

The k-NN classifier was trained with $k = 5$. This value was selected based on a grid search approach, optimising for accuracy. The performance of the

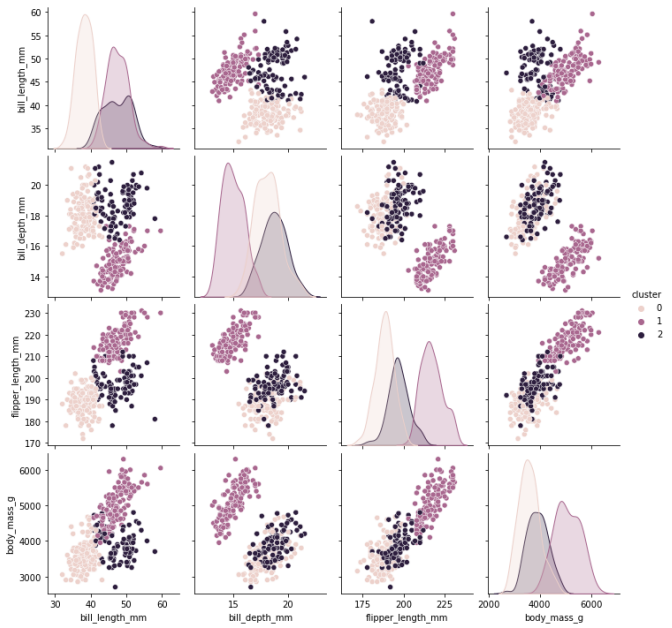

Figure 2: K-Means clustering results. Clusters are visualised and compared with actual species.

k-NN model is summarised in Table 2.

The k-NN classifier was trained with $k = 5$. This figure was chosen using a grid search algorithm that prioritised accuracy. The performance of the k-NN model is reported in Table 2.

	Precision	Recall	F1-score	Support
Adelie	0.90	0.70	0.79	54
Chinstrap	0.78	0.89	0.83	36
Gentoo	0.87	1.00	0.93	47
Accuracy	0.85			137

Table 2: Classification report for k-NN classifier.

1.4.2 Random Forest

The Random Forest classifier was trained using 100 estimators, which is a common method for balancing performance and computational efficiency. Table 3 describes the performance of the Random Forest model.

	Precision	Recall	F1-score	Support
Adelie	1.00	0.96	0.98	54
Chinstrap	0.95	1.00	0.97	36
Gentoo	1.00	1.00	1.00	47
Accuracy	0.99			137

Table 3: Classification report for Random Forest classifier.

1.5 Comparison and Conclusion

A Dummy Classifier was used to create a baseline model that predicted the most common class. The performance of this baseline model is reported in Table 4. The baseline model's accuracy was 26%, which corresponded to the distribution of the most common class. Comparatively, as demonstrated in Tables 2 and 3, the k-NN and Random Forest classifiers outperformed the baseline model with scores of 85% and 99%, respectively. This significant disparity emphasises the need of adopting more complicated models for classification tasks, which use the dataset's key properties to produce accurate predictions.

	Precision	Recall	F1-score	Support
Adelie	0.00	0.00	0.00	54
Chinstrap	0.26	1.00	0.42	36
Gentoo	0.00	0.00	0.00	47
Accuracy	0.26			137

Table 4: Classification report for the baseline Dummy Classifier.

In this work, I used both unsupervised and supervised learning approaches to predict penguin species. The first data analysis revealed distinct patterns among the species, which were subsequently highlighted using K-Means clustering. The k-Nearest Neighbours and Random Forest classifiers were trained and tested, and both outperformed the baseline model.

The Random Forest model's high accuracy indicates that it is well-suited to this dataset, most likely because of its ability to deal with complicated feature interconnections and significance. The results show that both techniques can effectively classify penguin species based on their physical characteristics. Future research might investigate different components or more complex models to enhance forecast accuracy.

2 Ethics in Data Science and Artificial Intelligence

Ethical problems abound in the fast-growing fields of data analytics and artificial intelligence (AI), with data security and the amplification of biases in datasets as major concerns. These issues can erode trust in technology and harm individuals and society so it is important to utilise robust data protection, transparency, fairness and accountability policies.

The Cambridge Analytica case demonstrates the ethical issue surrounding data protection. The political consulting business got millions of Facebook users' personal information without their consent and used it to influence voting decisions in many political campaigns (e.g. 2016 US presidential election). This was problematic because customers were unaware that their information was being collected and used without their consent, highlighting the sensitivity of personal information and its potential misuse. Data exploitation has serious real-world effects, including the potential to undermine democracy. The core issue is a lack of openness and inadequate protection of personal privacy.

Addressing these issues necessitates a comprehensive strategy that includes stiffer laws, increased transparency, and stronger security measures. Enforcing strict data protection standards (e.g. The EU GDPR) gives individuals more control over their personal information (European Parliament and Council of the European Union, 2016). Businesses must be open about their data gathering techniques and present clear, easily understandable privacy policies. Advanced security measures, such as encryption and anonymisation, are required to safeguard data from breaches and unauthorised access (Information Commissioner's Office, n.d.).

Another major ethical issue is the amplification of biases in historical datasets. For example, the Chicago Police Department's predictive policing algorithms have been found to disproportionately target black persons, frequently employing biased past crime data and reinforcing social stereotypes (Richardson, Schultz, & Crawford, 2019). The ethical issue here is systematic inequality, such that AI systems trained on biased data reinforce preconceived notions, resulting in unfair treatment of some populations and widening socioeconomic disparities. Regular bias audits and fairness evaluations can aid in the detection and reduction of biased tendencies (Barocas, Hardt, & Narayanan, 2019). Creating diverse and inclusive datasets can help remove discrimination and promote justice (Buolamwini & Gebru, 2018). Transparency and accountability in AI decision-making processes are essential for building confidence and assuring ethical outcomes (Diakopoulos, 2016).

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning.* http://fairmlbook.org.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of machine learning research* (Vol. 81, p. 77-91). Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56-62. doi: 10.1145/2844110

European Parliament and Council of the European Union. (2016). *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)* (Vol. L 119). Retrieved from http://data.europa.eu/eli/reg/2016/679/oj

Information Commissioner's Office. (n.d.). *Guide to the general data protection regulation (gdpr).* https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/. (Accessed: 2024-05-28)

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, *94*, 15-55. Retrieved from https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/