Exploratory Data Analysis

The dataset [1], comprising 344 entries across 9 columns, includes species, island, bill measurements, flipper length, body mass, and sex, representing penguins from three species across the Palmer Archipelago.

Figure 1's pairplot, indicates that flipper length and body mass are key in distinguishing penguin species, notably differentiating Gentoos from Adelies and Chinstraps. Bill length also helps to separate species, particularly Gentoos from others, while bill depth, though less significant alone, enhances classification when used with additional features. Hence, employing a multifaceted approach to feature selection is advisable for accurate classification.

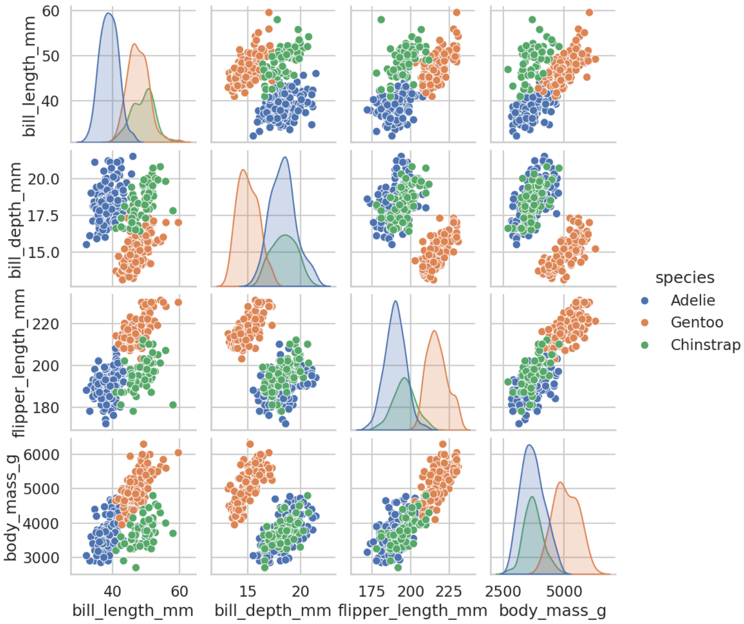

Figure 1: Pairplot of numerical features

Further Inspection and Pre-Processing

Most columns in the dataset are fully populated; however, a few entries are missing in the following columns: bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, and sex. Additionally, there exists a class imbalance with the Chinstrap species having significantly fewer records compared to the Adelie and Gentoo species. Recognizing these imbalances is crucial for the classification process as they may influence the model's performance. Upon further inspection, it is observed that Gentoo penguins are exclusive to Biscoe Island, while Chinstrap penguins are found only on Dream Island. Therefore, if a penguin is encountered on Torgersen Island, it can be immediately identified as an Adelie.

Initially, I removed 'rowid' and 'year' columns for being irrelevant to classification. Each numerical feature had two missing values, while the 'sex' column had eleven. For the missing values in the numerical features, I imputed the median value due to its robustness against outliers. Regarding outliers, a maximum of 6 were identified and left in the dataset, as I was not confident that they weren't natural variations and the limited number did not influence the later classification.

For the 'sex' column I implemented a ***novel*** approach. As missing entries constituted 3% of the column, they were predicted using a Random Forest Classifier. 'Species' and 'island' were label-encoded for model compatibility. The 'sex' data was split into `known` and `unknown` groups; the known set was processed, label-encoded, and divided into a 70-30 train-test split. A RandomForestClassifier, without hyperparameter tuning but achieving 90% accuracy, predicted the unknown 'sex' values. Post-prediction, 'species' and 'island' were converted back to their original formats. The classification report indicated a balanced prediction outcome between 'female' and 'male' classes.

	Bill Length (mm)		Bill Depth (mm)		Flipper Length (mm)		Body Mass (g)	
Species	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Adelie	38.83	2.69	18.34	1.22	190.00	6.54	3703.00	457.92
Chinstrap	48.83	3.34	18.42	1.14	195.82	7.13	3733.09	384.34
Gentoo	47.48	3.08	15.00	1.00	217.02	6.71	5067.74	510.45

Table 1: Summary of penguin species characteristics

Clustering

After processing the dataset, k-means clustering was implemented. This began with one-hot encoding for the categorical variables 'sex' and 'island' and excluding 'species' as the target; this increased the dimensionality of the dataset. The dataset was normalized using StandardScaler [2], as the K-means algorithm is sensitive to feature scales. Post-encoding and normalization, Principle Component Analysis (PCA) reduced the newly increased dimensionality, selecting two principal components to preserve 90% of the variance. An elbow plot, of inertia vs k-value, determined the optimal cluster number as three. With $k = 3$, k-means segregated the data into distinct species-related clusters. Feature importance, derived from PCA, pinpointed critical attributes: body mass and flipper length for the first component; bill depth and sex for the second, reinforcing prior observations from pairplot (Figure 2, right). The clusters separated species effectively, Figure 2 (left): Cluster 0 mainly included Gentoos, Cluster 1 Adelies, and Cluster 2 a mix of Adelies and Chinstraps, validating the clustering approach against the biological distinctions previously observed; Gentoos have more distinct features, whilst Adelies and chinstrap have greater overlap.

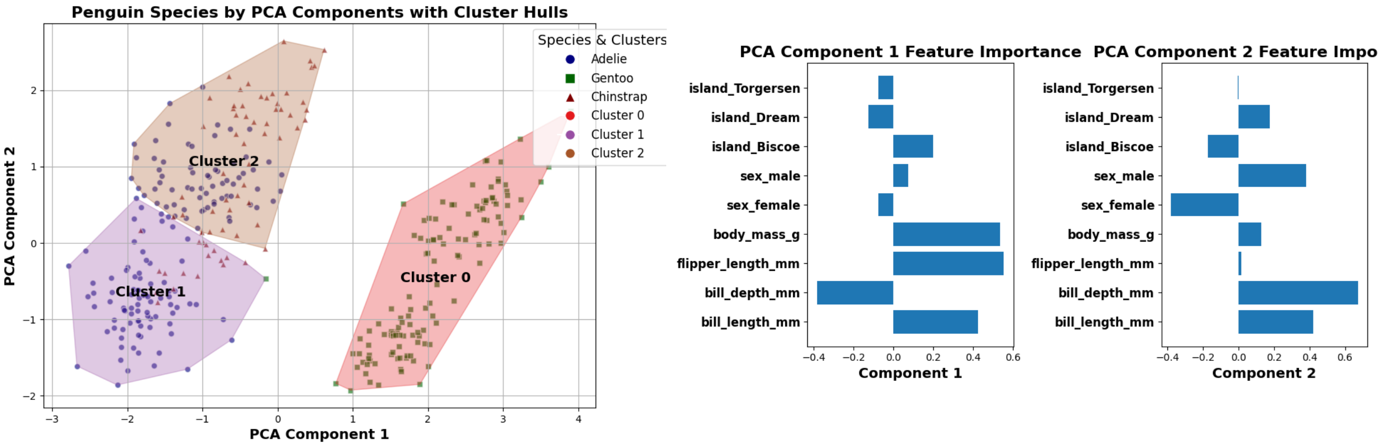

Figure 2: (Left) K-means clustering delineating penguin species. **(Right)** Feature importance for each PCA dimension.

Model Selection, Supervised Learning

Next, I partitioned the data into an 70-30 train-test split using a fixed seed to ensure consistent evaluation across models. Given the dataset's simplicity, I established a majority class baseline using scikit-learn's `DummyClassifier` with the `'most_frequent'` strategy, reflecting the prevalent Adelie class. This approach yielded a baseline accuracy of `43%`, setting a benchmark for evaluating the performance enhancements of more complex models.

Naive Bayes

Initially, I explored Naive Bayes classification, due to its relative simplicity and lack of optimisable hyperparameters. It is a probabilistic approach based on Bayes' theorem. The classifier predicts the probability that a given instance belongs to a particular class, C_k, by calculating:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Here, $P(C_k|x)$ is the posterior probability of class C_k given predictors x, $P(x|C_k)$ is the likelihood, $P(C_k)$ is the class prior probability, and $P(x)$ is the predictor prior probability.

Multinomial Logistic Regression

Multinomial logistic regression was selected due to observable linear relationships of some of the features and its interpretability. The model calculates class probabilities using the softmax function:

$$P(Y = k|X = x_i) = \frac{e^{\beta_k^T x_i}}{\sum_{j=1}^{K} e^{\beta_j^T x_i}},$$

optimizing parameters β_k for each class k by maximizing the log-likelihood,

$$\ell(\beta) = \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P(Y = k|X = x_i).$$

Regularization counters overfitting by modifying the log-likelihood to

$$L(\beta) = \ell(\beta) - \lambda \sum ||\beta_k||_p,$$

with $p = 1$ or 2 for L1 or L2 regularization, respectively. Hyperparameter selection, specifically the regularization strength λ, where $C = 1/\lambda$ and penalty type (p), was pivotal. Using GridSearchCV with 5-fold cross-validation, I explored differing combinations to enhance generalization. Initial trials with 'C': [0.001, 0.01, 0.1, 1, 10] pinpointed $C = 10$ and L2 penalty as optimal, indicating moderate regularization that balances bias and variance effectively while keeping all features in play. Expanding the parameter grid to include 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000] shifted optimal settings to $C = 100$ and L1 penalty. However, this led to a decrease in test accuracy, suggesting some overfitting. Thus, I selected $C = 10$ and L2 for its balance and robustness against overfitting, maintaining equitable feature weighting.Furthermore, learning curves for training and cross-validation were plotted, with converging score around 0.85 as the number of training examples increased; indicating improved model generalization with more data.

Model Evaluation

Both models were far superior to the simple majority-class baseline model, demonstrating that they identified distinct patterns in the data as opposed to merely predicting the dominant class each time.

As shown by Table 2 the MLR classifier outperforms the NB classifier in terms of precision, recall, and F1-scores across the different classes, as well as in the overall accuracy (99.0% for MLR vs. 97.1% for NB). Specifically, MLR shows higher precision and F1 scores for the 'Adelie' and 'Chinstrap' classes. This indicates that the MLR model is better at identifying the correct species without as many false positives and is better balanced in terms of precision and recall, especially for these two species. The 'Gentoo' class, however, has perfect scores for both models, suggesting that features distinguishing this species from the others are well-represented and clear enough for both classifiers to make perfect or near-perfect classifications; as was ascertained by earlier data explorations. Additionally, the feature importance analysis in Figure 3, further validates predictions from the initial data explorations; with features like bill length having the highest weighted importance in MLR.

Both models performed exceptionally well, with Naive Bayes erring on two samples where the MLR did not, and both models sharing one misclassification— a Chinstrap penguin with ***rowid***: 307. Despite this specimen not being marked as an outlier, its features deviate from the average by more than one standard deviation: bill length 40.9 mm, bill depth 16.6 mm, flipper length 187 mm, body mass 3200 g. However, given that this penguin is a female, which typically have smaller dimensions, these measures might align with the norm for female Chinstraps. Misclassification likely occurred due to its resemblance to an Adelie penguin's characteristic sizes.

MLR most likely outperforms NB in this scenario, as it does not rely on the assumption of feature independence like NB does. Additionally, MLR is more adaptable to various data distributions since it doesn't presuppose a specific feature distribution, unlike NB, which typically assumes a Gaussian distribution for continuous variables. Furthermore, LR can better manage the class imbalances through methods like log-likelihood optimisation.

Metric	**Logistic Regression (LR)**			**Naive Bayes (NB)**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Adelie	0.978	1.000	0.989	0.977	0.956	0.966
Chinstrap	1.000	0.947	0.973	0.900	0.947	0.923
Gentoo	1.000	1.000	1.000	1.000	1.000	1.000
Accuracy	0.990			0.971		
Macro Avg	0.993	0.982	0.987	0.959	0.968	0.963
Weighted Avg	0.991	0.990	0.990	0.972	0.971	0.971

Table 2: Comparison of Multinomial Logistic Regression and Naive Bayes

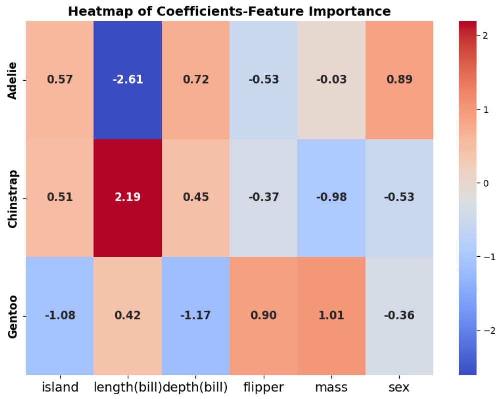

Figure 3: Heatmap showing feature contribution to MLR model.

Question 2

The existence of bias within training datasets, particularly that of gender bias, is a major concern. The prevalent cultural and linguistic norms are usually shaped by dominant ideologies, often leaving marginalized groups on the fringes. Hence, a lack of consideration towards the handling of training data, or the simple scarcity of quality data, can lead to unaccountable AI models unknowingly promoting biases. The extensive work of feminist scholars in highlighting the manifestation of subliminal gender ideology through language can be considered . For example, expressions such as 'single mum' and 'working mother' not only reflect prevailing social perceptions of women but also influence models to internalise these subtle patterns as textual features [3]. This can become a problem in any number of settings, especially when AI is implemented to hire candidates for companies. For example, a striking case involved Amazon's AI recruiting tool, found to be biased against women. This tool downgraded resumes including words that women are more likely to use in favour of resumes heavy in male-oriented language; words such as 'executed' and 'captured', highlighting how models can perpetuate gender biases embedded in the data they are trained on [4]. Furthermore, the biases can come in less obvious forms; such as clustering candidates based on locale, whereby ethnic minorities tend to congregate.

One way to address these issues is to advocate for transparency, such that decisions at every step of the process are justified and a greater onus is placed on the importance of correct handling of socially-curated information. This can mean that data collection processes, alongside justification of the design choices that have been made, should be published alongside the development of tools. Furthermore, design teams can incorporate a diverse range opinions, reflecting on the training data; and constantly update and re-evaluate the datasets used to train models; such that, verdicts do not become static and outdated.

Another ethical issue underpinning AI, is the use of personal data to train models. It is only now becoming a topic of debate, as to whether the use of people's readily available data to train AI models is ethically just. The Clearview AI controversy highlights these critical issues, as it involved the unauthorised scraping of billions of images to create a facial recognition database used by law enforcement [5]. This raised significant privacy concerns, as individuals' photos were used without consent, not to mention the risks of misidentification and bias, particularly affecting marginalised groups.

The EU AI Act directly addresses concerns similar to those raised by the Clearview AI case by establishing a comprehensive legal framework for AI, categorising systems by risk and imposing stringent requirements on high-risk application. Under the new legislation, use cases such as Clearview would be classified as 'Unnaccetable Risks' and outright banned [6]. Furthermore, the act mandates transparency, data governance, bias mitigation, human oversight, and strict adherence to privacy laws like the GDPR. For example, the Act requires descriptions of data sources, limitations, foreseeable risks, and mitigations for foundation models. These measures aim to ensure ethical AI development, prevent unauthorised data use, and protect individuals from bias and discrimination.

References

[1] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. palmerpenguins: Palmer archipelago (antarctica) penguin data, 2020. R package version 0.1.0.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] S. Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *ACM/IEEE 1st International Workshop on Gender Equality in*

Software Engineering, 2018.

[4] Rachel Goodman. Why amazon's automated hiring tool discriminated against women, 2018.

[5] Shiona McCallum. Clearview ai fined in uk for illegally storing facial images, 2022.

[6] E.-G. Breuer. Understanding the eu ai act as a framework for responsible innovation. https://www.cloudflight.io/en/blog/understanding-the-eu-ai-act-as-a-framework-for-responsible-innovation/, 2023.