Penguin Species Classification

Priyanshu Kumawat
Data Science & Artificial Intelligence
International Institute of Information Technology
Naya Raipur, India
priyanshuk21102@iiitnr.edu.in

Mrinal Bhan
Data Science & Artificial Intelligence
International Institute of Information Technology
Naya Raipur, India
mrinal21102@iiitnr.edu.in

***Abstract*—Penguins are a charismatic and ecologically important group of birds that inhabit the southern hemisphere. Accurately classifying penguin species is crucial for monitoring and managing their populations, but it can be challenging due to their similar physical appearances and overlapping ranges. In this project, we developed a machine learning model to classify six penguin species based on their physical features and habitat island. We used a dataset from Kaggle that included body measurements and habitat information for Adelie_Male, Adelie_Female, Chinstrap_Male, Chinstrap_Female, Gentoo_Male and Gentoo_Female penguins. We explored multiple machine learning algorithms, including multiple linear regression, logistic regression, k-means clustering, and k-nearest neighbor classification, to identify the most effective approach for classifying the penguin species.**

***Index Terms*—Penguins, species classification, regression, classification, clustering**

I. Introduction

A. Background Information

Penguins are a fascinating and iconic group of birds that inhabit the southern hemisphere, with many species facing numerous threats due to climate change, overfishing, and habitat destruction. Accurately classifying penguin species is essential for effective conservation efforts, as different species may have unique ecological requirements and face distinct challenges. However, penguin species classification can be challenging due to the birds' similar physical appearances and overlapping ranges. Traditional methods for identifying penguin species, such as manual observation or genetic testing, can be time-consuming and expensive, particularly in remote or inaccessible regions. In recent years, advances in machine learning and artificial intelligence have shown promise for classifying and monitoring penguin populations more efficiently and accurately.

B. Overview

In this project, we aimed to develop a machine learning model to classify six penguin species based on their physical features and habitat island. We explored multiple algorithms, including multiple linear regression, logistic regression, k-means clustering, and k-nearest neighbor classification, to identify the most effective approach for classifying the penguin species. Our results demonstrate the potential of machine learning for penguin species classification.

II. Dataset Description

A. Source

The dataset used in this project was obtained from Kaggle, a platform for hosting machine learning and data science datasets. The dataset includes physical body measurements and other variables. The variables measured for each penguin include culmen length, culmen depth, flipper length, body mass and the island where the penguin was observed. The dataset contains 344 observations.

B. Description of variables

The following is a description of the variables included in the dataset:

- species: The species of penguin
- island: The island where the penguin was observed, with three possible values: Biscoe, Dream, and Torgersen.
- culmen_length_mm: The length of the penguin's bill in millimeters.
- culmen_depth_mm: The depth of the penguin's bill in millimeters.
- flipper_length_mm: The length of the penguin's flipper in millimeters.
- body_mass_g: The mass of the penguin's body in grams.

C. Exploratory Data Analysis (EDA)

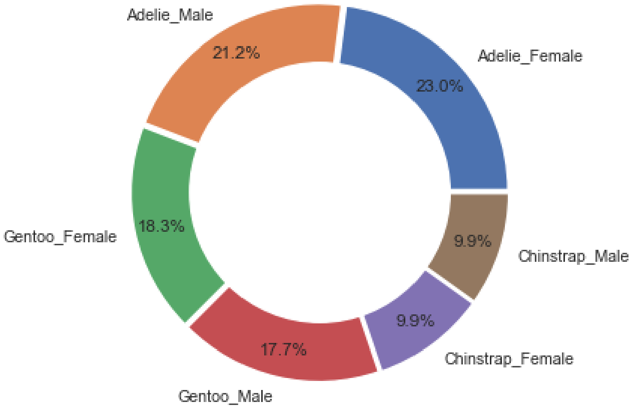


Fig. 1. Species Distribution in the dataset.

The donut chart above illustrates the distribution of penguin species in the dataset. Although it is evident that the classes are somewhat unbalanced, given the relatively small size of the dataset, we can proceed with our analysis without balancing the classes.

Data imputation is a method for retaining the majority of the dataset's data and information by substituting missing data with a different value. As the dataset contained few null values, we used the variable mean to replace them. After this imputation step, the dataset was free of null values. Table I below displays the statistical summary of the modified dataset.

TABLE I
STATISTICAL SUMMARY

Statistical values	Culmen length	Culmen depth	Flippers length	Body mass
count	344	344	344	344
mean	43.92	17.15	200.91	4201.61
std	5.443	1.969	14.02	799.608
min	32.1	13.1	172.0	2700.0
median	44.25	17.30	197.0	4050.0
max	59.6	21.5	231.0	6300.0

A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. They can tell us about the direction of the relationship, the form (shape) of the relationship, and the degree (strength) of the relationship between two variables.

$$r=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2 \sum\left(y_i-\bar{y}\right)^2}}$$

TABLE II
CORRELATION MATRIX

	Culmen length	Culmen depth	Flippers length	Body mass
culmen_length	1.000000	0.235053	0.656181	0.595149
culmen_depth	0.235053	1.000000	0.583851	0.472039
flipper_length	0.656181	0.583851	1.000000	0.871281
body_mass	0.595149	0.472039	0.871281	1.000000

To visualize the distribution of the different variables in the dataset, we used box plots. A box plot is a graphical representation of a dataset that shows the median, quartiles, and outliers. The box represents the interquartile range (IQR), which is the middle 50% of the data, while the whiskers extend to the minimum and maximum values within 1.5 times the IQR. Outliers, which are data points that fall outside of this range, are displayed as individual points. Box plots are useful for identifying the range, median, and skewness of a distribution, as well as for detecting potential outliers or extreme values. By using box plots, we were able to visualize the distribution of the different variables in the dataset and identify any potential outliers or unusual patterns.

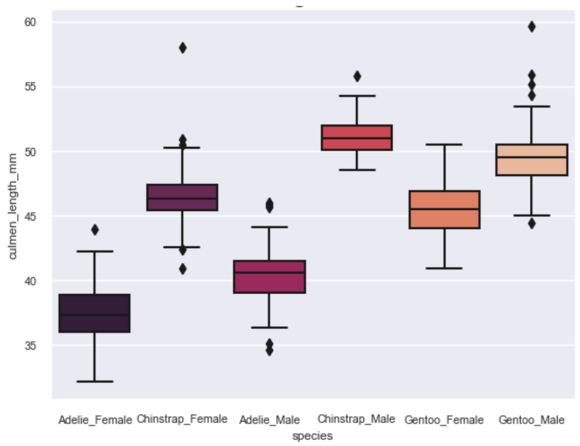


Fig. 2. Distribution of Culmen length by species.

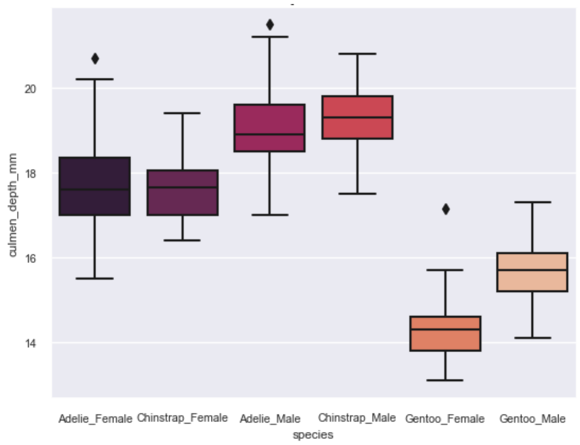


Fig. 3. Distribution of Culmen depth by species.

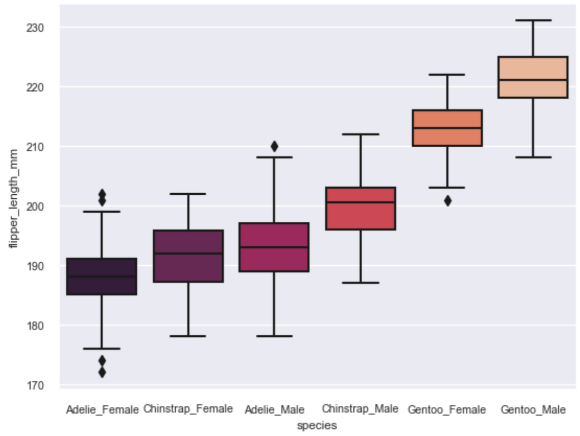


Fig. 4. Distribution of Flippers length by species.

Outliers can potentially have an impact on the analysis and the resulting classification models. Some ways in which outliers can affect the model are skewed distribution, biased classification, overfitting and hence reduce model performance.

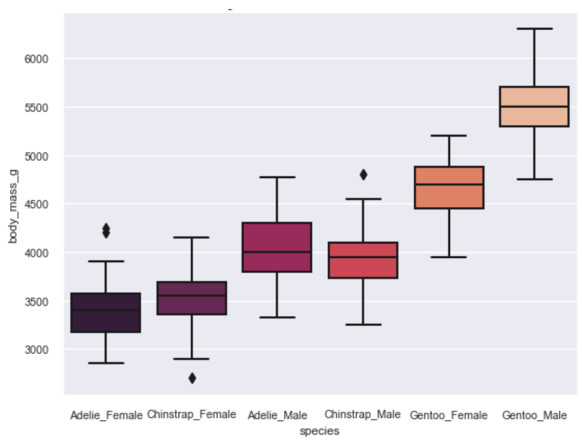


Fig. 5. Distribution of Body mass by species.

III. Methods

A. Data Preprocessing

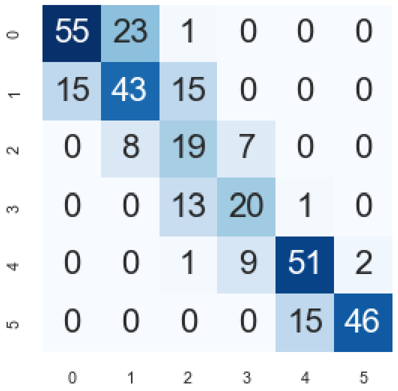
Data preprocessing was performed on the penguin dataset to ensure that the data was in a suitable format for analysis. This involved checking for and handling missing data, transforming or scaling the data , and encoding categorical variables. The preprocessing step is important to ensure that the data is clean, consistent, and appropriate for the analysis methods used.
Few commonly used scaling methods are -

- Standardization: This method scales the data to have a mean of 0 and a standard deviation of 1, so that the resulting distribution is a standard normal distribution.
- Min-max scaling: This method scales the data to a fixed range, usually between 0 and 1. This method preserves the relative relationships between the data points but can be sensitive to outliers.
- Robust scaling: This method is similar to standardization, but uses median and interquartile range instead of mean and standard deviation. This method is less sensitive to outliers than standardization.

In the case of penguin species classification, Robust Scaler and Standard Scaler will be preferred as there are features with high variance or outliers, as they are less sensitive to these factors than MinMax Scaler. Standard Scaler is also more suitable when the distribution of the features is not known or is not normal. Robust Scaler is preferred when the features have outliers or the data has a skewed distribution.

B. Machine Learning Models

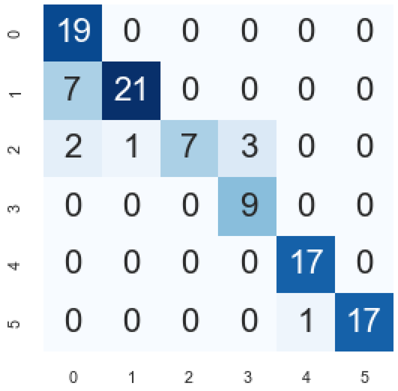
Multiple Linear Regression - Multiple linear regression is a statistical method used to model the relationship between two or more predictor variables and a response variable. It assumes that the response variable is continuous and normally distributed, and that the relationship between the predictor variables and the response variable is linear.



	0	1	2	3	4	5
0	55	23	1	0	0	0
1	15	43	15	0	0	0
2	0	8	19	7	0	0
3	0	0	13	20	1	0
4	0	0	1	9	51	2
5	0	0	0	0	15	46

Fig. 6. Confusion Matrix

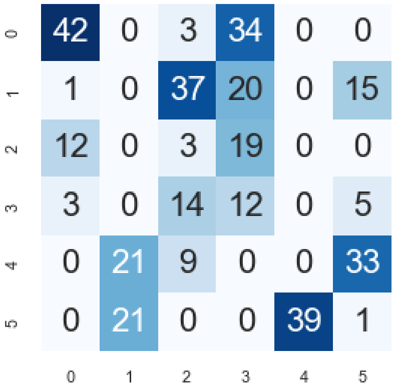
Multinomial Logistic Regression - Multinomial Logistic Regression is a statistical method used to analyze relationships between a categorical dependent variable with two or more categories, and multiple independent variables. In Multinomial Logistic Regression, the dependent variable has three or more categories. The aim is to predict the probabilities of each category given a set of independent variables. The method uses the maximum likelihood estimation to estimate the parameters of the model. The coefficients of the independent variables represent the effect of each variable on the probability of being in a particular category. The output of Multinomial Logistic Regression includes coefficients for each independent variable, which indicate the direction and strength of the relationship between the variable and the outcome. These coefficients are estimated by maximizing the likelihood function.



	0	1	2	3	4	5
0	19	0	0	0	0	0
1	7	21	0	0	0	0
2	2	1	7	3	0	0
3	0	0	0	9	0	0
4	0	0	0	0	17	0
5	0	0	0	0	1	17

Fig. 7. Confusion Matrix

K-means Clustering Algorithm - K-means clustering is a unsupervised machine learning technique used for clustering similar data points into k number of clusters. The algorithm starts by selecting k number of random centroids for each cluster, then assigning each data point to its nearest centroid, based on the distance measure. Then, the centroid of each cluster is recomputed based on the new data points assigned to it. This process is repeated until the centroids no longer move, or a maximum number of iterations is reached.



	0	1	2	3	4	5
0	42	0	3	34	0	0
1	1	0	37	20	0	15
2	12	0	3	19	0	0
3	3	0	14	12	0	5
4	0	21	9	0	0	33
5	0	21	0	0	39	1

Fig. 8. Confusion Matrix

IV. Results and Conclusion

A. Evaluation Metrics

- Accuracy: measures the percentage of correctly classified instances out of all instances in the dataset.

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

- Precision: measures the proportion of true positives (correctly predicted positive instances) out of all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- Recall: measures the proportion of true positives (correctly predicted positive instances) out of all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- F1-score: harmonic mean of precision and recall. It takes into account both precision and recall, and is a useful metric when there is an imbalance between the number of positive and negative instances in a dataset. The F1 score ranges between 0 and 1, with 1 indicating perfect precision and recall.

$$F1 \text{ Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

B. Comparing the models

Evaluating machine learning models is a critical step in the model development process as it allows you to measure the model's performance and determine how well it is able to make accurate predictions on new, unseen data.

TABLE III
Performance Comparison

	Accuracy	Precision	Recall	F1-Score
Linear Regression	68.02	0.671	0.666	0.663
Logistic Regression	86.54	0.8879	0.872	0.858
K-means Clustering	21.8	0.199	0.2175	0.2072

C. Conclusion

In conclusion, this project focused on the classification of penguin species based on their physical characteristics and habitat island. Three machine learning techniques, namely multiple linear regression, logistic regression, and K-means clustering, were applied to classify penguin species. Additionally, K-means clustering was used for unsupervised clustering analysis.

The evaluation metrics were used to compare the performance of these models, and it was found that the Logistic Regression outperformed the other models with an F1 score of 0.858. Furthermore, it was observed that the accuracy of the K-means clustering model was low, which might be attributed to the high variance in the dataset.

Overall, this project demonstrates the importance of choosing an appropriate machine learning technique for classification tasks and the significance of data preprocessing in improving model performance.

References

[1] Dataset
https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data

[2] K-means clustering, K-neighbors classification
https://www.kaggle.com/code/ayushikaushik/k-means-clustering-k-neighbors-classification

[3] Performance of model
https://scikit-learn.org/stable/modules/model_evaluation.html