University of Asia Pacific

Report No: 03

Report Name: Penguin Species Classification via Logistic Regression

Submitted to

Noor Mairukh Khan Arnob

Lecturer

Department of Computer Science & Engineering

Submitted by

Md. Arifur Rahman Akash

Roll-21201091, Section – B2, 52nd Batch,

Department of Computer Science & Engineering

Course Title: Artificial Intelligence and Expert Systems Lab

Course Code: CSE 404

Problem Title: Classify penguin species based on physical features using machine learning.

Problem Description: The goal of this project is to build a machine learning model to classify three species of penguins based on physical measurements such as bill length, flipper length, body mass, and more.
The classification is performed using a Logistic Regression algorithm.

Tools and Languages:

- **Programming Language**: Python 3
- **Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, palmerpenguins
- **Algorithm**: Logistic Regression
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1-Score

Dataset Description:
The Palmer Penguins dataset, collected by Dr. Kristen Gorman and made available through the palmerpenguins package, contains 344 samples.
Features include:

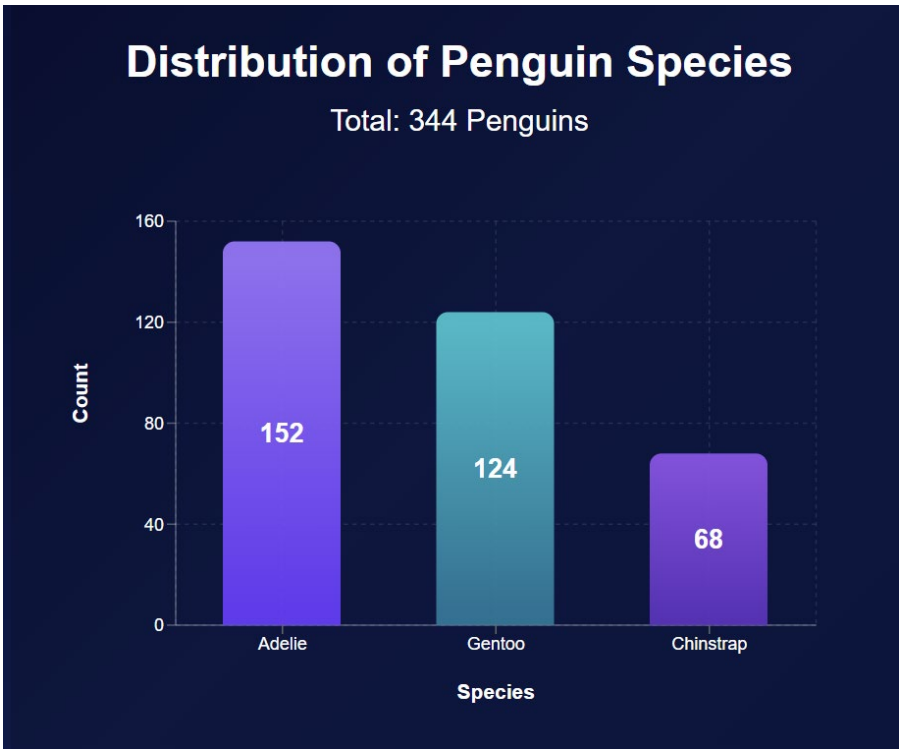
- Bill Length (mm)
- Bill Depth (mm)
- Flipper Length (mm)
- Body Mass (g)
- Sex
- Island
- Year

Target Variable: Species
Missing Values: Present in some measurement fields and categorical features (handled during preprocessing).

Exploratory Data Analysis (EDA):

1. Species Distribution

- We first examined the distribution of the penguin species.

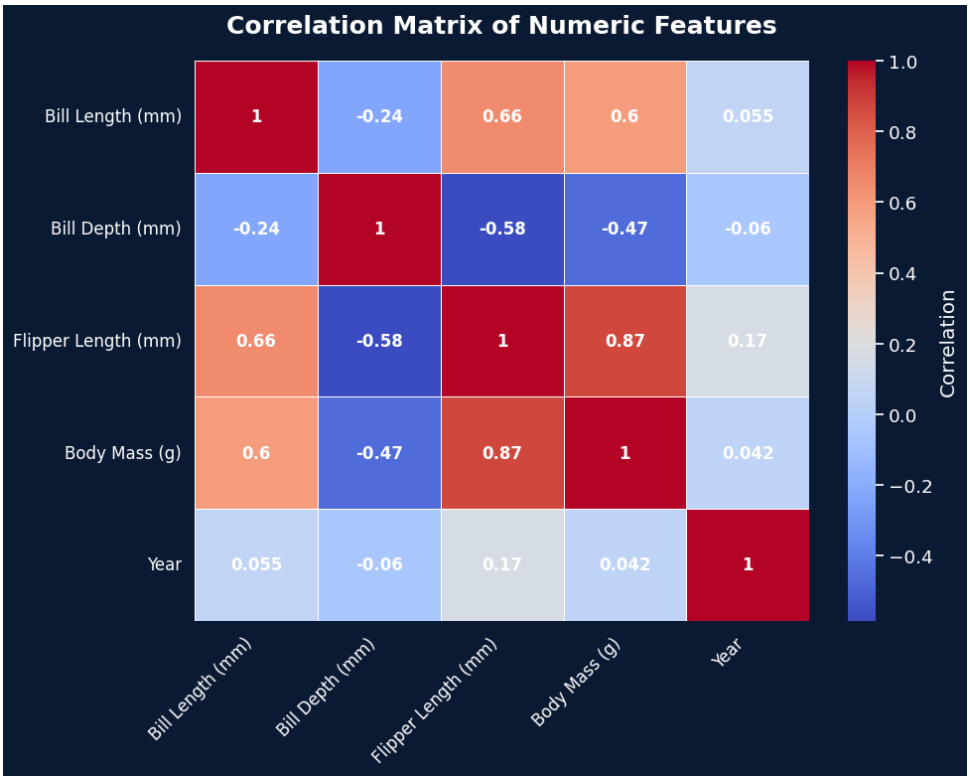


2. Physical Characteristics by Species

- We analyzed how physical measurements vary across species.

3. Correlation Analysis

- Correlation heatmap reveals strong relationships between features.



Data Preprocessing:

Missing Values: Rows with missing values were dropped.

Feature Engineering:

- Numeric Features: Median imputation and StandardScaler.
- Categorical Features: Most-frequent imputation and OneHotEncoder.

Pipeline: All preprocessing combined into a ColumnTransformer.

Model Training:

Data Split: 70% Training, 30% Testing (with stratification).

Classifier: Logistic Regression (random_state=42, max_iter=1000).

Pipeline: Preprocessing and model training were combined into a single pipeline.

Model Evaluation:

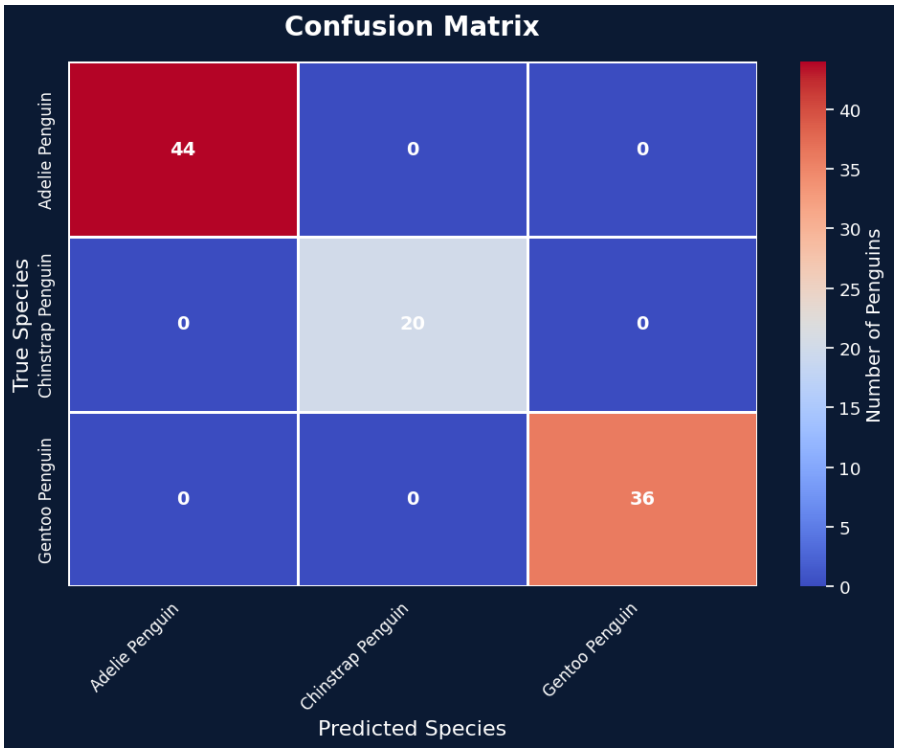
1. Performance Metrics

The trained model achieved 100% accuracy on the test set.

Class	Precision	Recall	F1-Score	Support
Adelie	1.00	1.00	1.00	44
Chinstrap	1.00	1.00	1.00	20
Gentoo	1.00	1.00	1.00	36
accuracy			1.00	100
macro avg	1.00	1.00	1.00	100
weighted avg	1.00	1.00	1.00	100

2. Confusion Matrix

The confusion matrix confirms no misclassifications.



Feature Importance Analysis:

We examined feature importance based on Logistic Regression coefficients for each species.

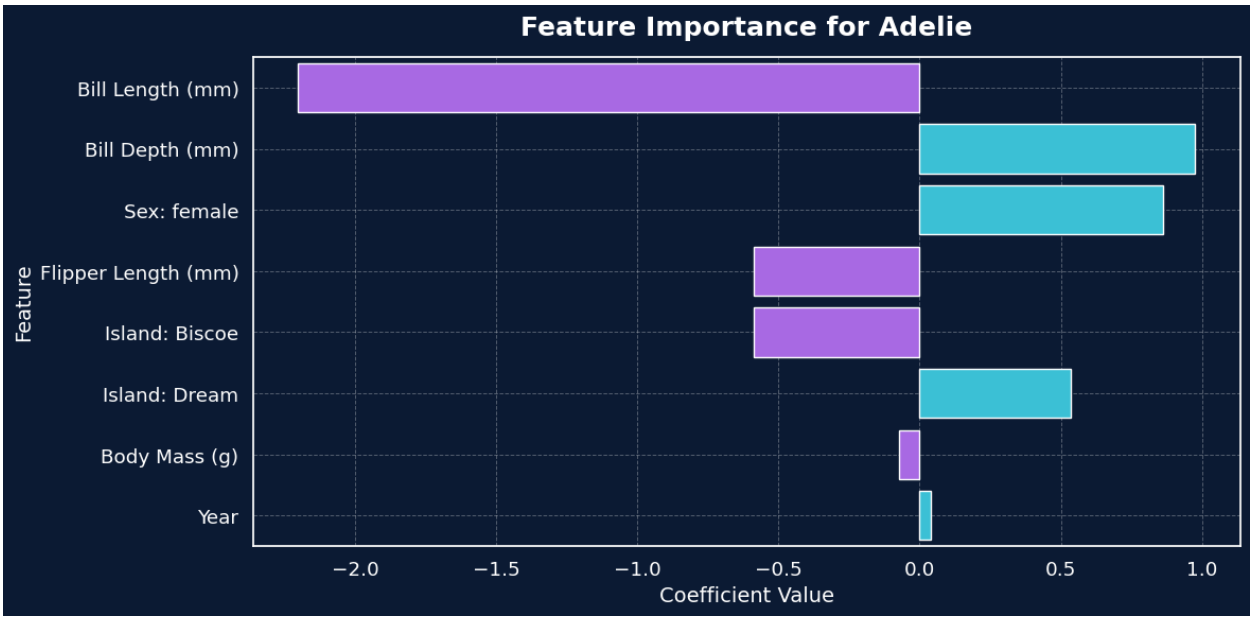


Fig: Feature importance for classifying Adelie Penguins.

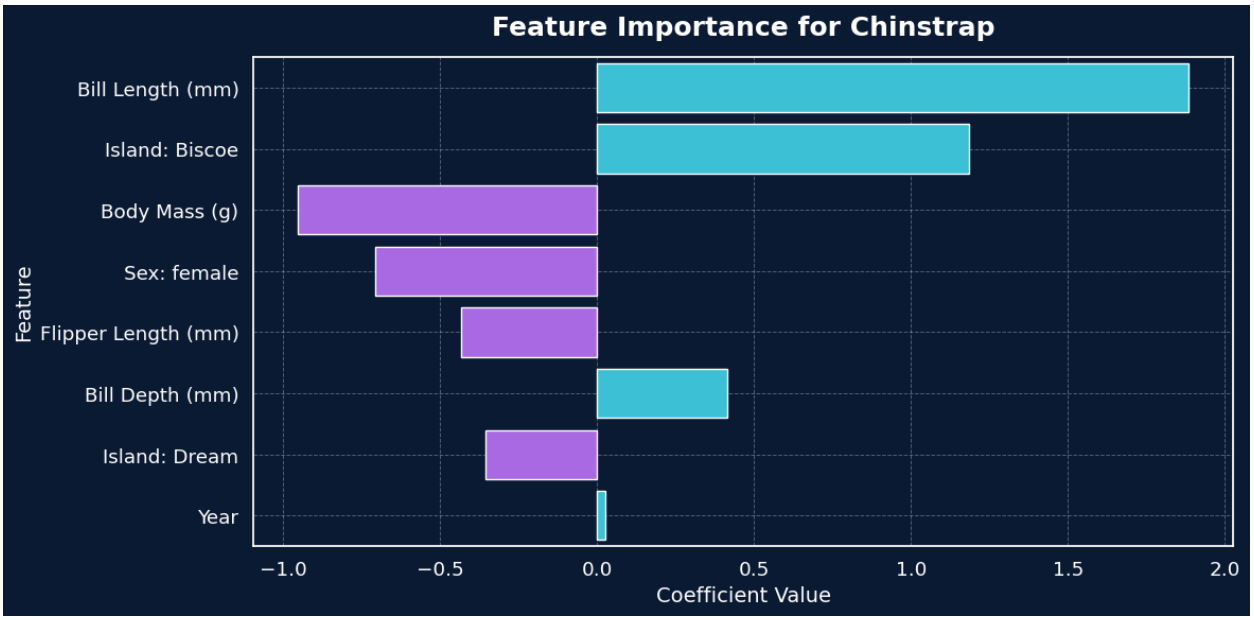


Fig: Feature importance for classifying Chinstrap Penguins.

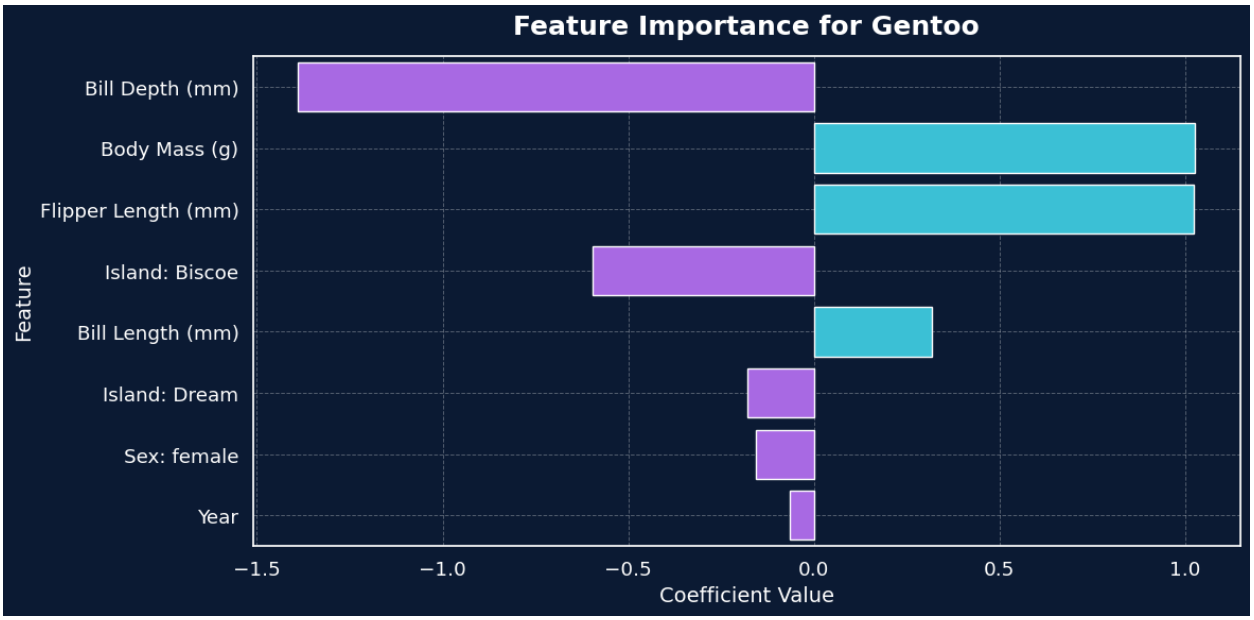


Fig: Feature importance for classifying Gentoo Penguins.

Conclusion

The Logistic Regression model perfectly classified the penguin species based on physical measurements.

It demonstrated that simple, interpretable models can achieve high accuracy when features are well-separated.

Challenges

- Handling missing data without introducing bias.
- Balancing train-test split to maintain class distribution.
- Correctly preprocessing numeric and categorical features.
- Interpreting feature importance in a multi-class logistic regression setup.

Source Code