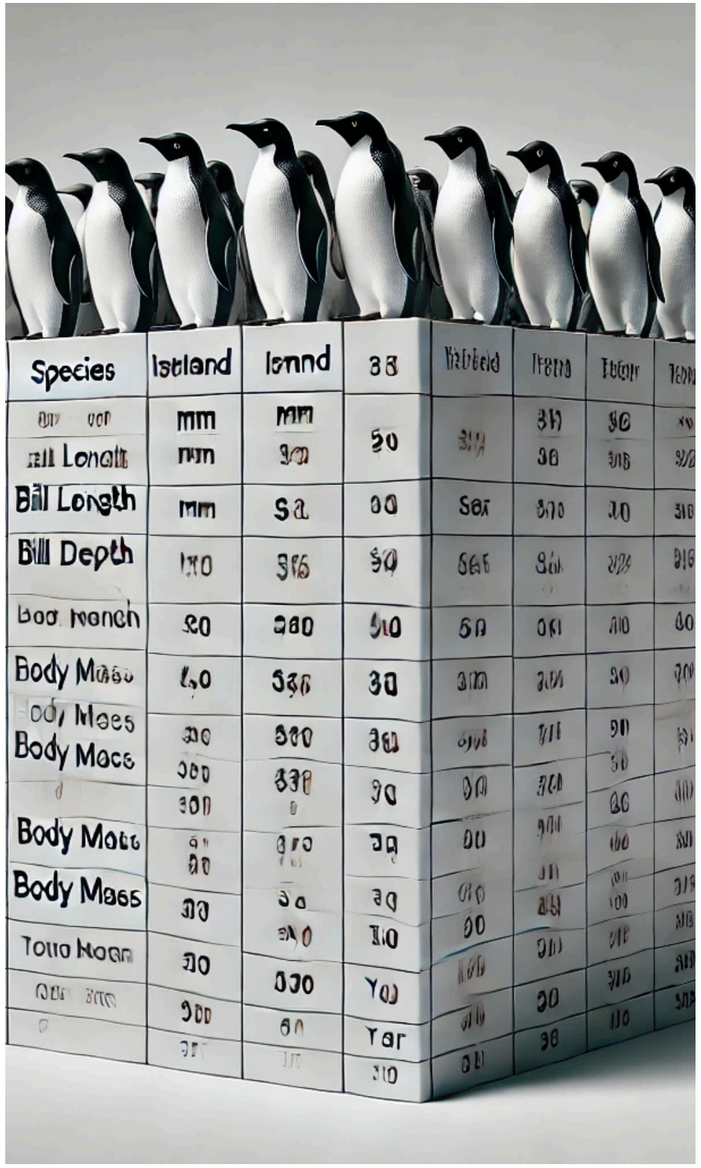
CSC3206 Artificial Intelligence

Palmer Penguin Classification



Brief overview of the dataset

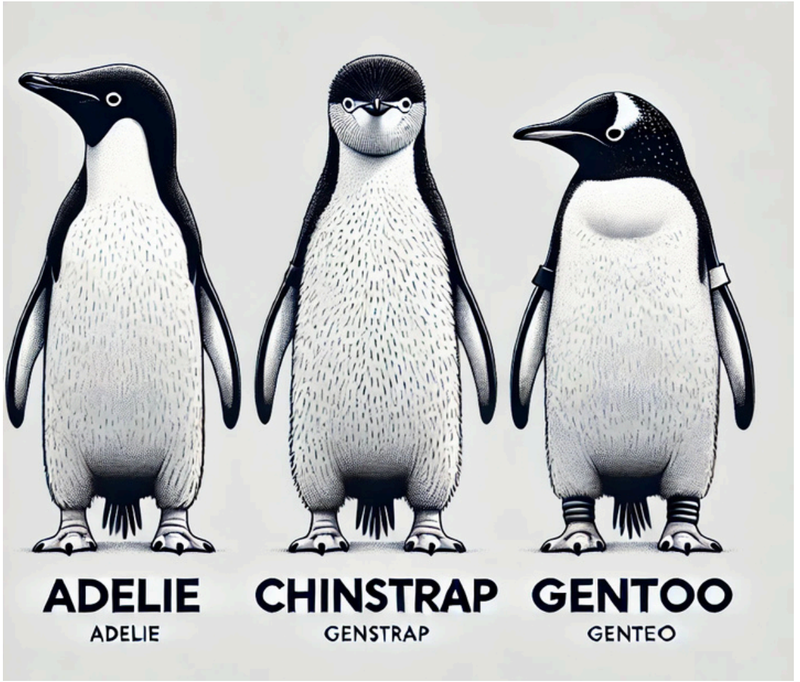
- The Palmer Penguins dataset is a biological dataset collected as part of ecological research in the Palmer Archipelago, Antarctica.
- It serves as an alternative to the well-known Iris dataset, often used for statistical and machine learning practices.

Source

- The dataset is sourced from the Palmer Archipelago Penguin Monitoring Project.

Features and Variables

- **Species**: Three species of penguins: Adelie, Chinstrap, and Gentoo.
- **Island**: Locations in the Palmer Archipelago (Biscoe, Dream, Torgersen).
- **Physical Measurements:**
  - Bill length and depth (in mm).
  - Flipper length (in mm).
  - Body mass (in grams).
- **Sex**: Male or female.
- **Year**: Year of observation (2007-2009).



Purpose of the Dataset

- Designed for exploratory data analysis and teaching machine learning concepts.
- Ideal for visualizing relationships and statistical patterns in ecological data.
- Unlike traditional datasets, it provides a realistic yet straightforward introduction to ecological data analysis.

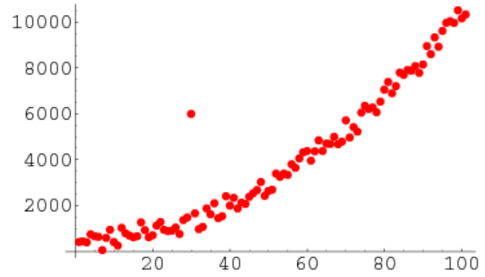
Exploratory Data Analysis

What is it?

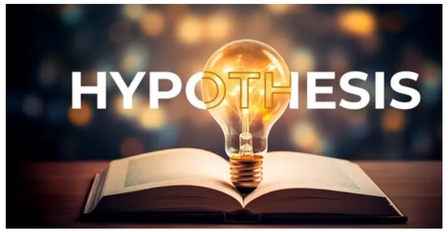
- discover patterns -
- to spot anomalies -
- to test hypothesis -
- to check assumptions

Why is EDA important?

- Understand the data set



- Identify revelant patterns in the dataset & Identify any Anomalies and Outliers



- Verify Hypothesis

- Gain insights about potential features

These 2 pictures display how many records are in the dataset

```
[4]: (344, 8)
```

Full Dataset:

[240]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN	2007
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007
...	...	...	...	...	...	...	...	...
339	Chinstrap	Dream	55.8	19.8	207.0	4000.0	male	2009
340	Chinstrap	Dream	43.5	18.1	202.0	3400.0	female	2009
341	Chinstrap	Dream	49.6	18.2	193.0	3775.0	male	2009
342	Chinstrap	Dream	50.8	19.0	210.0	4100.0	male	2009
343	Chinstrap	Dream	50.2	18.7	198.0	3775.0	female	2009

344 rows × 8 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   bill_length_mm     342 non-null    float64
 3   bill_depth_mm      342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                333 non-null    object 
 7   year               344 non-null    int64  
dtypes: float64(4), int64(1), object(3)
memory usage: 21.6+ KB
```

The meta data of the table is dispalyed

[6]:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	342.000000	342.000000	342.000000	342.000000	344.000000
mean	43.921930	17.151170	200.915205	4201.754386	2008.029070
std	5.459584	1.974793	14.061714	801.954536	0.818356
min	32.100000	13.100000	172.000000	2700.000000	2007.000000
25%	39.225000	15.600000	190.000000	3550.000000	2007.000000
50%	44.450000	17.300000	197.000000	4050.000000	2008.000000
75%	48.500000	18.700000	213.000000	4750.000000	2009.000000
max	59.600000	21.500000	231.000000	6300.000000	2009.000000

The dataset is summarsied

```
Class distribution:
species
Adelie       152
Gentoo       124
Chinstrap     68
Name: count, dtype: int64
```

```
[8]: species              0
     island               0
     bill_length_mm       2
     bill_depth_mm        2
     flipper_length_mm    2
     body_mass_g          2
     sex                 11
     year                 0
     dtype: int64
```

The dataset is sorted by their species

Cleaning the Dataset

All empty rows are dropped

Classification is used for this data set

This is how the cleaned dataset looks like

```
(333, 8)
```

[248]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	male	2007
...	...	...	...	...	...	...	...	...
339	Chinstrap	Dream	55.8	19.8	207.0	4000.0	male	2009
340	Chinstrap	Dream	43.5	18.1	202.0	3400.0	female	2009
341	Chinstrap	Dream	49.6	18.2	193.0	3775.0	male	2009
342	Chinstrap	Dream	50.8	19.0	210.0	4100.0	male	2009
343	Chinstrap	Dream	50.2	18.7	198.0	3775.0	female	2009

333 rows × 8 columns

Summary of the cleaned dataset

[249]:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	333.000000	333.000000	333.000000	333.000000	333.000000
mean	43.992793	17.164865	200.966967	4207.057057	2008.042042
std	5.468668	1.969235	14.015765	805.215802	0.812944
min	32.100000	13.100000	172.000000	2700.000000	2007.000000
25%	39.500000	15.600000	190.000000	3550.000000	2007.000000
50%	44.500000	17.300000	197.000000	4050.000000	2008.000000
75%	48.600000	18.700000	213.000000	4775.000000	2009.000000
max	59.600000	21.500000	231.000000	6300.000000	2009.000000

The dataset is classified again into their species but there are no missing values

```
 Updated Class distribution:
species
Adelie       146
Gentoo       119
Chinstrap     68
Name: count, dtype: int64
```

[251]:

	0
species	0
island	0
bill_length_mm	0
bill_depth_mm	0
flipper_length_mm	0
body_mass_g	0
sex	0
year	0

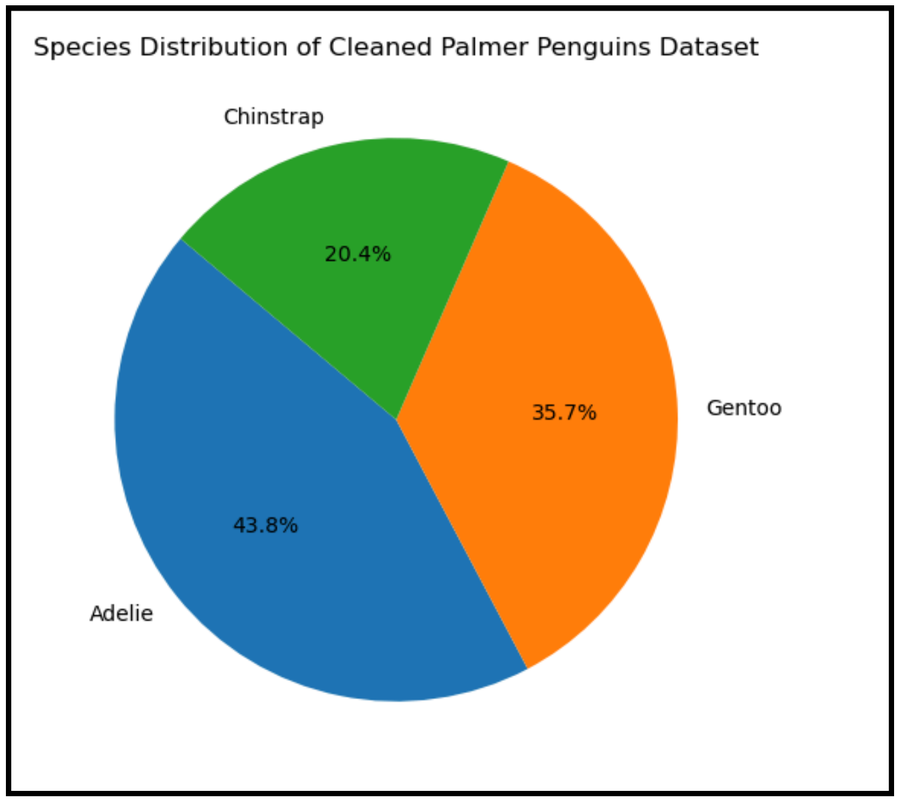
dtype: int64

EDA Visualizations

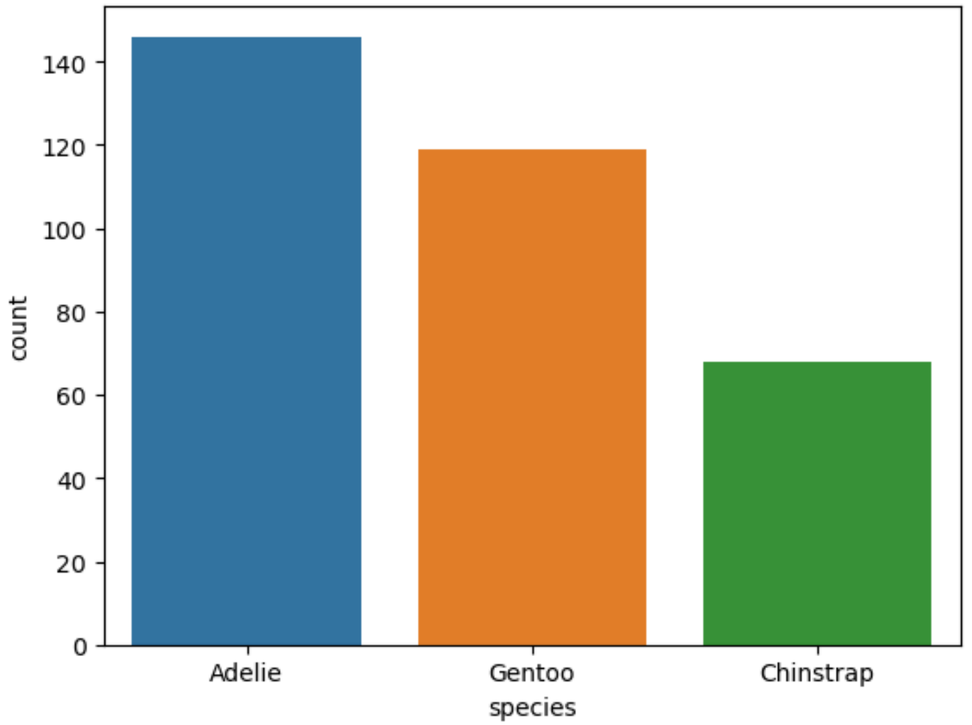
The year column removed

Filter for Adelie, gentoo and Chinstrap penguins are applied

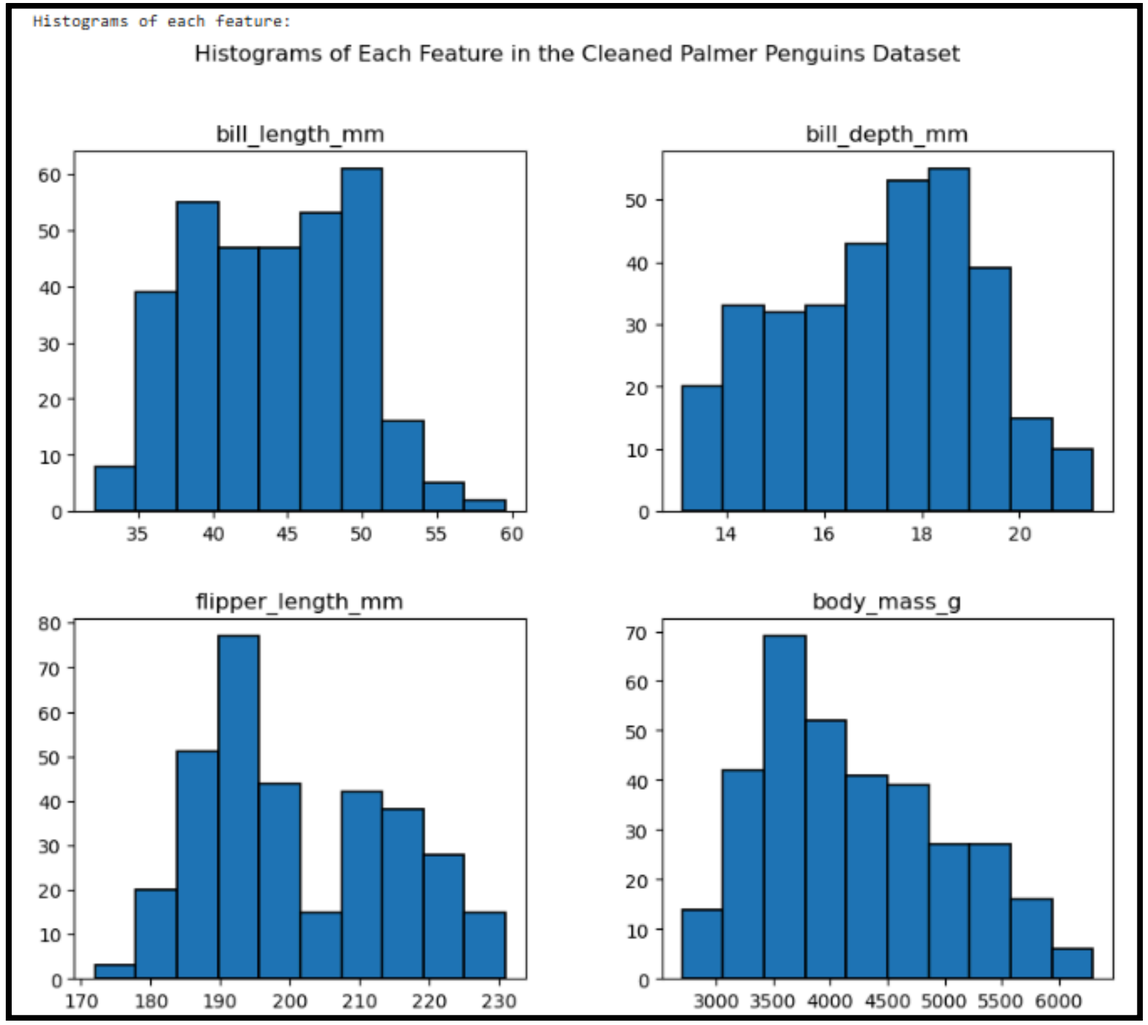
Class Distribution Pie Chart:

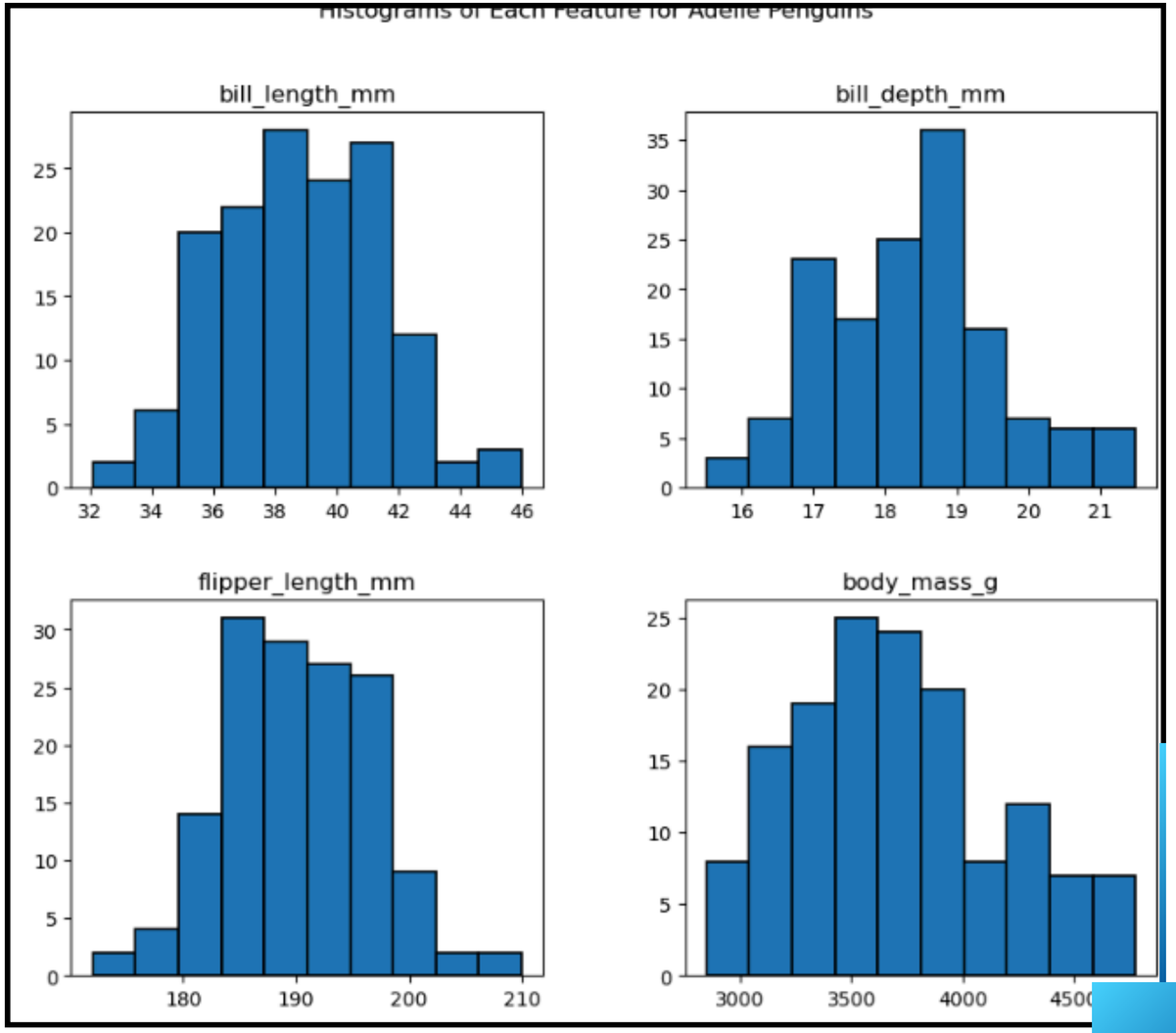


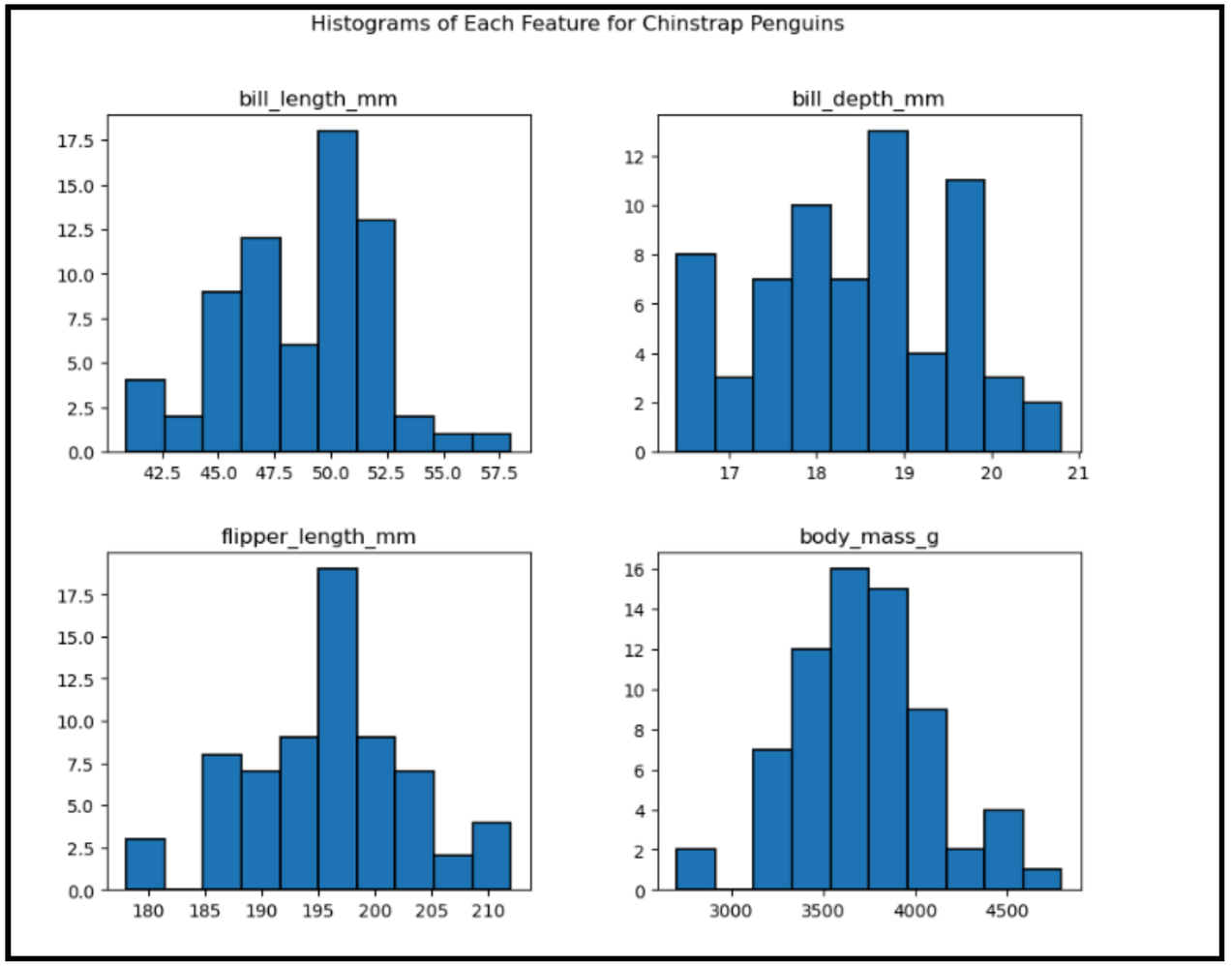
Class Distribution Count

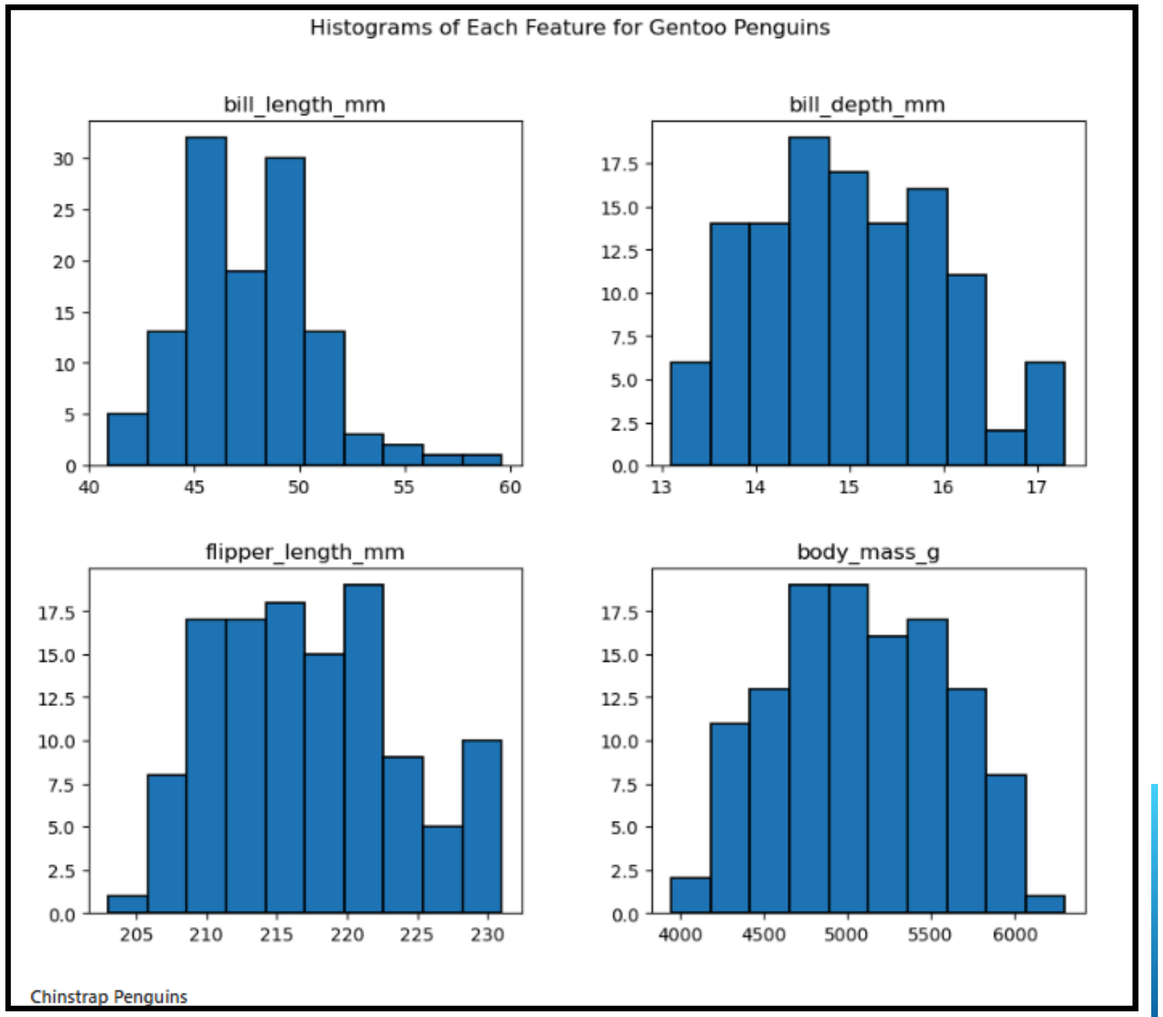


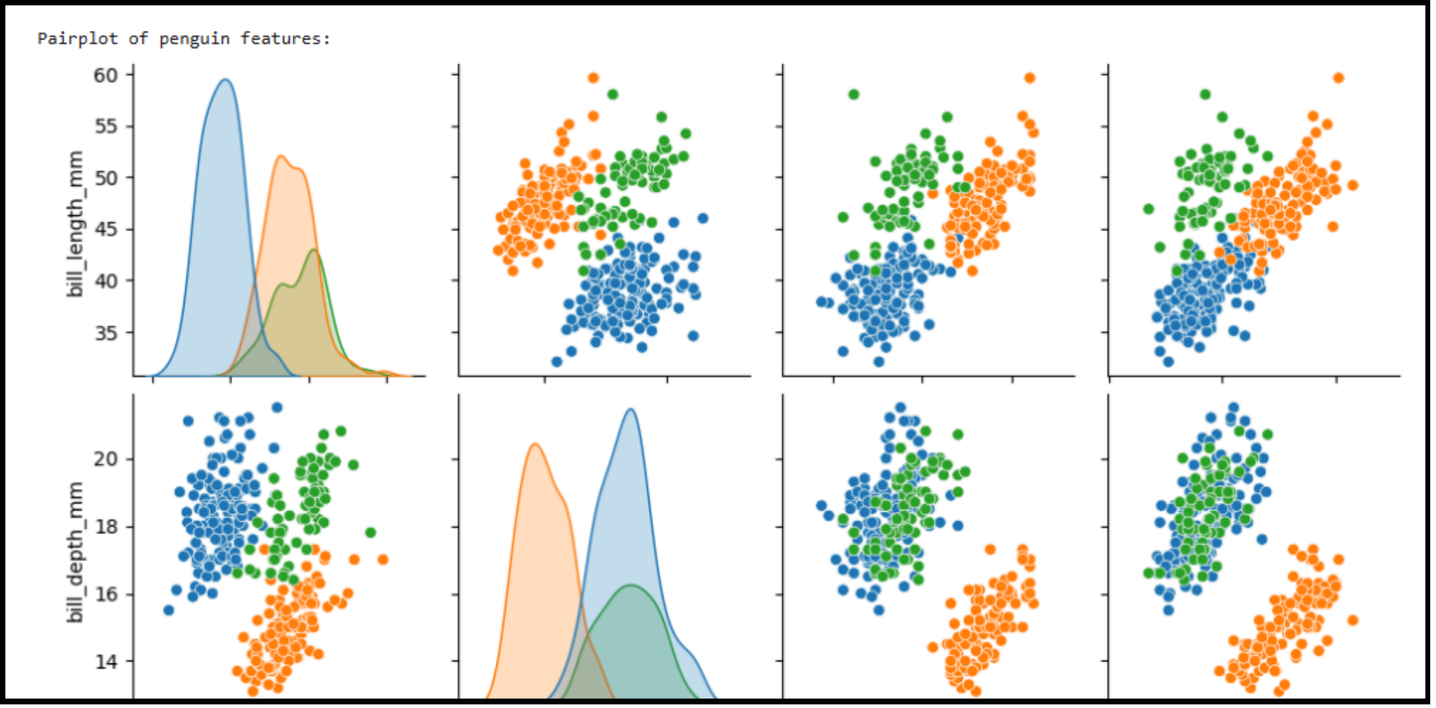
Histograms

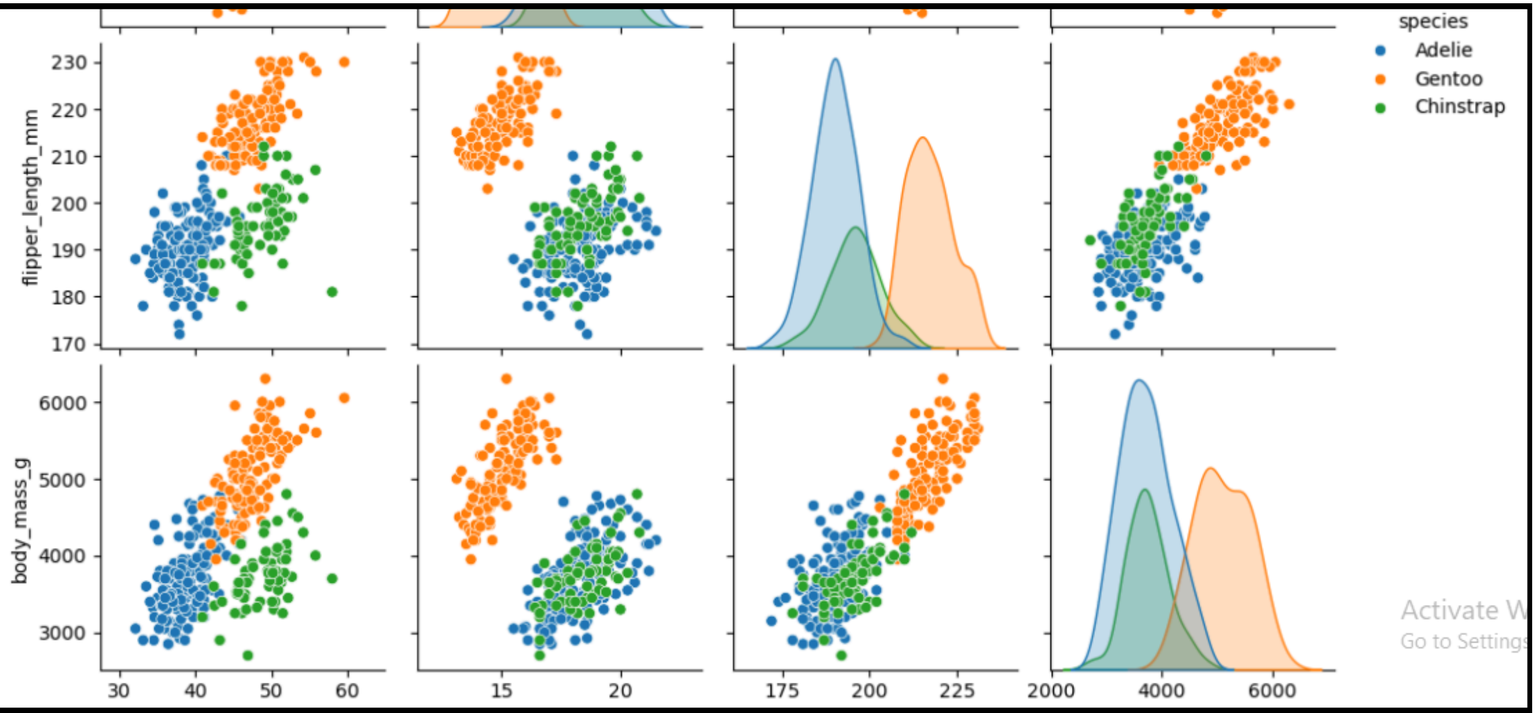

Histograms of each feature:
Histograms of Each Feature in the Cleaned Palmer Penguins Dataset
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g


bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g


Histograms of Each Feature for Chinstrap Penguins
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g

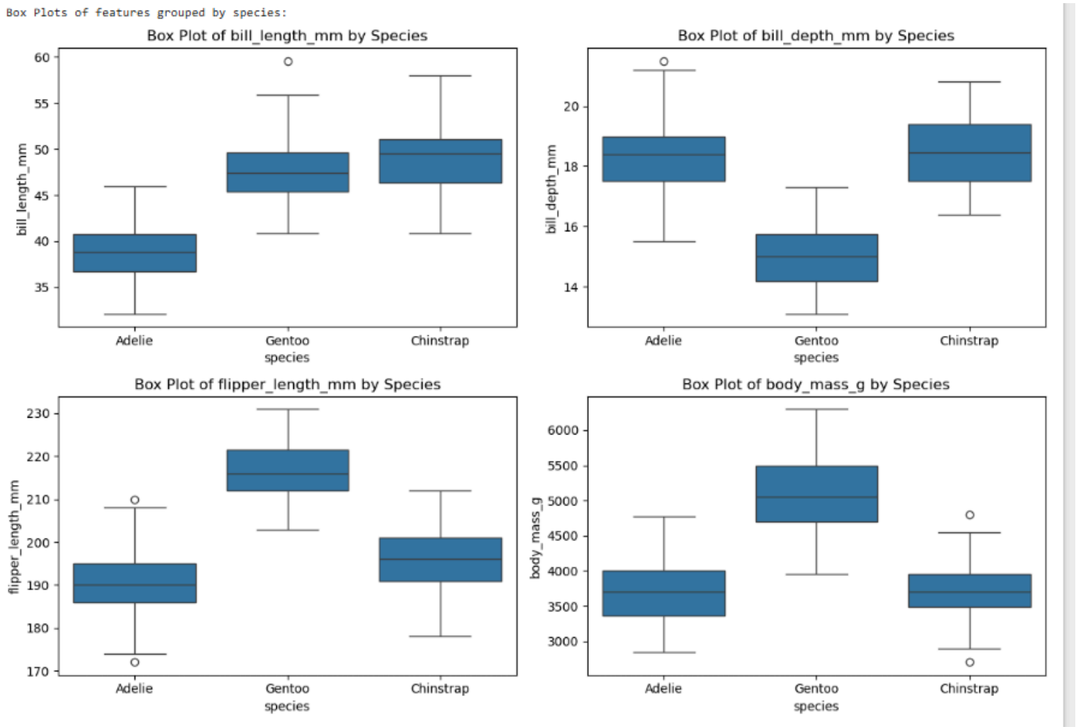

Histograms of Each Feature for Gentoo Penguins
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
Chinstrap Penguins



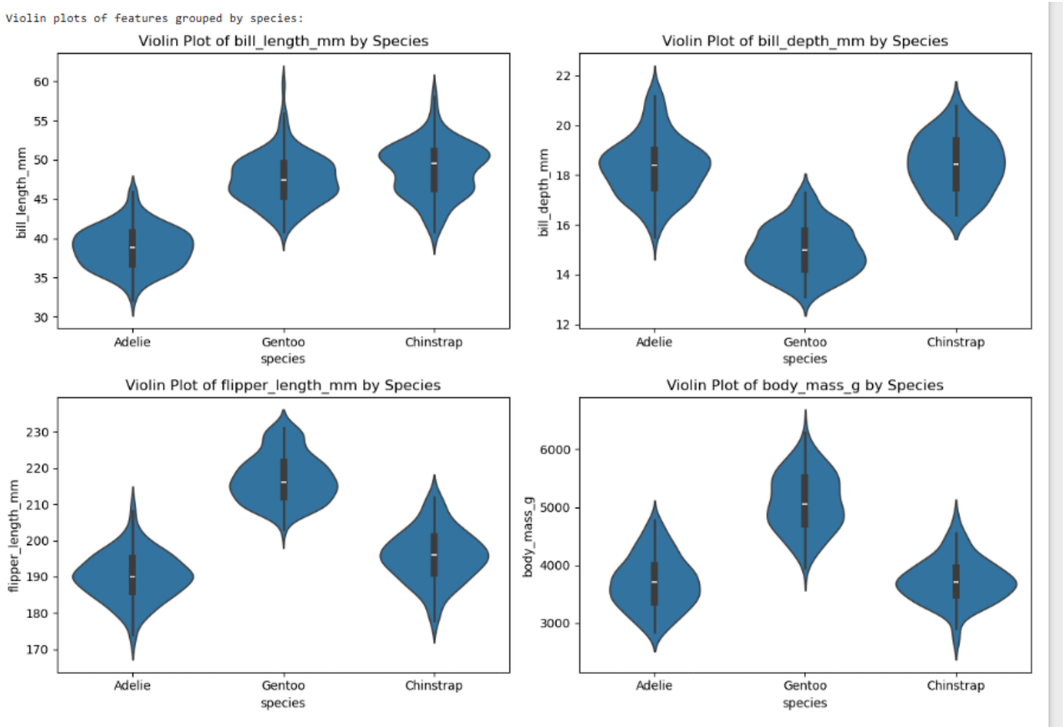


Pairplot of penguin features:

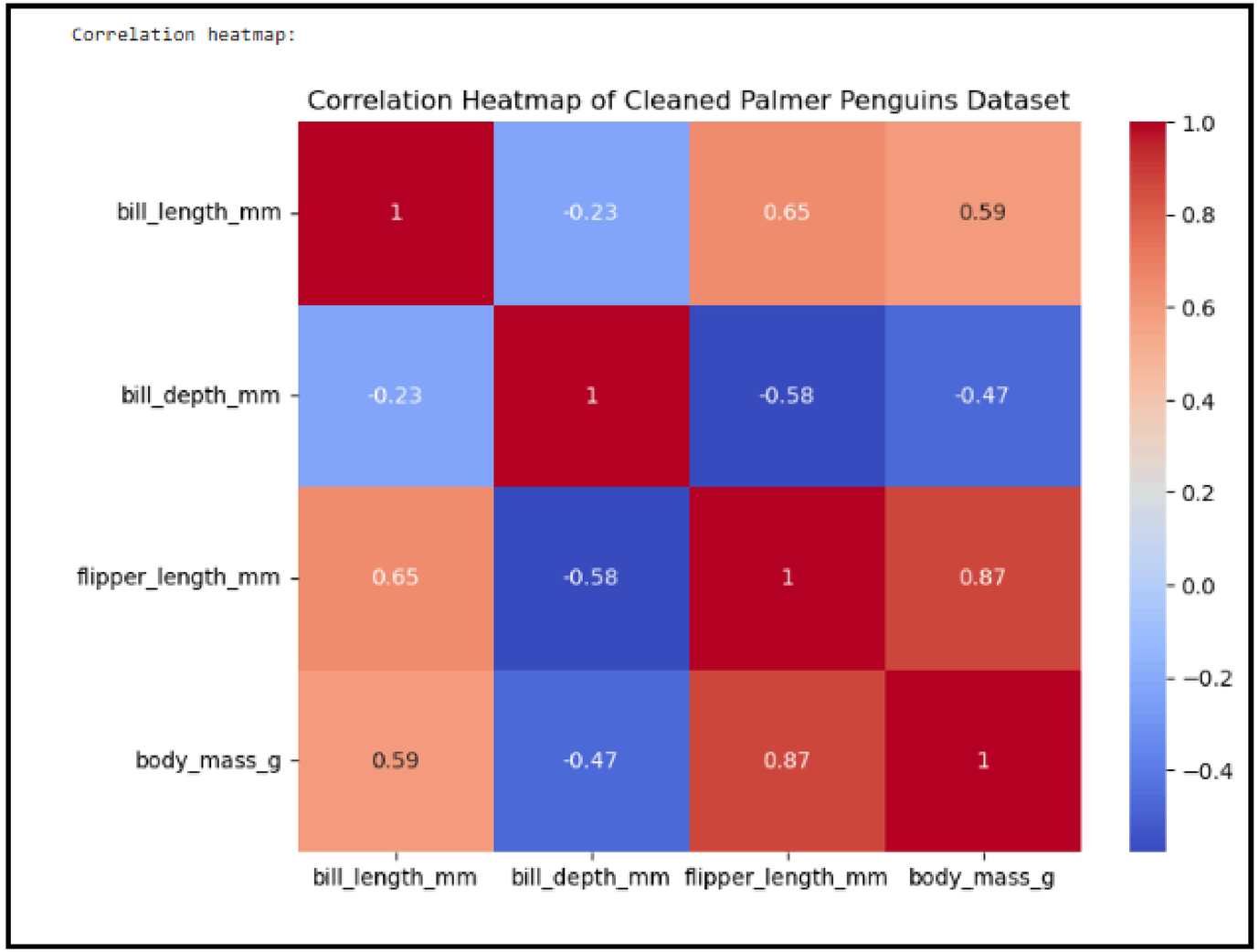
Box Plots of features grouped by species:



Violin plots of features grouped by species:



Correlation heatmap:



Correlation heatmap:

Methodology

Decision Tree

KNN

Why Decision Tree & KNN?

Goal

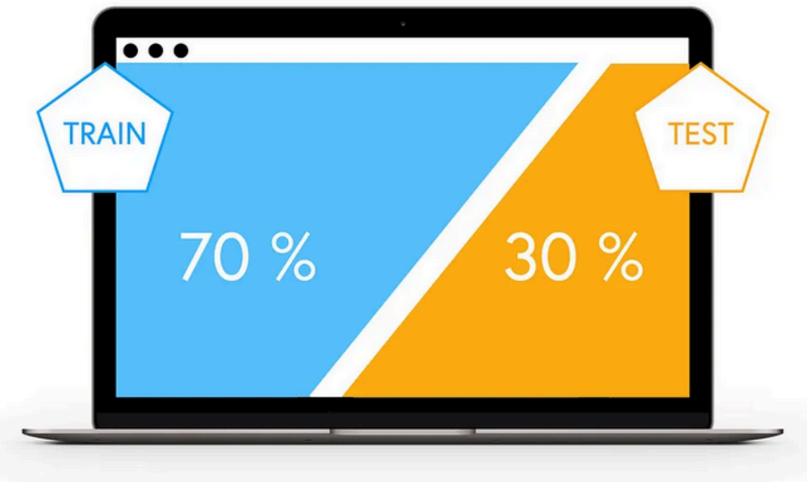
To **identify** the penguin species

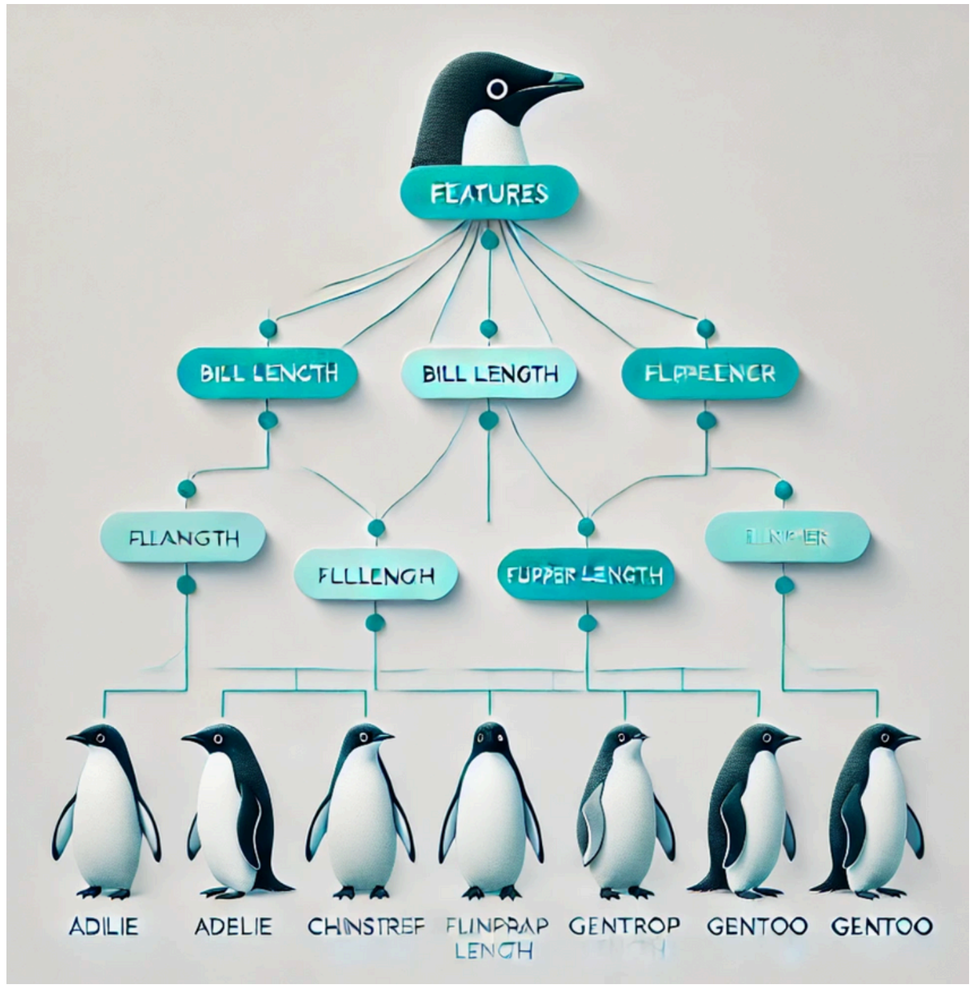
Suitability

- These algorithms are specifically designed for classification and can handle categorical target variables
- Target variable in the Palmer Penguins dataset is **categorical** (Species)
- Regression is not suitable as is used to predict **continuous** numerical values

Preparation

1. Reorganize the cleaned penguin Dataset to a dictionary structure for easier access to the specific required data
    - **Attributes:** Penguin features Eg. Bill length, Flipper length, etc.
        - Purpose: Store input used to identify the target
    - **Target:** Penguin Species Column of Dataset
        - Purpose: Target used to compare predicted vs actual
    - **Target Name:** Unique specie names (Adelie, Gentoo, or Chinstrap)
        - Purpose: To display the penguin types in plots
2. Train-Test Split
    - Split data into training and testing sets
    - Training Set = 70% Testing Set = 30%
    - x_train, x_test : Penguin features
    - y_train, y_test: Target (Species)





Decision Tree

Recursively splits the data based on features to create a tree-like structure.

Gini Impurity Measure is used to select the best features for splitting the tree

fit() method used to construct the decision tree, by selecting the best feature until the stopping condition is met

predict() method traverses the tree until a leaf node has been reached

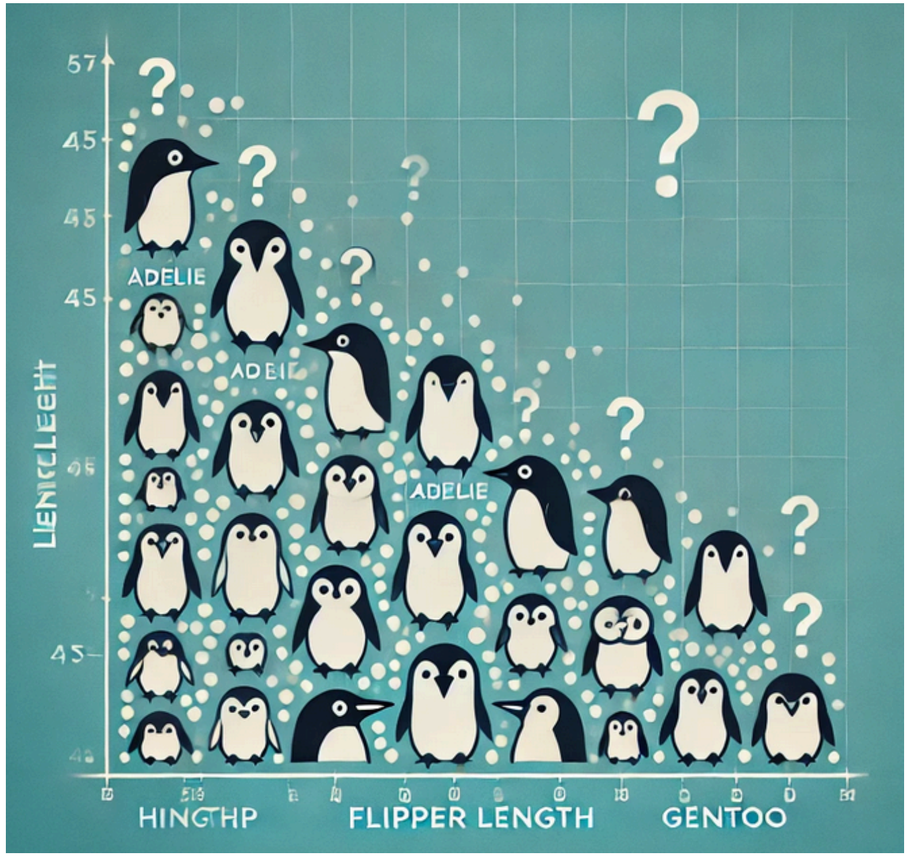
K-Nearest Neighbours (KNN)

Uses majority class of proximity to make predictions about data points

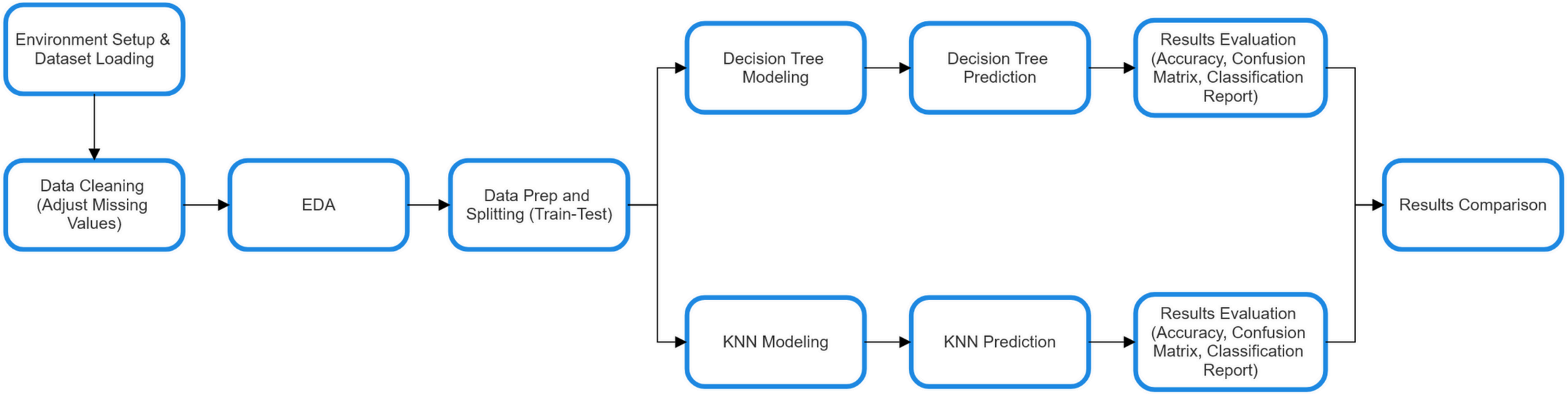
Set k value as k=5

fit() method used to store the training data in the plot

predict() method used to predict the species of a new penguin based on distance and k value



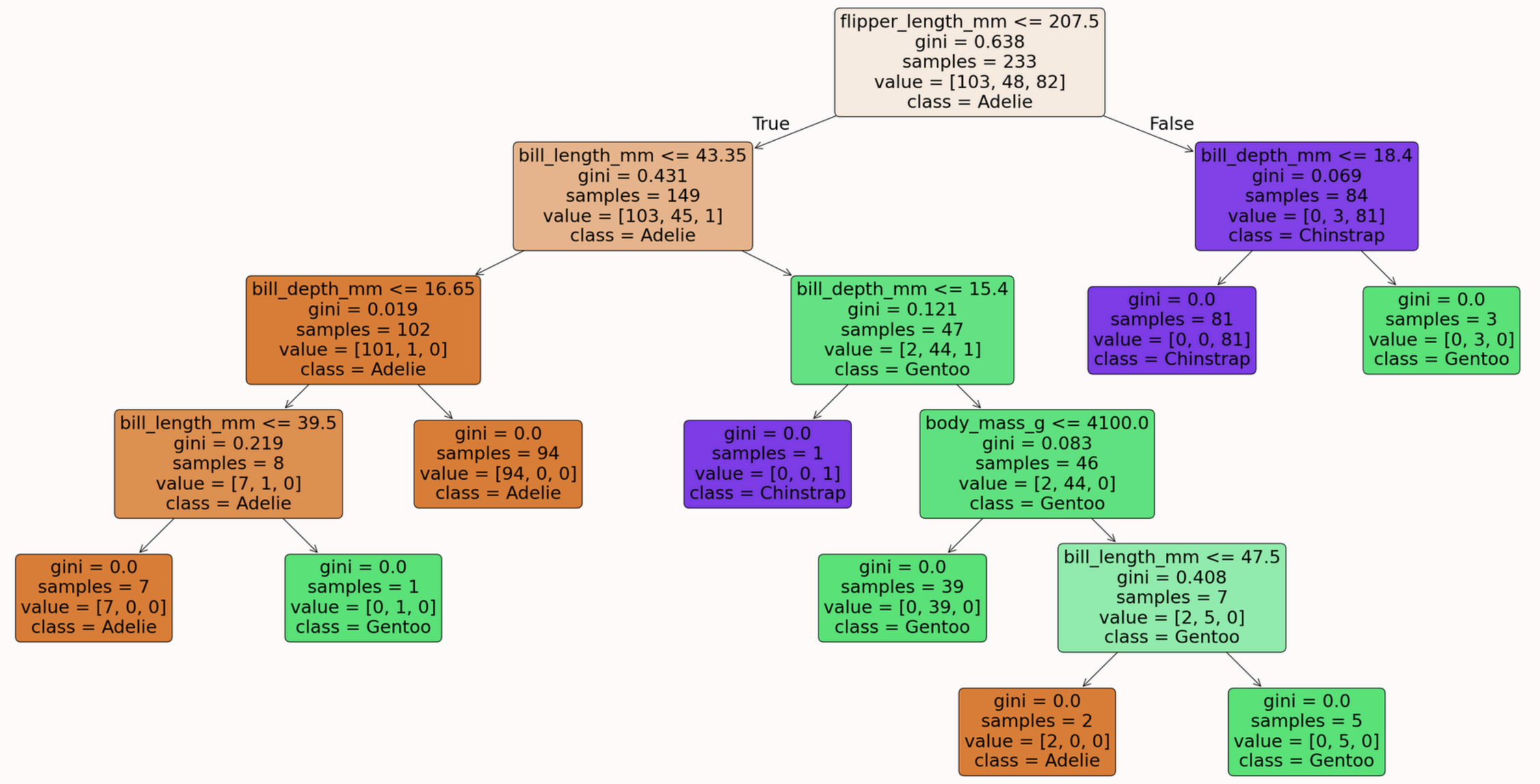
Flowchart


Environment Setup & Dataset Loading
Data Cleaning (Adjust Missing Values)
EDA
Data Prep and Splitting (Train-Test)
Decision Tree Modeling
Decision Tree Prediction
Results Evaluation (Accuracy, Confusion Matrix, Classification Report)
KNN Modeling
KNN Prediction
Results Evaluation (Accuracy, Confusion Matrix, Classification Report)
Results Comparison

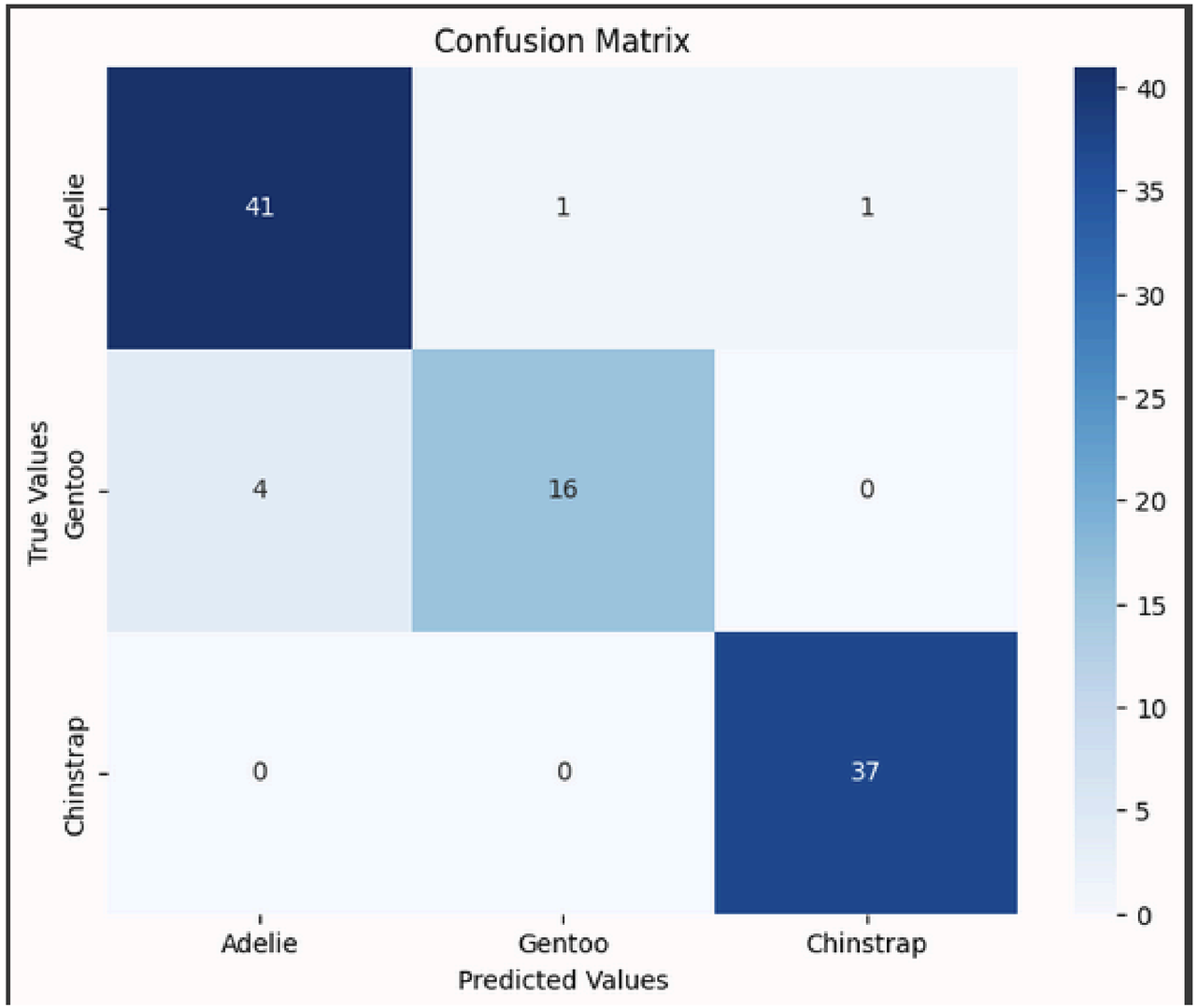
Results

What was the outcome?

Decision Tree



Decision Tree Confusion Matrix



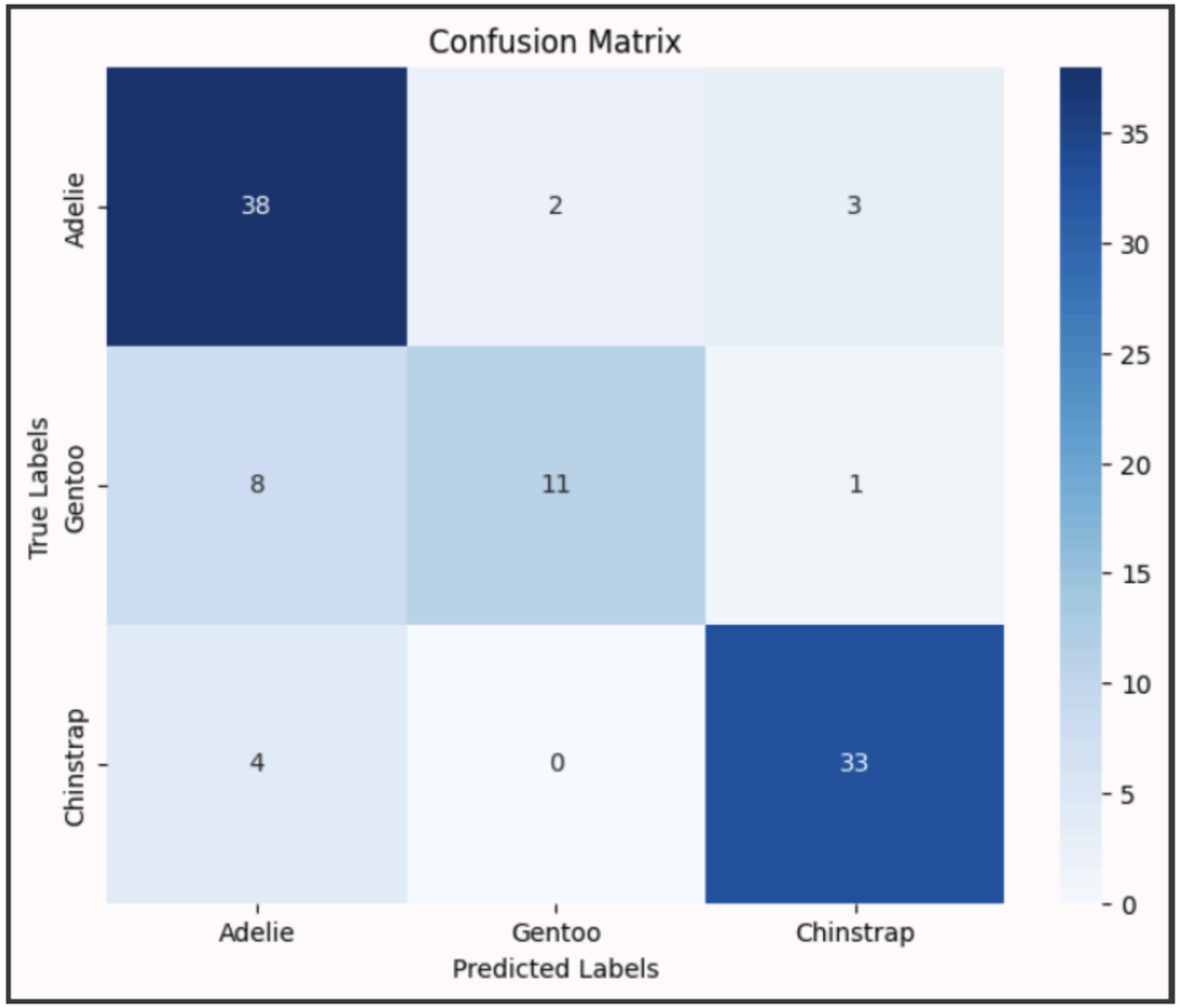

```
Classification Report:

              precision    recall  f1-score

      Adelie       0.91      0.95      0.93
      Gentoo       0.94      0.80      0.86
   Chinstrap       0.97      1.00      0.99

    accuracy                           0.94
   macro avg       0.94      0.92      0.93
weighted avg       0.94      0.94      0.94
```

K-Nearest Neighbour Confusion Matrix



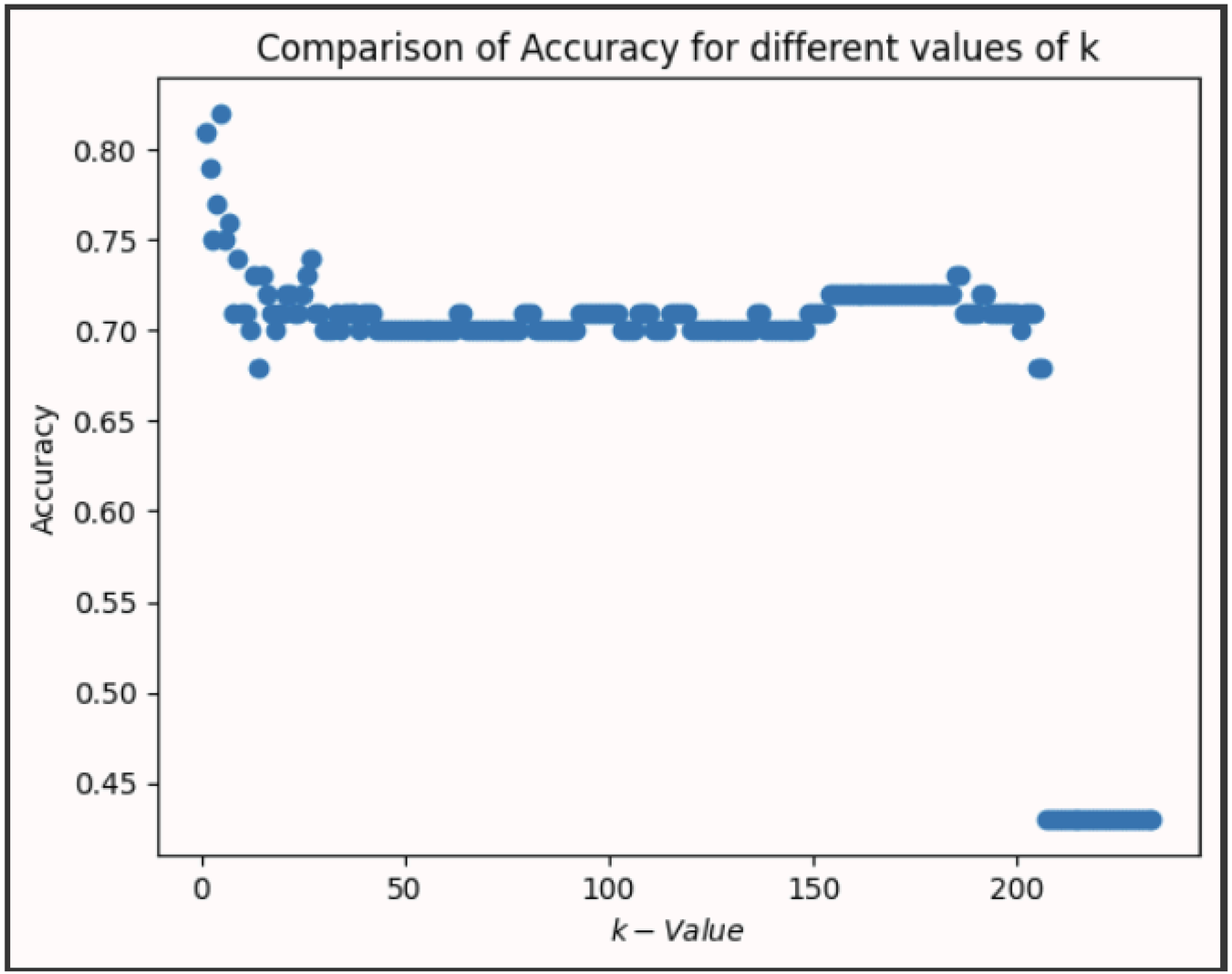

```
Classification Report:

              precision    recall  f1-score

      Adelie       0.76      0.88      0.82
      Gentoo       0.85      0.55      0.67
   Chinstrap       0.89      0.89      0.89

    accuracy                           0.82
   macro avg       0.83      0.78      0.79
weighted avg       0.83      0.82      0.81
```

K-Nearest Neighbour Changing K-Value



When k - value increases from 0 to length of train data, a decline in accuracy can be seen from k(5) to k(25)

Beyond k(25), the accuracy stays relatively constant until k > 200 where the accuracy decreases below 0.45

Based on results, the accuracy is highest when k = 5

Comparison

Decision Tree Result

Classification Report:

	precision	recall	f1-score
Adelie	0.91	0.95	0.93
Gentoo	0.94	0.80	0.86
Chinstrap	0.97	1.00	0.99
accuracy			0.94
macro avg	0.94	0.92	0.93
weighted avg	0.94	0.94	0.94

K - Nearest Neighbour Result (K = 5)

Classification Report:

	precision	recall	f1-score
Adelie	0.76	0.88	0.82
Gentoo	0.85	0.55	0.67
Chinstrap	0.89	0.89	0.89
accuracy			0.82
macro avg	0.83	0.78	0.79
weighted avg	0.83	0.82	0.81

Conclusion

Key Findings

Decision Tree > KNN in this case for accuracy identifying the penguins

Why?

- Small dataset
- Easier modeling & understanding

Reflection

- The **Decision Tree** showcased strong interpretability and robustness, making it well-suited for the dataset's characteristics. However, its performance might be constrained when applied to larger or noisier datasets without proper tuning.

- **KNN**, though conceptually simple, was able to maintain comparable accuracy due to simplicity namely changing the value of k (number of neighbors)

Improvements

Try more advanced techniques

- Testing with other algorithms, such as Random Forest or SVM could provide new insights.

Challenges

Data Preprocessing

- Ensuring the dataset was well-prepared for both models required significant attention to scaling and handling missing values.