ALBUKHARY INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTING AND INFORMATICS

Real-World Data Mining Project

Course Code	**CCS2313**	Course Name	**Data Mining and Analytics**
Group Members	Thinley Yeshey Choden AIU22102188 Dema Yuden AIU22102234		

Assessment Marks:

No.	Criteria	Weight	Marks
1			
2			
3			
	Total Marks		

Comments:

Date of Submission: 30/01/2025

Lecturer’s Name: Sir Farhan Mohsin

Table of contents

1.0 Problem Identification & Dataset Selection

1.1 Dataset Selection

For this project, we have chosen the Palmer Penguins Dataset Extended from kaggle. It provides ecological data on three penguin species in the Palmer Archipelago, Antarctica.

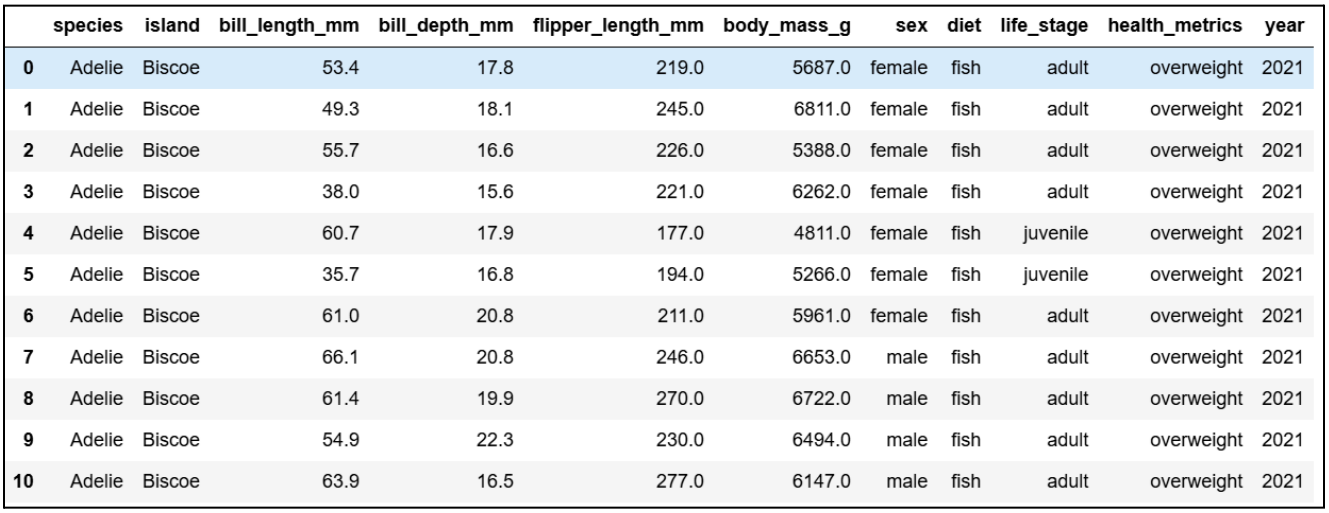
Link to dataset:

https://www.kaggle.com/datasets/samybaladram/palmers-penguin-dataset-extended

1.2 Dataset Description

Palmer Penguins Dataset was collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. The Palmer penguins dataset contains size measurements for three penguin species observed on three islands in the Palmer Archipelago, Antarctica. The **Palmer Penguins Extended Dataset** builds upon the original Palmer Penguins dataset by incorporating additional features that provide a richer and more realistic view of penguin biology and ecology. This extended version offers enhanced opportunities for both exploratory analysis and advanced machine learning tasks.

Table 1. Overview of the Palmer Penguins Extended dataset



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	diet	life_stage	health_metrics	year
0	Adelie	Biscoe	53.4	17.8	219.0	5687.0	female	fish	adult	overweight	2021
1	Adelie	Biscoe	49.3	18.1	245.0	6811.0	female	fish	adult	overweight	2021
2	Adelie	Biscoe	55.7	16.6	226.0	5388.0	female	fish	adult	overweight	2021
3	Adelie	Biscoe	38.0	15.6	221.0	6262.0	female	fish	adult	overweight	2021
4	Adelie	Biscoe	60.7	17.9	177.0	4811.0	female	fish	juvenile	overweight	2021
5	Adelie	Biscoe	35.7	16.8	194.0	5266.0	female	fish	juvenile	overweight	2021
6	Adelie	Biscoe	61.0	20.8	211.0	5961.0	female	fish	adult	overweight	2021
7	Adelie	Biscoe	66.1	20.8	246.0	6653.0	male	fish	adult	overweight	2021
8	Adelie	Biscoe	61.4	19.9	270.0	6722.0	male	fish	adult	overweight	2021
9	Adelie	Biscoe	54.9	22.3	230.0	6494.0	male	fish	adult	overweight	2021
10	Adelie	Biscoe	63.9	16.5	277.0	6147.0	male	fish	adult	overweight	2021

1. **Number of Records:** 3430
2. **Data Types:**

- **Structured:** Numeric, categorical and ordinal features
- **Features:**

Dependent variable:

- **Species (Categorical):** Penguin species - Adelie, Chinstrap, or Gentoo

Independent variables:

- **Island (Categorical):** Island where the penguin was found - Biscoe, Dream, or Torgersen
- **Bill Length (Numeric):** Beak length (in mm)
- **Bill Depth (Numeric):** Beak depth (in mm)
- **Flipper Length (Numeric):** Flipper length (in mm)
- **Body Mass (Numeric):** Body weight (in grams)
- **Sex (Categorical):** Penguin gender - Male, Female
- **Diet (Categorical):** Primary diet of the penguin - Fish, Krill, Parental, Squid
- **Life Stage (Categorical):** Life stage at time of observation - Adult, Juvenile, Chick
- **Health metrics (Categorical):** Health status - Overweight, Healthy, Underweight
- **Year (Ordinal):** Year the data was collected - 2021, 2022, 2023, 2024, 2025

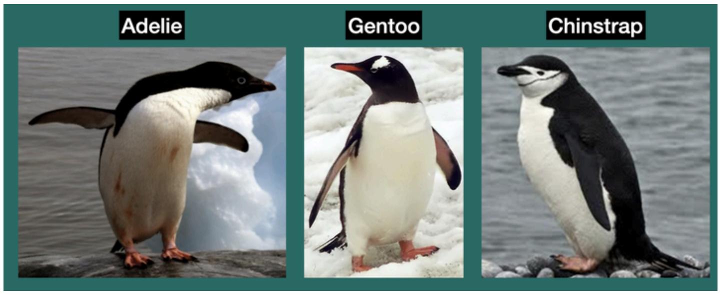


Figure 1. Three different Penguin species.

1.3 Problem Description

The problem involves a classification problem with the objective to develop a predictive model for penguin species classification and health status prediction using physical, diet and environment descriptors including bill length, bill depth, body mass, sex, life stage, and diet.

1.4 Objectives

1. **Penguin Species Classification:** Develop a predictive model to accurately classify penguin species using physical, dietary, and environmental descriptors such as bill length, bill depth, body mass, and sex.

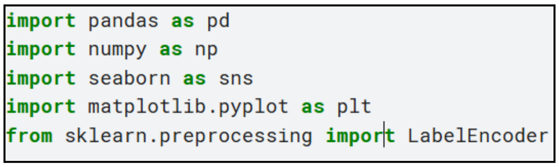
2. **Health Metrics Prediction:** Predict the health status of penguins based on features including physical characteristics, diet, life stage, and environmental factors.

1.5 Proposed Data Mining Model

- **Selected Model:** Random Forest Classifier.
- **Model Goal:** The model seeks to discover penguin species differentiation patterns in terms of physical penguin traits and their health status through various diet habits. Information gained through such a model could inform both ecologic studies and conservation programs through an enhanced grasp of species distributions.
- **Justification:**
    - Handles both categorical and numerical data effectively.
    - Easy to determine feature importance, aiding in ecological insights.
    - Robust against overfitting due to its ensemble approach.
    - Performs well on datasets with non-linear relationships.

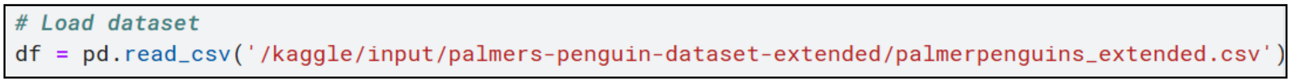
2.0 Data Preprocessing

We begin by importing the necessary libraries and loading the dataset.



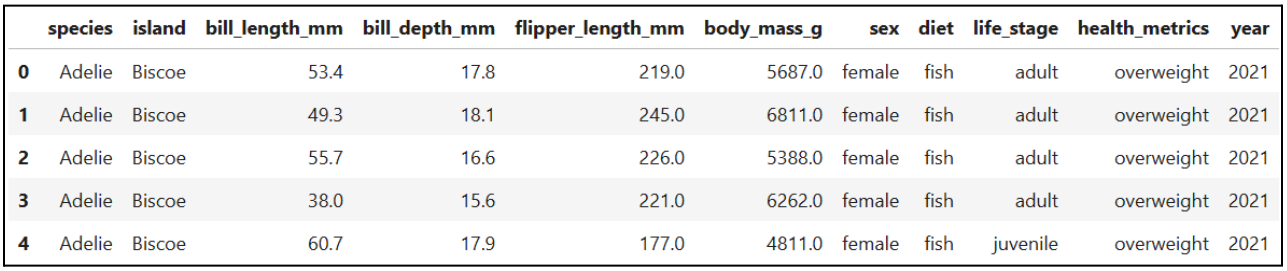
```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
```

Figure 2. Code for importing necessary libraries.



```python
# Load dataset
df = pd.read_csv('/kaggle/input/palmers-penguin-dataset-extended/palmerpenguins_extended.csv')
```

Figure 3. Code for loading the dataset on kaggle.



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	diet	life_stage	health_metrics	year
0	Adelie	Biscoe	53.4	17.8	219.0	5687.0	female	fish	adult	overweight	2021
1	Adelie	Biscoe	49.3	18.1	245.0	6811.0	female	fish	adult	overweight	2021
2	Adelie	Biscoe	55.7	16.6	226.0	5388.0	female	fish	adult	overweight	2021
3	Adelie	Biscoe	38.0	15.6	221.0	6262.0	female	fish	adult	overweight	2021
4	Adelie	Biscoe	60.7	17.9	177.0	4811.0	female	fish	juvenile	overweight	2021

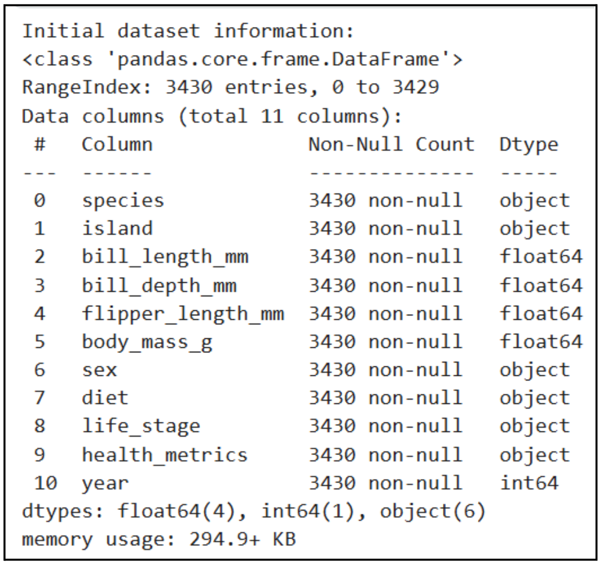
Figure 4. Initial output dataset.

Then we go through 8 major steps:

1. Initial Exploration
2. Data Cleaning
3. Exploratory Data Analysis (EDA)
4. Encoding Categorical Variables
5. Skewness and Outlier Detection
6. Reduce Outlier Impact using Log Transformation
7. Finalize dataset
8. Saving Preprocessed Data

Step 1: Initial Exploration

1. **Dataset Information using df.info()**

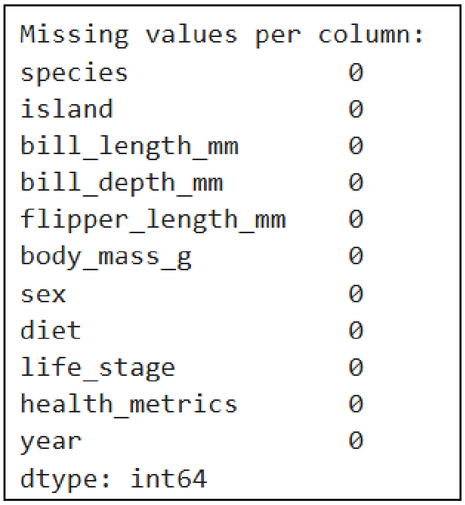


```
Initial dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3430 entries, 0 to 3429
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            3430 non-null   object 
 1   island             3430 non-null   object 
 2   bill_length_mm     3430 non-null   float64
 3   bill_depth_mm      3430 non-null   float64
 4   flipper_length_mm  3430 non-null   float64
 5   body_mass_g        3430 non-null   float64
 6   sex                3430 non-null   object 
 7   diet               3430 non-null   object 
 8   life_stage         3430 non-null   object 
 9   health_metrics     3430 non-null   object 
 10  year               3430 non-null   int64  
dtypes: float64(4), int64(1), object(6)
memory usage: 294.9+ KB
```

Figure 5. Output of dataset information.

The extended penguin dataset has 3430 rows and 11 columns consisting of object datatype for the categorical columns and float data type for the numerical columns.

2. **Missing values were identified per column using df.isnull().sum():**



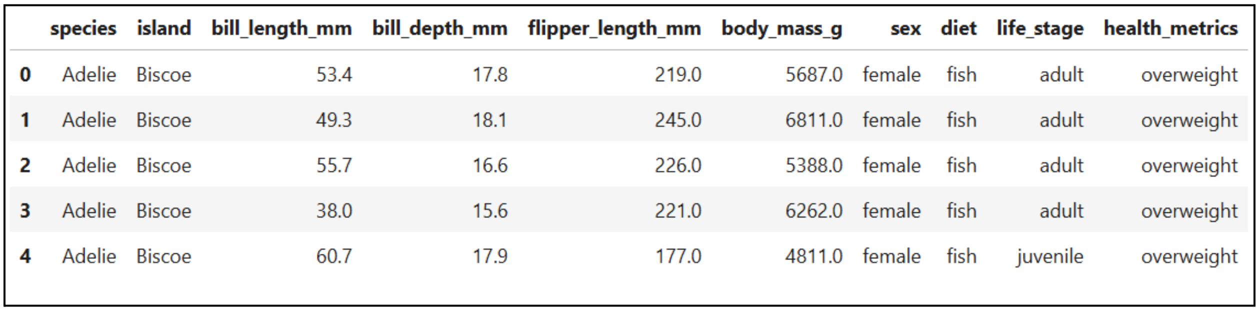
```
Missing values per column:
species              0
island               0
bill_length_mm       0
bill_depth_mm        0
flipper_length_mm    0
body_mass_g          0
sex                  0
diet                 0
life_stage           0
health_metrics       0
year                 0
dtype: int64
```

Figure 6. Output of missing values.

There were no missing values found in the dataset.

Step 2: Data Cleaning

1. Irrelevant Column Removal

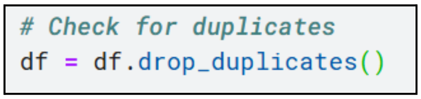


	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	diet	life_stage	health_metrics
0	Adelie	Biscoe	53.4	17.8	219.0	5687.0	female	fish	adult	overweight
1	Adelie	Biscoe	49.3	18.1	245.0	6811.0	female	fish	adult	overweight
2	Adelie	Biscoe	55.7	16.6	226.0	5388.0	female	fish	adult	overweight
3	Adelie	Biscoe	38.0	15.6	221.0	6262.0	female	fish	adult	overweight
4	Adelie	Biscoe	60.7	17.9	177.0	4811.0	female	fish	juvenile	overweight

Figure 7. Output of dataset after removal of year column.

The year column was dropped as it is not relevant for species classification in our context. Unless we are using time series data for future prediction, adding the year column will just add noise in the model training.

2. Duplicate Removal



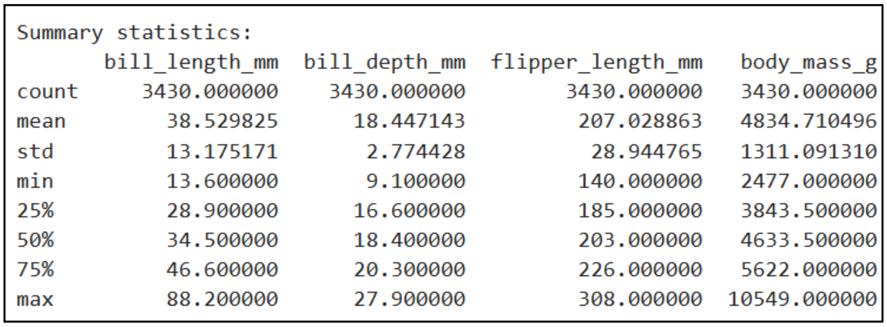
```
# Check for duplicates
df = df.drop_duplicates()
```

Figure 8. Code to drop duplicate rows.

Duplicates were checked and removed to ensure dataset integrity.

Step 3: Exploratory Data Analysis (EDA)

1. Summary Statistics using df.describe()



```
Summary statistics:
       bill_length_mm  bill_depth_mm  flipper_length_mm   body_mass_g
count     3430.000000    3430.000000        3430.000000   3430.000000
mean        38.529825      18.447143         207.028863   4834.710496
std         13.175171       2.774428          28.944765   1311.091310
min         13.600000       9.100000         140.000000   2477.000000
25%         28.900000      16.600000         185.000000   3843.500000
50%         34.500000      18.400000         203.000000   4633.500000
75%         46.600000      20.300000         226.000000   5622.000000
max         88.200000      27.900000         308.000000  10549.000000
```

Figure 9. Output of summary statistics of the dataset.

This Figure provides insights for bill length, bill depth, flipper length, and body mass. On average, bill length is 38.53 mm, while flipper length and body mass are much larger at 207.03 mm and 4,834.71 g, respectively. The dataset shows variability, with flipper length and body mass being the most variable (standard deviations of 28.94 mm and 1,311.09 g).

The ranges reveal that bill length spans from 13.6 mm to 88.2 mm, flipper length from 140 mm to 308 mm, and body mass from 2,477 g to 10,549 g. Most values are clustered between the 25th and 75th percentiles: 28.9–46.6 mm for bill length, 185–226 mm for flipper length, and 3,843.5–5,622 g for body mass.

2. Pairplot

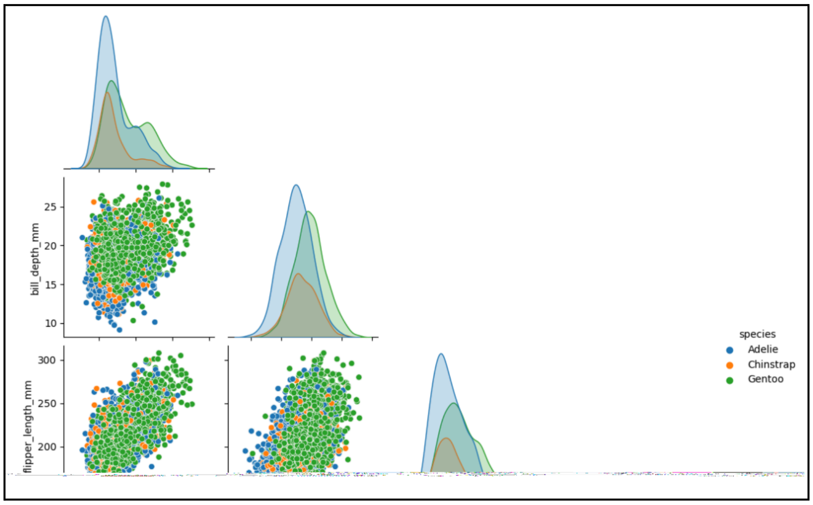


Figure 10. Pairplot.

Key relationships among numerical features were visualized using a pairplot. The scatter plots and KDE distributions reveal clear relationships between body mass, bill length, and flipper length across species. There is a noticeable positive correlation between flipper length and body mass, suggesting that larger penguins tend to have longer flippers. However, overlapping density distributions indicate some similarity in feature ranges between species, which may pose challenges for classification.

3. Violin Plot

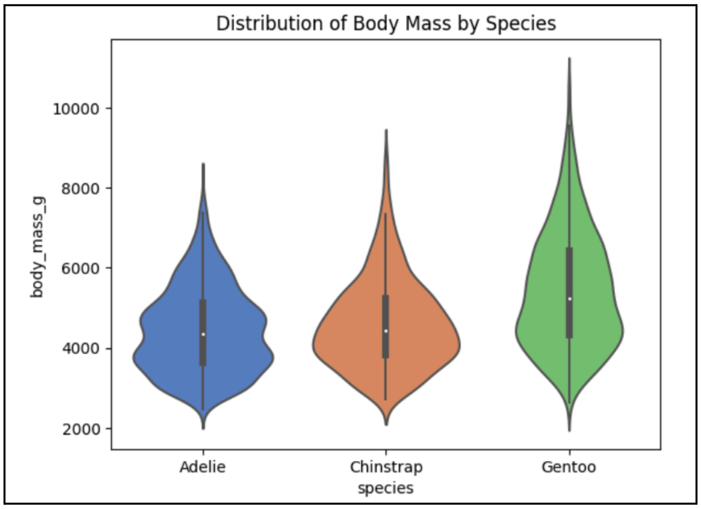


Figure 11. Violin Plot showing distribution of body mass by species.

The body mass distribution varies significantly between species, with Gentoo penguins generally being the heaviest, followed by Chinstrap and then Adelie. This aligns with biological expectations, as Gentoo penguins are typically larger. The violin plot also highlights that Adelie and Chinstrap penguins have more compact distributions, while Gentoo exhibits a wider spread in body mass. This variability could impact predictive modeling efforts when distinguishing species based on body mass.

4. Box Plot

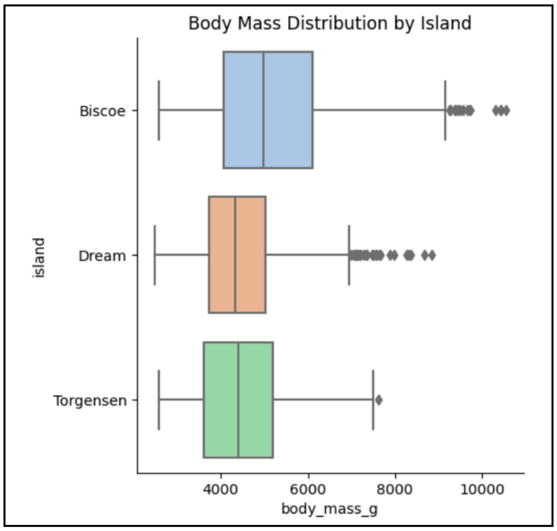
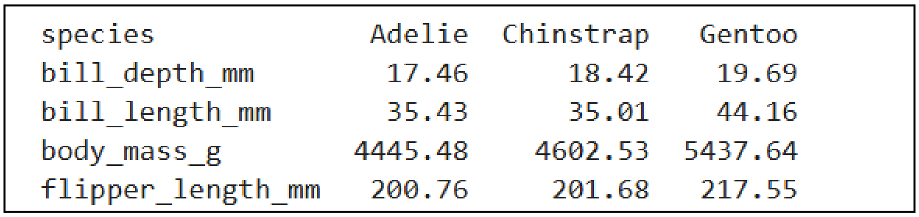


Figure 12. BoxPlot showing body mass distribution across islands.

Penguins on Biscoe Island exhibit the highest median body mass, with a wider range and several high-value outliers (>9000g). Dream Island penguins have a lower median body mass than Biscoe, but also show a number of high-value outliers. Torgersen Island penguins have the lowest median body mass, with a relatively more symmetric distribution and fewer extreme values.

The differences in body mass across islands suggest that island location may serve as an indirect predictor for species classification, given that different species inhabit specific islands.

5. Pivot Table



```
species             Adelie  Chinstrap   Gentoo
bill_depth_mm        17.46      18.42    19.69
bill_length_mm       35.43      35.01    44.16
body_mass_g        4445.48    4602.53  5437.64
flipper_length_mm   200.76     201.68   217.55
```

Figure 13. Pivot table showing mean numerical feature values grouped by species.

The observed differences in bill depth, bill length, body mass, and flipper length suggest that Gentoo penguins are the largest species, followed by Chinstrap, with Adelie being the smallest. The increasing trend in bill size from Adelie to Gentoo may indicate adaptations to different diets, with larger bills potentially aiding in handling a broader range of prey. Similarly, Gentoo penguins also exhibit higher body mass (5437.64 g) and longer flipper length.

6. Correlation Heatmap

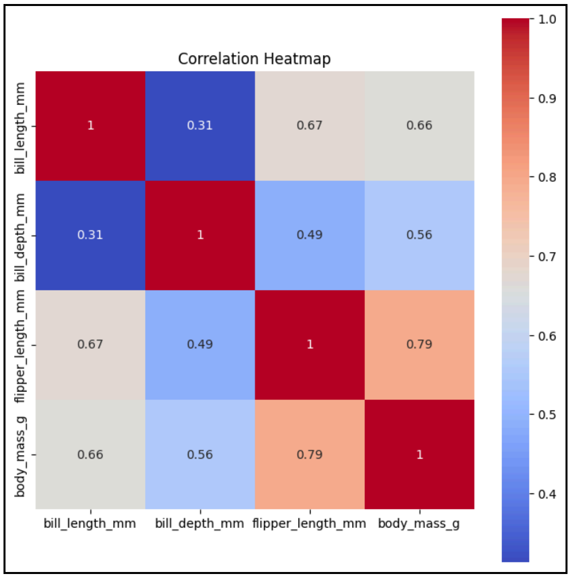


Figure 14. Correlation among numerical features.

There is a strong positive correlation between flipper length and body mass of the penguins, with a correlation coefficient of 0.79. The bill length and flipper length show a moderate correlation of 0.67, followed by a moderate correlation of 0.66 between bill length and body mass.

7. Pairplot with KDE

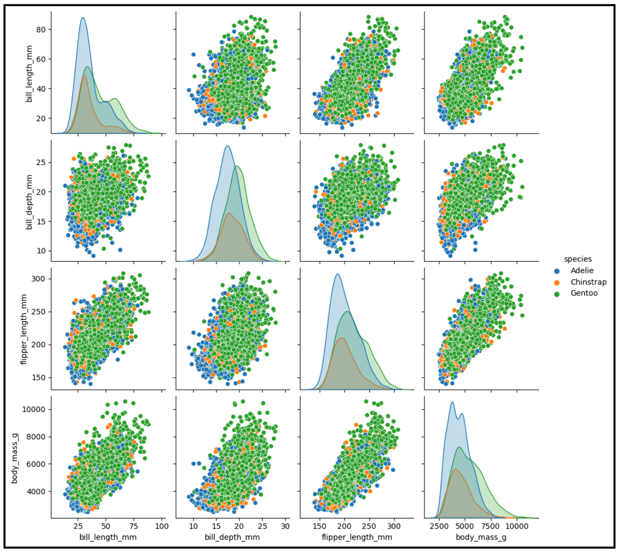


Figure 15. Scatterplot and KDE for species and key features.

The pairplot provides insights into the relationships between numerical features (bill length, bill depth, flipper length, and body mass) across the three penguin species (Adelie, Chinstrap, Gentoo).

- **Feature Distributions (Diagonal KDE Plots):**

Gentoo penguins (green) generally have higher values across most features, especially in body mass and flipper length. Adelie (blue) and Chinstrap (orange) penguins show more overlap, making them harder to distinguish based solely on one feature.

- **Pairwise Feature Relationships (Scatter Plots):**

Body mass and flipper length show a strong positive correlation, suggesting that larger flippers are associated with heavier penguins. Less correlation is observed for bill depth with body mass, indicating that bill depth alone may not be as strong a predictor of species.

The scatter plot clusters indicate that a multivariate classification model (such as Random Forest) will likely perform well since species show distinct feature distributions. Features such as flipper length and body mass appear to be strong predictors, while bill depth may contribute less to classification accuracy.

8. **Countplot for Health Metrics Distribution:**

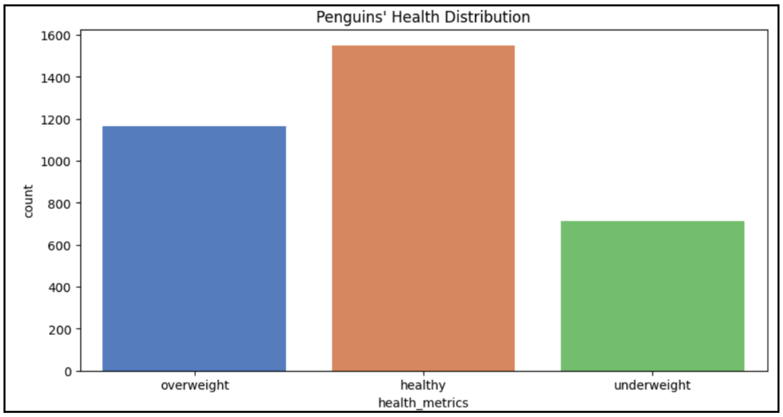


Figure 16. Countplot showing the distribution of different health categories in the dataset.

There is a high count of healthy penguins followed by overweight and then underweight penguins in the dataset.

9. Bill Length vs. Health Status:

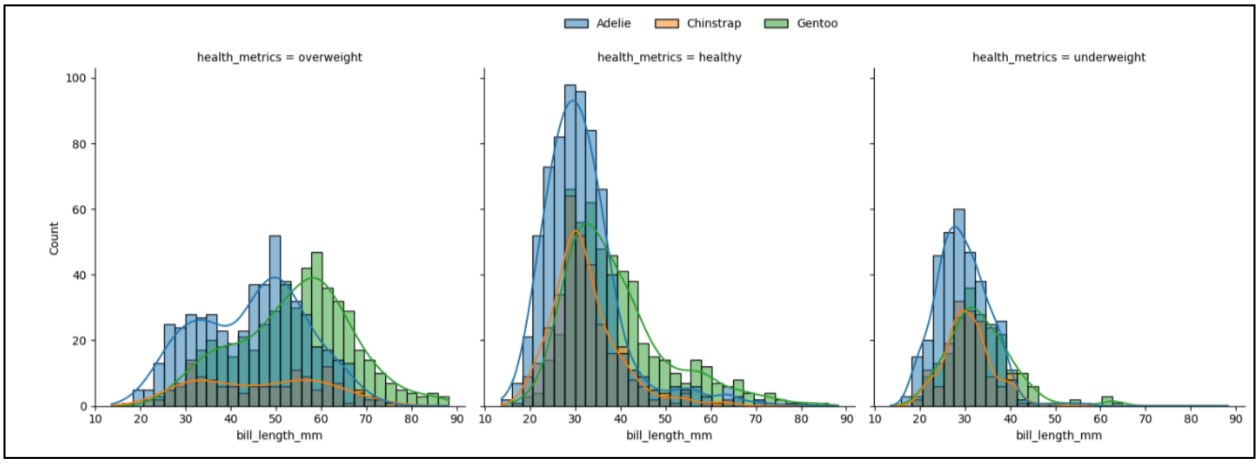


Figure 17. A visualization of how bill length varies among different health conditions across species.

Overweight Penguins tend to have a wider range of bill lengths, with Gentoo showing a strong presence in the higher bill length range. For healthy Penguins, the distributions seem more centered, with Adelie and Chinstrap clustering around the 30-50 mm range. Underweight Penguins follow a similar trend but have fewer counts overall and a slightly lower bill length distribution.

10. Scatterplot for Flipper Length & Body Mass by Health:

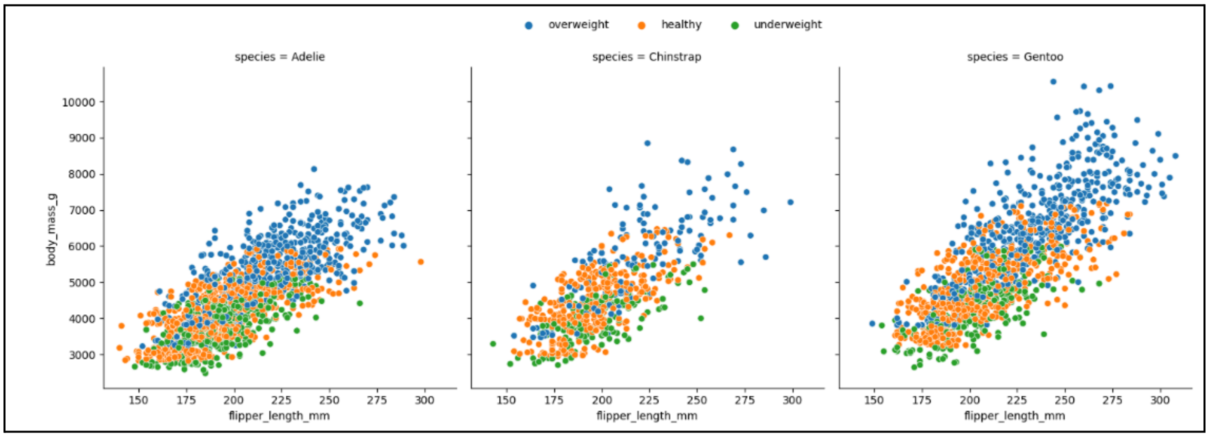
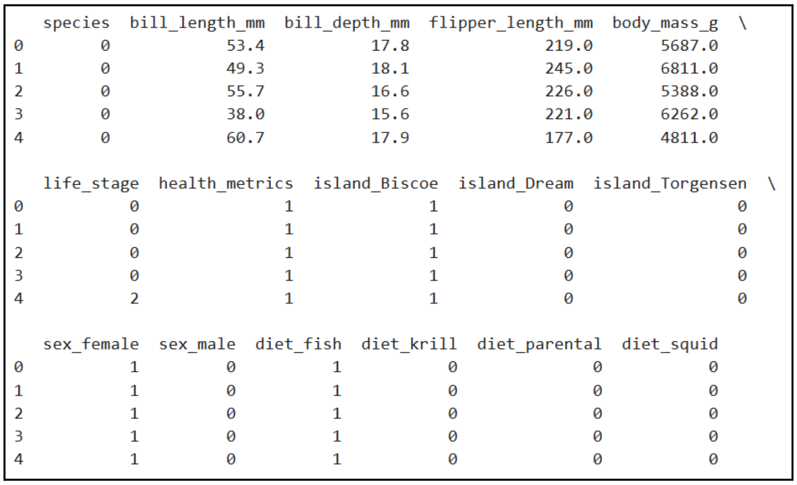


Figure 18. Scatterplot displaying how flipper length and body mass differ across health categories.

There is a clear positive correlation between flipper length and body mass across all species. Gentoo penguins have the highest distribution of larger body mass and flipper lengths, while Adelie penguins tend to be on the lower end. Across species, overweight penguins tend to be clustered in the upper range of body mass, while underweight ones remain in the lower region. There is some variation within each species, but overall, this suggests that larger flippers are generally associated with higher body mass.

Step 4: Encoding Categorical Variables



```
   species  bill_length_mm  bill_depth_mm  flipper_length_mm  body_mass_g  \
0        0            53.4           17.8              219.0       5687.0
1        0            49.3           18.1              245.0       6811.0
2        0            55.7           16.6              226.0       5388.0
3        0            38.0           15.6              221.0       6262.0
4        0            60.7           17.9              177.0       4811.0

   life_stage  health_metrics  island_Biscoe  island_Dream  island_Torgensen  \
0           0               1              1             0                 0
1           0               1              1             0                 0
2           0               1              1             0                 0
3           0               1              1             0                 0
4           2               1              1             0                 0

   sex_female  sex_male  diet_fish  diet_krill  diet_parental  diet_squid
0           1         0          1           0              0           0
1           1         0          1           0              0           0
2           1         0          1           0              0           0
3           1         0          1           0              0           0
4           1         0          1           0              0           0
```

Figure 19. Output dataset after encoding categorical variables.

Categorical variables were encoded using label encoding to transform them into numerical format.

- Ordinal Variables (life_stage, health_metrics) were label encoded, assigning numerical values based on their order. For instance, life_stage values like juvenile, adult, and elder were mapped to 0, 1, 2, respectively, ensuring the model recognizes their progression. Similarly, health_metrics was encoded to reflect increasing health levels (Healthy = 0, Overweight = 1, Underweight = 2).
- Nominal Variables, such as island, sex, and diet, were processed using One-Hot Encoding, creating separate binary columns for each category. This prevents the model

from assuming any artificial ranking between categories. In the output, for example: island_Biscoe is 1 for the first row, indicating the penguin is from Biscoe Island, while island_Dream and island_Torgensen are 0. Similarly, diet_fish is 1, showing the penguin primarily consumes fish, while other diet columns (e.g., diet_krill, diet_squid) are 0.

- Target Variable (species) was encoded using Label Encoding, mapping each species to a numerical label (Adelie = 0, Chinstrap = 1, Gentoo = 2). This allows the model to handle the classification task effectively without introducing relationships between the species.

Step 5: Skewness and Outlier Detection

1. **Histograms**

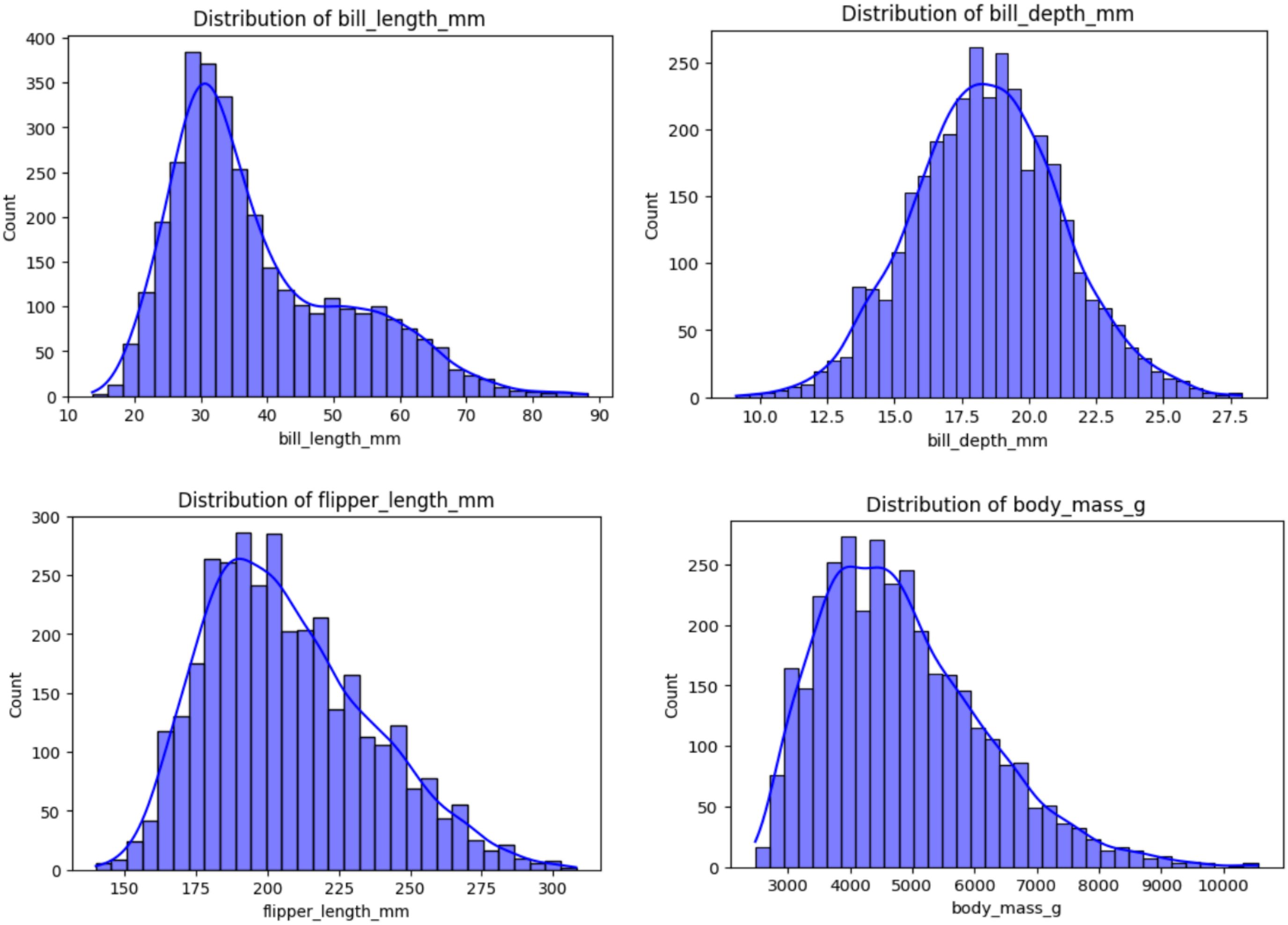


Figure 20. Histograms for numerical features to assess skewness.

- The bill depth distribution is the only one that appears normally distributed, while the others (bill length, flipper length, and body mass) show right-skewness. The presence of long tails in bill length, flipper length, and body mass suggests potential outliers on the higher end of the distribution. If further analysis (e.g., statistical modeling) assumes normality, transformations (like log transformation) might be needed for the skewed variables.

2. **Box Plots**

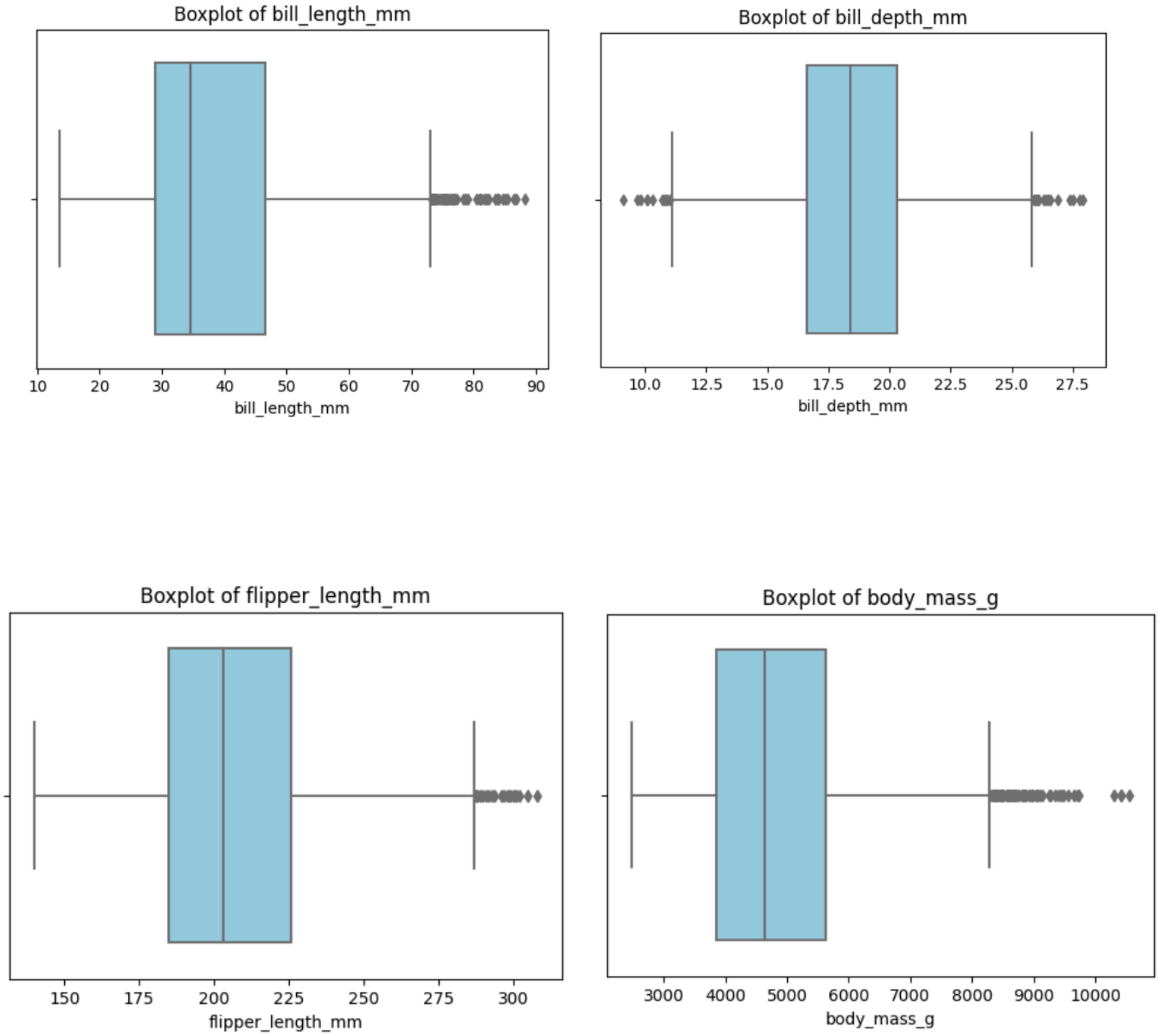


Figure 21. Boxplots to detect potential outliers.

1. Bill Length (bill_length_mm): The boxplot shows several outliers on the higher end. The distribution appears right-skewed, consistent with the histogram. Most data points are

within the interquartile range (IQR), but the presence of many high-value outliers suggests variability in bill length.

2. Bill Depth (bill_depth_mm): There are outliers on both the lower and higher ends. The central box is relatively symmetric, supporting the previous histogram's normal shape. The presence of both lower and upper outliers suggests some penguins have unusually small or large bill depths.
3. Flipper Length (flipper_length_mm): The distribution is right-skewed, with several outliers on the higher end. This confirms that while most penguins have shorter flipper lengths, some individuals have significantly longer ones.
4. Body Mass (body_mass_g): The boxplot reveals numerous high-end outliers. The median is closer to the lower quartile, reinforcing a right-skewed distribution. A large number of heavier penguins exist, which could indicate variations between species or measurement errors.

The right-skewed nature of bill length, flipper length, and body mass is evident from both histograms and boxplots. Significant outliers exist in all variables, particularly in bill length, flipper length, and body mass.

Step 6: Reduce Outlier Impact using Log Transformation

1. **Histogram after log transformation.**

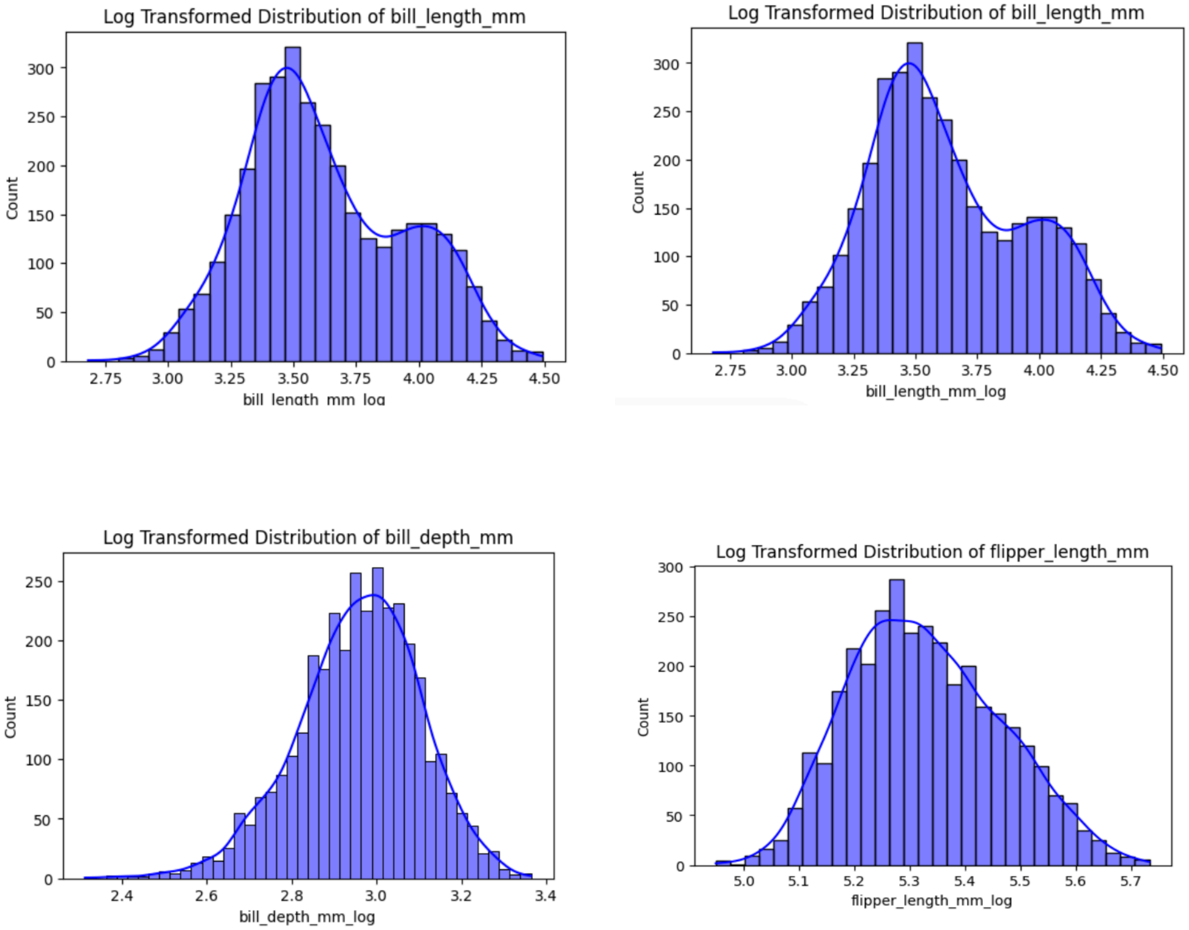


Figure 22. Histograms for numerical features after log transformation.

Log transformation effectively reduced skewness and normalized the data, making it more suitable for statistical analysis. Outliers have less impact on the transformed data, as extreme values are compressed. Distributions are now closer to normal, which is beneficial for models that assume normality (e.g., linear regression, t-tests).

2. Box plots after log transformation.

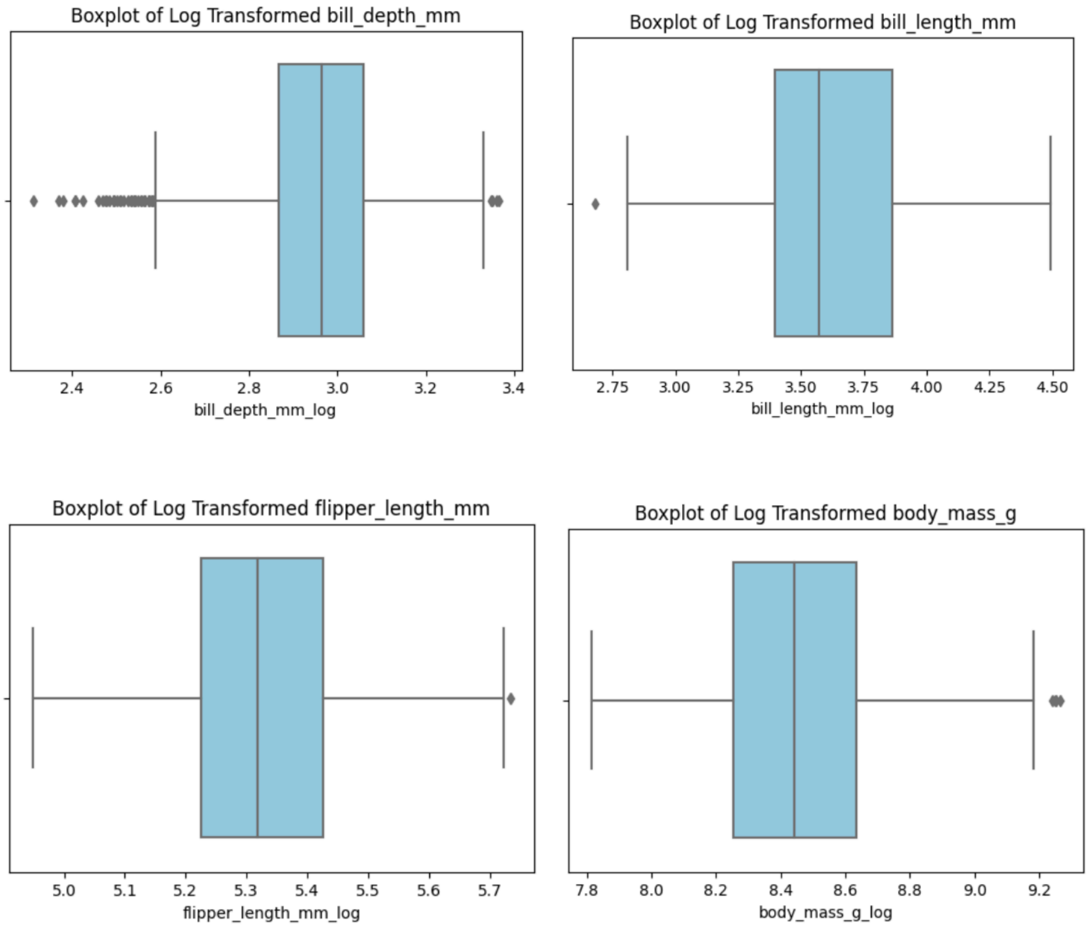
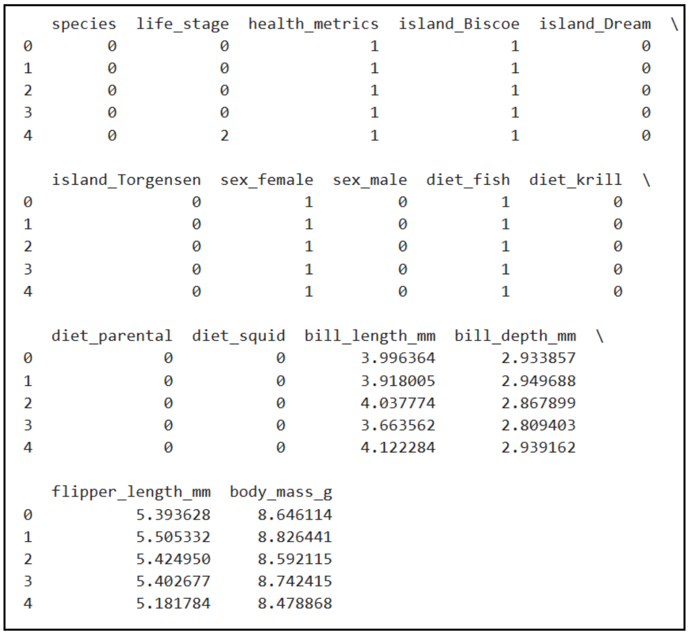


Figure 23. Box plots after log transformation.

Although outlier still in the bill_depth box plot, it has significantly reduced. While the other features have little to no outliers at all. This proves the effectiveness of log transformation in reducing the effect of outliers.

Step 7: Finalize dataset

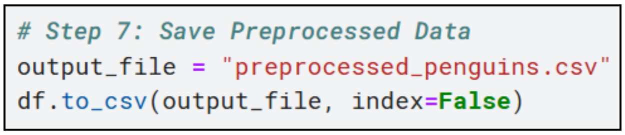


```
   species  life_stage  health_metrics  island_Biscoe  island_Dream  \
0        0           0               1              1             0
1        0           0               1              1             0
2        0           0               1              1             0
3        0           0               1              1             0
4        0           2               1              1             0

   island_Torgensen  sex_female  sex_male  diet_fish  diet_krill  \
0                 0           1         0          1           0
1                 0           1         0          1           0
2                 0           1         0          1           0
3                 0           1         0          1           0
4                 0           1         0          1           0

   diet_parental  diet_squid  bill_length_mm  bill_depth_mm  \
0              0           0        3.996364       2.933857
1              0           0        3.918005       2.949688
2              0           0        4.037774       2.867899
3              0           0        3.663562       2.809403
4              0           0        4.122284       2.939162

   flipper_length_mm  body_mass_g
0           5.393628     8.646114
1           5.505332     8.826441
2           5.424950     8.592115
3           5.402677     8.742415
4           5.181784     8.478868
```

Figure 24. Output dataset after replacing the log transformed values.

The original numerical columns were replaced with their log-transformed versions. First, the untransformed numerical features were dropped, and their corresponding log-transformed columns were renamed to match the original names.

Step 8: Saving Preprocessed Data



```python
# Step 7: Save Preprocessed Data
output_file = "preprocessed_penguins.csv"
df.to_csv(output_file, index=False)
```

Figure 25. Code for saving the preprocessed model.

The cleaned and transformed dataset was saved for model building.

Unlike models that rely on distance-based calculations (e.g., KNN, SVM, Linear Regression), Random Forest is a tree-based algorithm that makes decisions based on threshold splits rather than numerical magnitudes. That is why we did not do any feature scaling here.

3.0 Model Building

This section focuses on developing and testing a machine learning model for two primary tasks:

1. **Species Classification:** Classifying penguin species in terms of a range of physical traits.
2. **Health Status Prediction:** Predicting the health status of penguins using dietary and environmental factors.

We have performed both tasks with a **Random Forest Classifier** and have measured its performance with a range of evaluation metrics.

3.1 Model Selection

We selected Random Forest for its high robustness, its compatibility with both numerical and categorical values, and its effectiveness in overcoming overfitting. Besides, Random Forest classifiers have high performance in working with complex datasets with many features, making it a strong candidate for our classification problem.

3.2 Dataset Preparation

Features Selected:

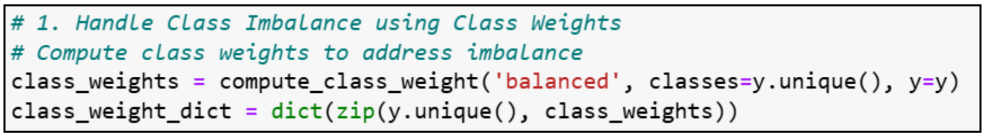
- For species classification: Physical attributes such as ‘bill_length_mm’, ‘bill_depth_mm’, ‘flipper_length_mm’, and ‘body_mass_g’, along with categorical features like ‘island_Biscoe’, ‘island_Dream’, ‘island_Torgensen’, ‘sex_female’, ‘'sex_male', ‘diet_fish’, ‘diet_krill’, ‘diet_parental’, and ‘diet_squid’.
- For health status prediction: Physical attributes, environmental and dietary attributes such as ‘bill_length_mm’, ‘bill_depth_mm’, ‘flipper_length_mm’, ‘body_mass_g’, ‘island_Biscoe’, ‘island_Dream’, ‘island_Torgensen’, ‘sex_female’, ‘sex_male’, ‘diet_fish’, ‘diet_krill’, ‘diet_parental’, and ‘diet_squid’]

Target Variables:

- Species Classification: The species column.
- Health Status Prediction: The health_metrics column.

Data Preprocessing:

- **Class Imbalance Handling:** Applied class weights in Random Forest to address imbalances in both tasks.



```
# 1. Handle Class Imbalance using Class Weights
# Compute class weights to address imbalance
class_weights = compute_class_weight('balanced', classes=y.unique(), y=y)
class_weight_dict = dict(zip(y.unique(), class_weights))
```

Figure 26. Code for handling class imbalance.

3.3 Methodology

Train-Test Splits:

- The data was split into nine different test sizes (from 10% to 90%) to evaluate the robustness and generalizability of the model for both tasks.

Model Training:

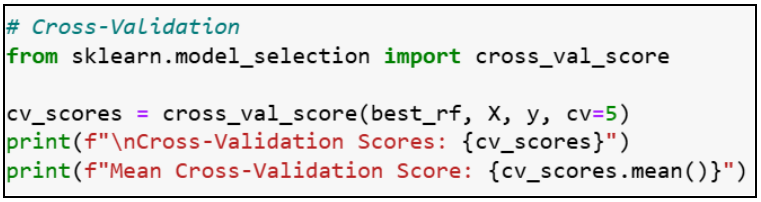
- Trained the Random Forest Classifier on the training data for both tasks, utilizing class weights to handle class imbalances.

Evaluation Metrics:

- For species classification and health status prediction, we used the following evaluation metrics:
    - **Accuracy:** Measure of overall correctness.
    - **Precision:** Proportion of correct positive predictions.
    - **Recall:** Proportion of actual positives correctly predicted.
    - **F1 Score:** Harmonic mean of precision and recall.

Cross-Validation:

- Performed 5-fold cross-validation to assess the stability of the models.



```
# Cross-Validation
from sklearn.model_selection import cross_val_score

cv_scores = cross_val_score(best_rf, X, y, cv=5)
print(f"\nCross-Validation Scores: {cv_scores}")
print(f"Mean Cross-Validation Score: {cv_scores.mean()}")
```

Figure 27. Code for performing cross validation.

3.4 Results Presentation

3.4.1 Species Classification

- **Confusion Matrix for Best Random Forest Model**

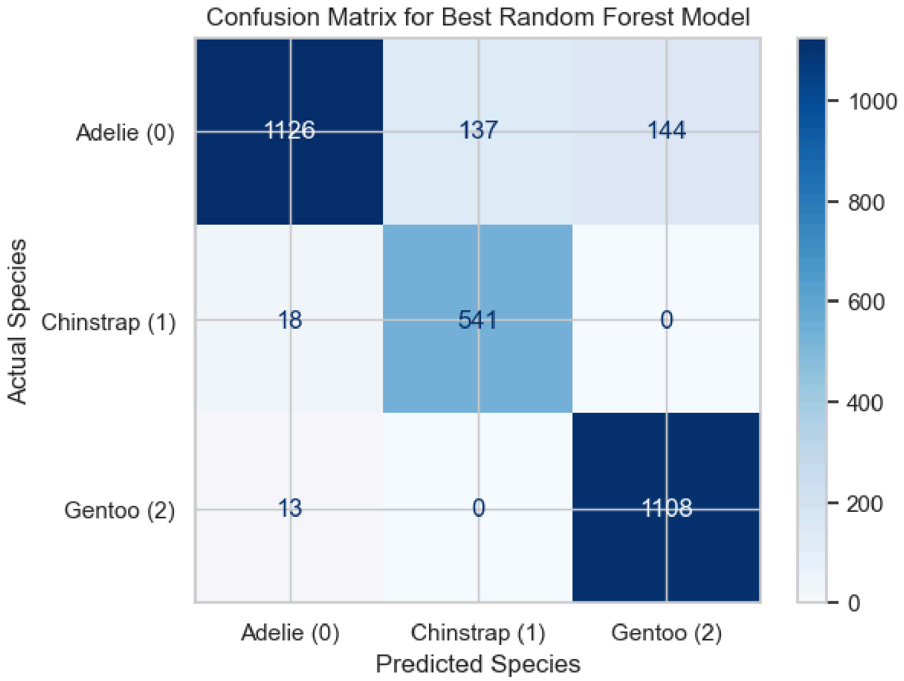


Figure 28. Confusion matrix for species classification.

From the figure above, we can conclude that the model correctly classified 1126 Adelie samples, misclassified 137 as Chinstrap, and 144 as Gentoo. Similarly, for Chinstrap, 541 were correctly classified, while 18 were wrongly predicted as Adelie. For Gentoo, 1108 were correctly classified, while 13 were wrongly predicted as Adelie. The confusion matrix highlights the model's strong performance in predicting Adelie species, with minor misclassifications across other classes.

- **Performance metrics**

The Random Forest model was evaluated on nine different train-test split ratios, ranging from 10:90 to 90:10. The results are summarized in the table below:

Table 2. Performance Metrics for Different Train-Test Split Ratios for Species Classification.

No.	Fold Split (Train:Test)	Accuracy	Precision	Recall	F1-Measure
1.	10:90	0.781341	0.780390	0.781341	0.779299
2.	20:80	0.776968	0.779740	0.776968	0.777902
3.	30:70	0.769679	0.769439	0.769679	0.766631
4.	40:60	0.761662	0.763357	0.761662	0.759512
5.	50:50	0.762099	0.761167	0.762099	0.758914
6.	60:40	0.758989	0.757271	0.758989	0.757176
7.	70:30	0.750521	0.746181	0.750521	0.747704
8.	80:20	0.740889	0.735297	0.740889	0.736264
9.	90:10	0.736961	0.740562	0.736961	0.735823

To visualize these results, a line graph was plotted showing how the performance metrics vary across different train-test splits:

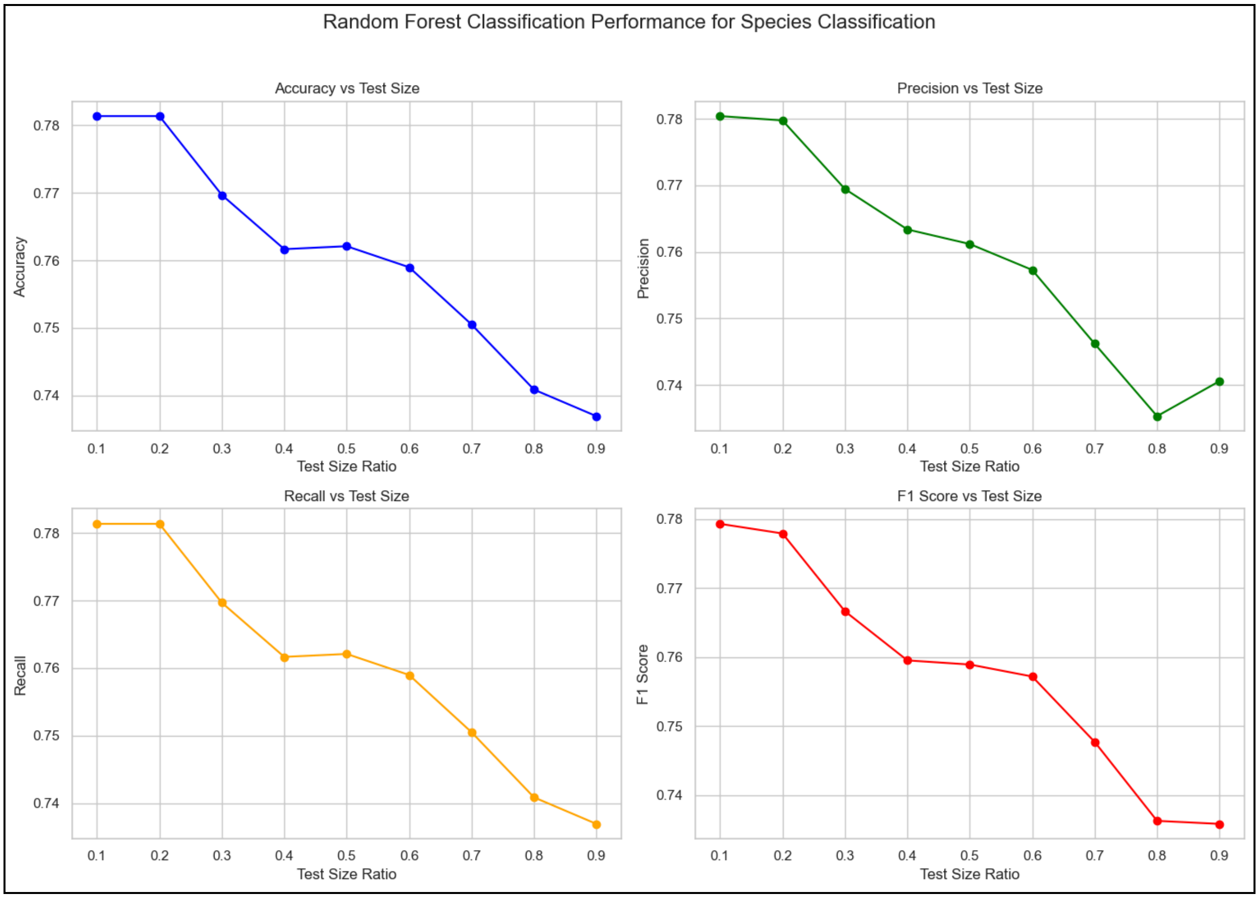


Figure 29. Line graph showing Model Performance vs. Train-Test Split Ratio.

3.4.2 Health Status Prediction

- **Confusion Matrix for Best Random Forest Model**

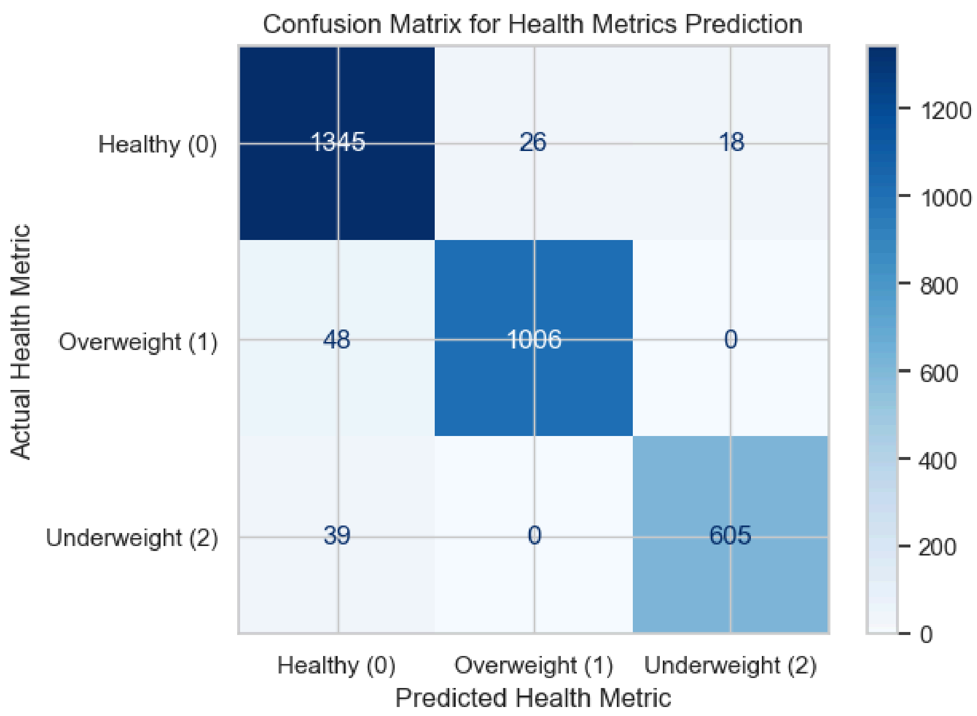


Figure 30. Confusion matrix for health metric prediction.

From the figure above, we can conclude that the model correctly classified 1345 Healthy samples, misclassified 26 as Overweight, and 18 as Underweight. Similarly, for Overweight, 1006 were correctly classified, while 48 were wrongly predicted as Healthy. For Underweight, 605 were correctly classified, while 39 were wrongly predicted as Healthy. The confusion matrix highlights the model's strong performance in predicting Healthy metrics, with minor misclassifications across other classes.

- **Performance metrics**

The Random Forest model was evaluated on nine different train-test split ratios, ranging from 10:90 to 90:10. The results are summarized in the table below:

Table 3. Performance Metrics for Different Train-Test Split Ratios for Health Metric Predictions.

No.	Fold Split (Train:Test)	Accuracy	Precision	Recall	F1-Measure
1.	10:90	0.804665	0.813321	0.804665	0.805174

2.	20:80	0.801749	0.807498	0.801749	0.801788
3.	30:70	0.801749	0.806507	0.801749	0.801679
4.	40:60	0.806122	0.810063	0.806122	0.806044
5.	50:50	0.799417	0.805810	0.799417	0.799521
6.	60:40	0.791545	0.800066	0.791545	0.791114
7.	70:30	0.782591	0.792566	0.782591	0.781716
8.	80:20	0.771137	0.782369	0.771137	0.769322
9.	90:10	0.738905	0.756397	0.738905	0.734411

To visualize these results, a line graph was plotted showing how the performance metrics vary across different train-test splits:

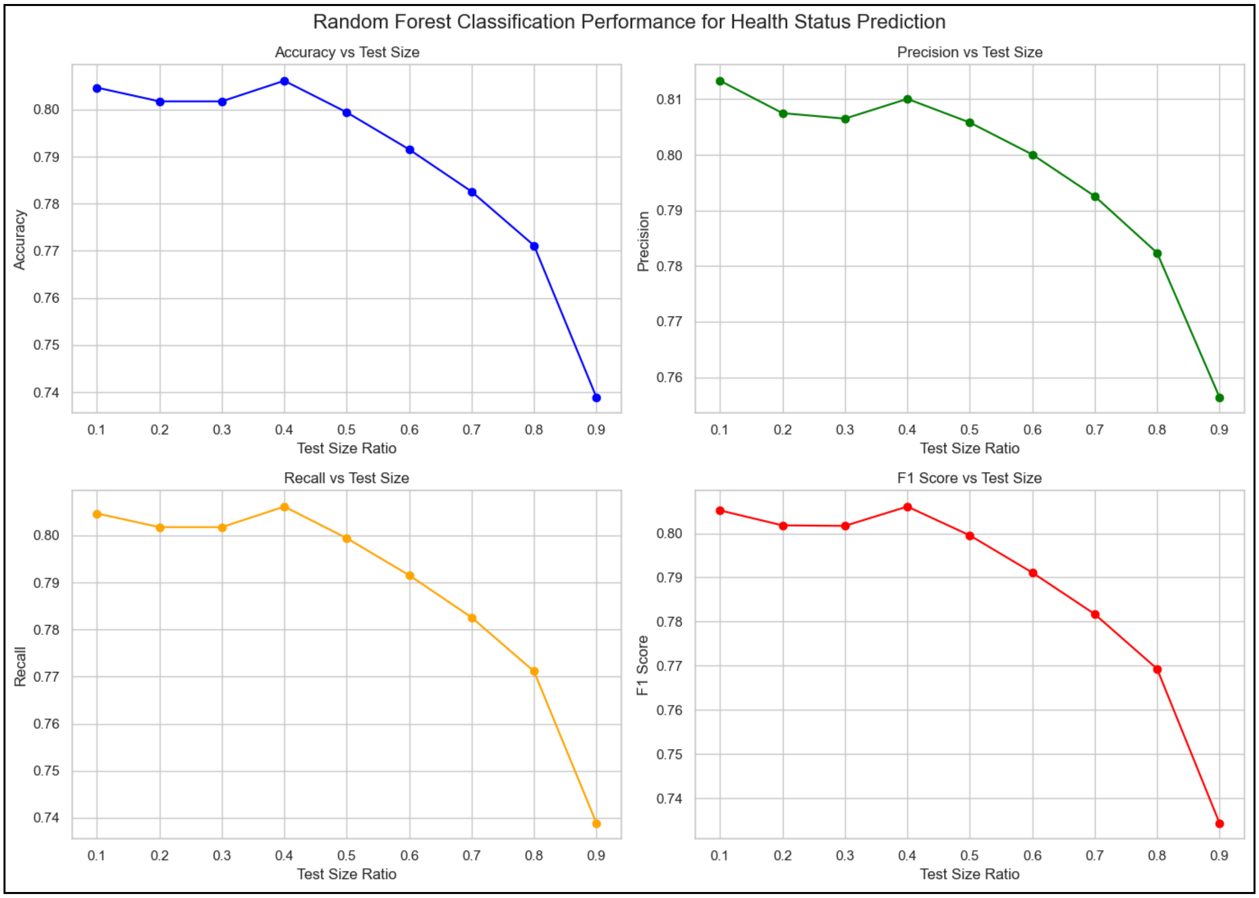


Figure 31. Line graph showing Model Performance vs. Train-Test Split Ratio.

3.5 Insight Analysis

3.5.1 Species Classification

1. **General Performance Trends**

   - **Accuracy, Precision, Recall, and F1 Score:**

     These metrics gradually decline as the test size increases, indicating that the model performs better with a larger training set. This trend is expected because larger training datasets provide more information for the model to learn complex patterns.

   - **Best Performance:**

     The highest accuracy is observed with a **test size of 0.1 (Accuracy: 78.13%)**, indicating the model's strong performance when trained on 90% of the data.

- **Worst Performance:**

  The lowest accuracy occurs at **test size 0.9 (Accuracy: 73.69%)**, which is likely due to insufficient training data.

2. **Optimal Train-Test Split**
   - The test size of 0.1 to 0.3 (10% to 30% test data) appears to provide a good balance between training data availability and reliable test set evaluation. The model performs consistently well in this range.
3. **Model Robustness**
   - The decline in performance is not drastic, which indicates the robustness of the Random Forest model. Despite increasing the test size to 90%, the accuracy still remains above 73%, suggesting that the model generalizes reasonably well.

3.5.2 Health Status Prediction

1. **General Performance Trends**
   - **Accuracy, Precision, Recall, and F1 Score:**

     The best performance is observed with **test size 0.4 (Accuracy: 80.61%)**, while performance steadily declines as the test size increases. The worst accuracy occurs at **test size 0.9 (Accuracy: 73.89%)**, indicating the importance of sufficient training data for robust predictions.
2. **Optimal Train-Test Split**
   - Performance metrics remain relatively stable for test sizes between 0.1 and 0.5, with a peak at test size 0.4 with and accuracy of 80.6%.
3. **Model Robustness**
   - The model demonstrates strong generalization for test sizes up to 0.5, with consistent accuracy, precision, and recall scores.
   - Beyond test size 0.6, performance degrades, indicating that the model struggles to generalize well with limited training data.

3.6 Feature Importance

3.6.1 Species classification

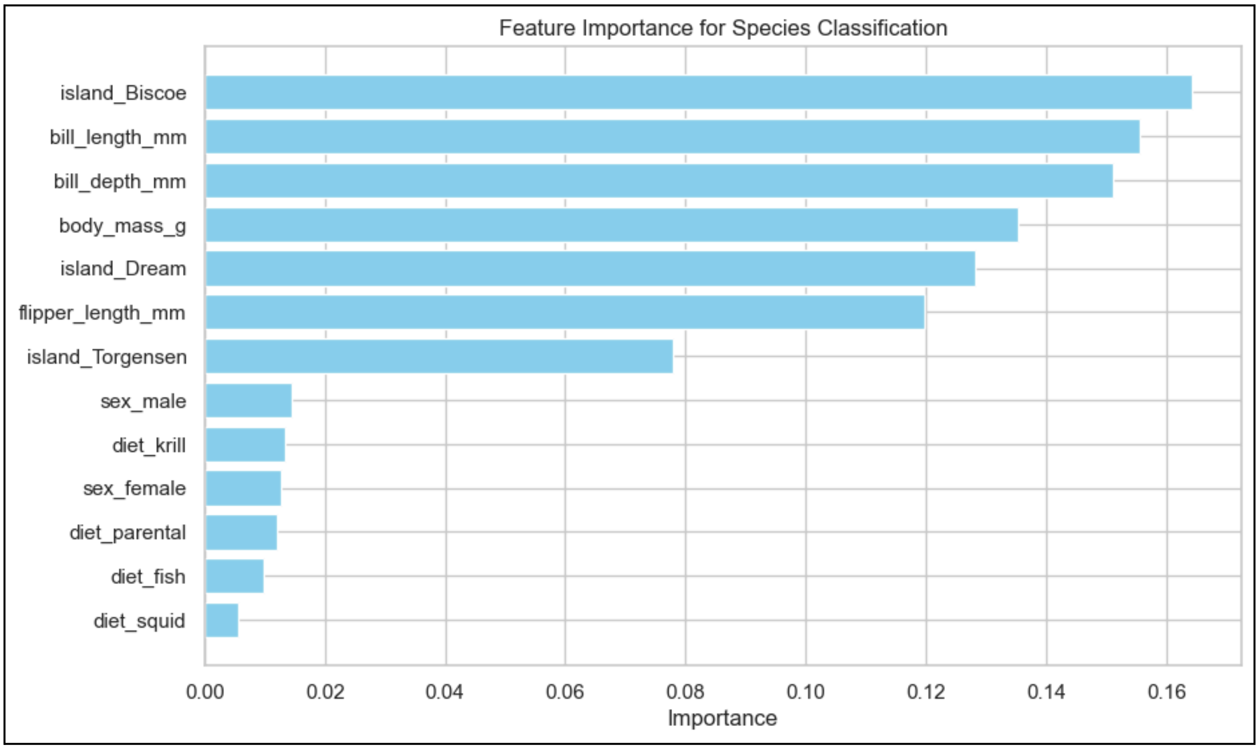


Figure 32. Ranking for feature importance for species classification.

The diagram shows that **island_Biscoe** emerges as the most important feature in species classification model because it directly correlates with the species distribution on the island. It's a strong predictor because geographical location often plays a crucial role in identifying species, especially if the species are isolated to particular regions.

Other features like body mass, bill length, and flipper length may still be important, but they don't have the same level of distinction provided by the island feature.

3.6.2 Health metrics prediction

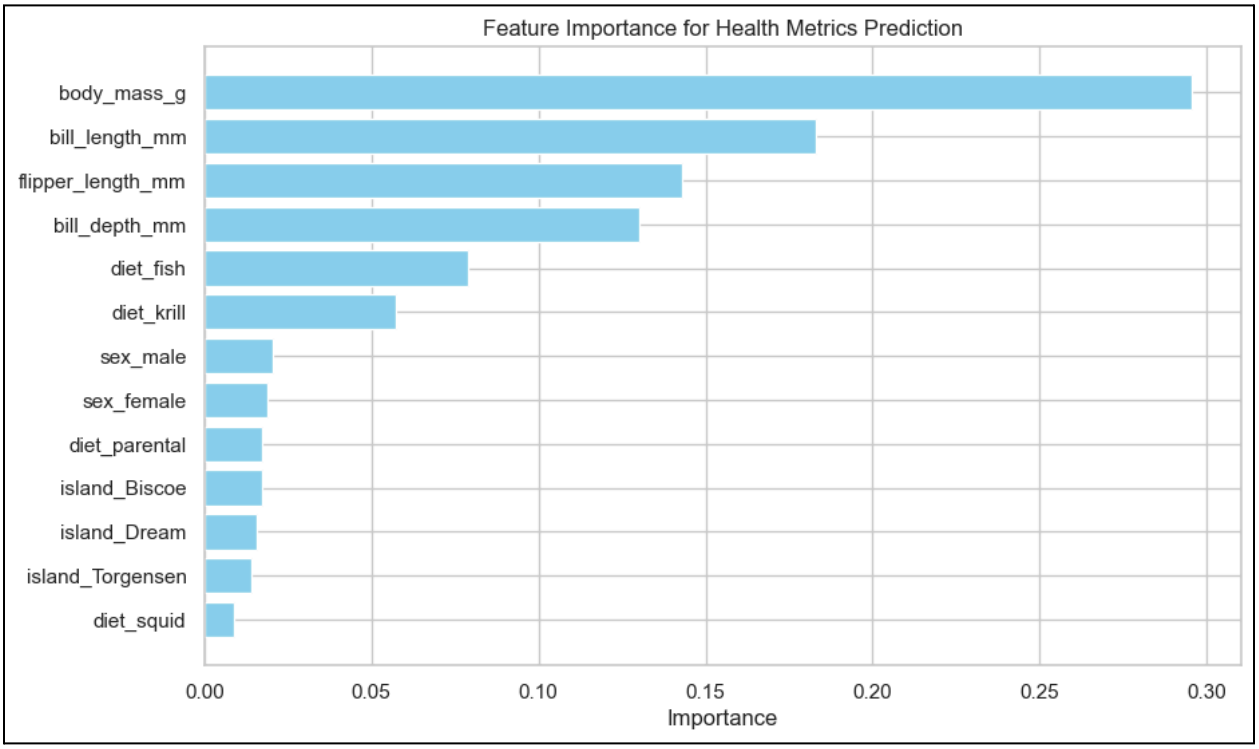


Figure 33. Ranking for feature importance for health metrics prediction.

The diagram essentially shows how **body_mass_g** has the most direct influence on the model's ability to predict health metrics. It is probably because body mass is a crucial biological measure that dictates the health condition in penguins.

4.0 Conclusion and Recommendations

4.1 Summary of Findings

This study successfully applied machine learning to classify penguin species and predict health status based on physical, environmental, and dietary attributes. Through rigorous data preprocessing, we ensured a clean and structured dataset, highlighting the importance of physical traits—such as body mass, bill size, and flipper length—as key differentiators in species classification. Additionally, island location played a crucial role, indicating species distribution patterns across different regions.

For species classification, the Random Forest model demonstrated strong performance, particularly when trained on larger datasets (test size 10-30%), where accuracy remained high. Similarly, for health status prediction, the model achieved peak accuracy at test size 40%, suggesting a balance between training data availability and model generalization. Despite minor misclassifications, the model showed robustness across different train-test splits.

4.2 Key Insights and Recommendations

- Feature Importance: The analysis confirmed that body mass is a dominant predictor for health status, while island location, flipper length, and bill size significantly contribute to species classification. Future studies should maintain these core features while exploring additional environmental variables (e.g., temperature, habitat conditions) to enhance prediction accuracy.
- Model Optimization: While the Random Forest classifier performed well, further improvements can be made through hyperparameter tuning (e.g., optimizing tree depth, adjusting the number of estimators) and exploring alternative models such as Gradient Boosting or Neural Networks for comparison.
- Future Exploration: Incorporating time-based or seasonal variations in diet and habitat could refine predictions and offer deeper ecological insights. Understanding the impact of external factors (e.g., climate change) on species distribution and health would be a valuable next step.

This study demonstrates that physical traits and geographic factors are strong predictors of both species classification and health status in penguins. The model generalizes well, with performance remaining stable even as test sizes increase, reinforcing the robustness of Random Forest. However, further enhancements such as fine-tuning parameters, testing additional models, and integrating environmental features could improve predictive accuracy. Future research should expand on these findings by incorporating a wider range of ecological and behavioral data to improve classification and health assessments.