**Supervised ML – Classification 1 (Palmer Penguins Dataset)**

**3.Palmer Penguins**

Colab Link - <https://colab.research.google.com/drive/10xy0ZIxJ5IB8BQdsledRiC3_2bR9IkCd?usp=sharing>

# --- 0. Setup: Import Libraries ---

A black rectangular object with white text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a computer

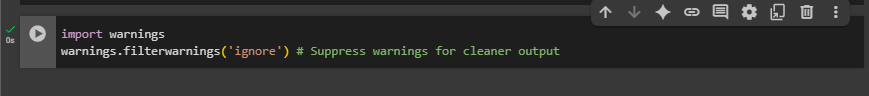
AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

# Note: MAE, MSE, RMSE, R-squared are for regression tasks and won't be applied here,

# But are crucial for evaluating regression models.



1. **Data Loading & Initial Inspection-**

**# Load the Palmer Penguins dataset from seaborn**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**# Display the first few rows**

**A screenshot of a computer

AI-generated content may be incorrect.**

**# Get a concise summary of the DataFrame**

**A screenshot of a computer

AI-generated content may be incorrect.**

**# Check for missing values**

**A screenshot of a computer

AI-generated content may be incorrect.**

**# Basic descriptive statistics for numerical columns**

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Data Cleaning, Preprocessing & Wrangling**

# --- Handling Missing Values ---

# For simplicity in this tutorial, we'll drop rows with any missing values.

# In a real-world scenario, you might impute them (e.g., mean, median, mode, or more advanced methods).

**A screen shot of a computer

AI-generated content may be incorrect.**

**# --- Feature Engineering (Conceptual) ---**

**# For this dataset, we won't create complex new features, but it's important**

**# to understand that this phase is where you might combine features,**

**# extract information (e.g., from dates), or apply mathematical transformations.**

**# Example: If we had 'year' and 'month' columns, we could engineer a 'season' feature.**

**# Or, if these were images, we'd extract features using CNNs.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**# --- Encoding Categorical Features ---**

**# 'species' is our target variable (y). 'island' and 'sex' are features (X).**

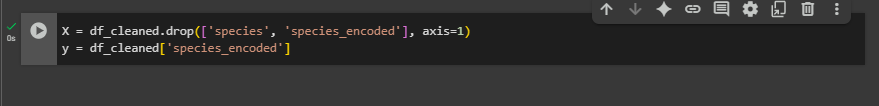
**# We need to convert these into numerical representations.**

**# Encode the target variable 'species'**

**A screen shot of a computer code

AI-generated content may be incorrect.**

**# Define features (X) and target (y)**

****

**# Identify categorical and numerical features for preprocessing**

**A black screen with orange text

AI-generated content may be incorrect.**

**# Create a preprocessing pipeline**

**# One-hot encode categorical features and scale numerical features**

**A computer screen shot of a black screen

AI-generated content may be incorrect.**

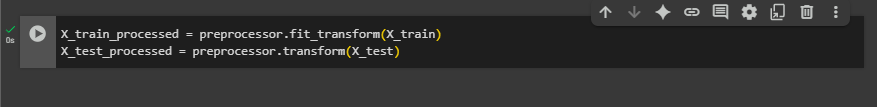
**# --- Train-Test Split ---**

**# Split the dataset into training and testing sets to evaluate model performance**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**# Apply preprocessing to training and testing data**

****

**# Get feature names after one-hot encoding for better interpretability**

**A screenshot of a computer

AI-generated content may be incorrect.**

**3.Exploratory Data Analysis (EDA) –**

**# Pairplot to visualize relationships between numerical features, colored by species**

**A screen shot of a computer code

AI-generated content may be incorrect.**

A screenshot of a graph

AI-generated content may be incorrect.

**# Correlation Heatmap for numerical features**

**A screen shot of a computer

AI-generated content may be incorrect.**

A diagram of a heat map

AI-generated content may be incorrect.

**# Distribution of Species**

**A black rectangular object with a white border

AI-generated content may be incorrect.**

A graph of different species

AI-generated content may be incorrect.

**# Boxplots of numerical features by species**

**A screen shot of a computer code

AI-generated content may be incorrect.**

A group of blue and white boxes

AI-generated content may be incorrect.

**4.Model Training with Different Algorithms**

**# Initialize different classifiers**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A computer screen shot of a computer

AI-generated content may be incorrect.**

**5. Model Evaluation**

**A computer screen shot of a program code

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

A group of blue squares

AI-generated content may be incorrect.

**# --- ROC Curve (Multiclass) ---**

**# For multiclass ROC, a common approach is One-vs-Rest (OvR)**

**# We calculate ROC for each class against all others.**

**A screen shot of a computer

AI-generated content may be incorrect.**

**A computer screen shot of a program

AI-generated content may be incorrect.**

A graph with numbers and lines

AI-generated content may be incorrect.

**A screenshot of a computer program

AI-generated content may be incorrect.**

**6. Prediction & Decision Making**

**# Choose the best performing model. In this case, many models achieved very high accuracy.**

**# Let's pick Random Forest as an example, as it's robust.**

**# --- Example Prediction on unseen data ---**

**# Let's create a hypothetical new penguin measurement**

**# (Make sure these values are within a reasonable range for penguins)**

**A screen shot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**# Preprocess the new data using the same preprocessor fitted on training data**

**# Make a prediction**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**# If we want the probability distribution:**

**A screen shot of a computer

AI-generated content may be incorrect.**

**# --- Decision Making ---**

**A screen shot of a computer

AI-generated content may be incorrect.**

**A black screen with white text

AI-generated content may be incorrect.**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **New Data** | **Prediction** | **Decision** |
| **1** | { 'bill\_length\_mm': 39.5, 'bill\_depth\_mm': 17.5, 'flipper\_length\_mm': 190.0, 'body\_mass\_g': 3500.0, 'island': 'Torgersen', 'sex': 'Male' } | Penguin 1: Predicted Species = Adelie | Adelie |
| **2** | **{ 'bill\_length\_mm': 46.0, 'bill\_depth\_mm': 21.0, 'flipper\_length\_mm': 210.0, 'body\_mass\_g': 4200.0, 'island': 'Biscoe', 'sex': 'Female' }** | Penguin 2: Predicted Species = Gentoo | Gentoo |
| **3** | **{ 'bill\_length\_mm': 50.5, 'bill\_depth\_mm': 15.3, 'flipper\_length\_mm': 222.0, 'body\_mass\_g': 5000.0, 'island': 'Dream', 'sex': 'Male' }** | Penguin 3: Predicted Species = Chinstrap | Chinstrap |
| **4** | **{ 'bill\_length\_mm': 36.2, 'bill\_depth\_mm': 18.9, 'flipper\_length\_mm': 181.0, 'body\_mass\_g': 3200.0, 'island': 'Torgersen', 'sex': 'Female' }** | Penguin 4: Predicted Species = Adelie | Adelie |
| **5** | **{ 'bill\_length\_mm': 42.1, 'bill\_depth\_mm': 19.5, 'flipper\_length\_mm': 200.0, 'body\_mass\_g': 3800.0, 'island': 'Biscoe', 'sex': 'Male' }** | **Penguin 5: Predicted Species = Adelie** | **Adelie** |
| **6** | **{ 'bill\_length\_mm': 45.6, 'bill\_depth\_mm': 20.2, 'flipper\_length\_mm': 210.0, 'body\_mass\_g': 3950.0, 'island': 'Dream', 'sex': 'Female' }** | **Penguin 6: Predicted Species = Chinstrap** | **Chinstrap** |
| **7** | **{ 'bill\_length\_mm': 48.0, 'bill\_depth\_mm': 16.8, 'flipper\_length\_mm': 218.0, 'body\_mass\_g': 4700.0, 'island': 'Biscoe', 'sex': 'Male' }** | **Penguin 7: Predicted Species = Gentoo** | **Gentoo** |
| **8** | **{ 'bill\_length\_mm': 37.8, 'bill\_depth\_mm': 18.5, 'flipper\_length\_mm': 192.0, 'body\_mass\_g': 3450.0, 'island': 'Torgersen', 'sex': 'Female' }** | **Penguin 8: Predicted Species = Adelie** | **Adelie** |
| **9** | **{ 'bill\_length\_mm': 41.3, 'bill\_depth\_mm': 17.7, 'flipper\_length\_mm': 196.0, 'body\_mass\_g': 3600.0, 'island': 'Dream', 'sex': 'Male' }** | **Penguin 9: Predicted Species = Adelie** | **Adelie** |
| **10** | **{ 'bill\_length\_mm': 52.0, 'bill\_depth\_mm': 20.5, 'flipper\_length\_mm': 225.0, 'body\_mass\_g': 5200.0, 'island': 'Biscoe', 'sex': 'Female' }** | **Penguin 10: Predicted Species = Gentoo** | **Gentoo** |

**Conclusion:**

* I learned how to use machine learning to classify data, specifically by analyzing physical characteristics of penguins to predict their species. This project gave me a solid understanding of the entire data science pipeline, from data preprocessing to model evaluation.
* The biggest difficulty I faced was handling the dataset's messy aspects, like missing values and duplicate rows. I had to learn to systematically clean and prepare the data before it could be used for modeling. Another challenge was visualizing high-dimensional data, which I overcame by using dimensionality reduction techniques like PCA.
* Through this project, I gained several key skills, including data cleaning and preprocessing, feature scaling, exploratory data analysis (EDA), and applying unsupervised learning algorithms for classification. I also improved my problem-solving skills and learned to interpret the results of a machine learning model.
* If I were offered a job, I could apply these skills to a wide range of tasks, such as analyzing customer data to identify trends, predicting product demand, or even detecting fraudulent activity by spotting anomalies in a data set. My experience with this project has given me a practical foundation in data analysis that is directly applicable to many real-world business problems.
* This experience has motivated me to continue learning about advanced machine learning topics like deep learning and natural language processing. I’m excited to explore new datasets and build more complex models to solve different kinds of challenges.