# SUPERVISED MACHINE LEARNING WITH PALMER ARCHIPELAGO (ANTARTICA) PENGUIN DATASET

Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine

by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1EuEOJQf1u2GyGMOKkWkAV6AwdAPpM0Qt?usp=sharing


Three species of Palmer Archipelago Penguin (Pfeifer et al., 2025).

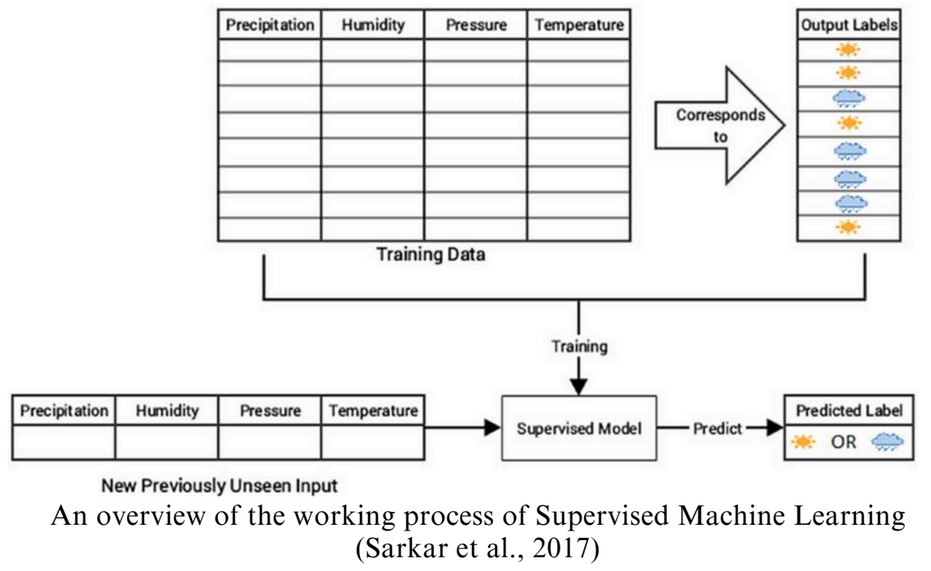
## INTRODUCTION

### PALMER ARCHIPELAGO PENGUIN DATASET

The Palmer Archipelago Penguin Dataset is an alternative to the well-known Iris dataset for statistical and data science education. It contains real-world body size measurements of three *Pygoscelis* penguin species—Adélie, Chinstrap, and Gentoo—collected from 2007 to 2009 in the Western Antarctic Peninsula as part of the US Long-Term Ecological Research (LTER) Network (Horst et al., 2022).

### SUPERVISED MACHINE LEARNING

Supervised Machine Learning is a method where a model learns from labeled data to recognize patterns. It uses this knowledge to predict the correct category or value for new data (Sarkar et al., 2017).




An overview of the working process of Supervised Machine Learning (Sarkar et al., 2017)

### RESEARCH PURPOSE

Determining the best Supervised Machine Learning model for the Iris dataset among Logistic Regression, K-NN, Decision Tree, Random Forest, and SVM.

## LITERATURE REVIEW

### SUPERVISED MACHINE LEARNING MODELS

**1. Logistic Regression**

Logistic Regression models the probability of binary classification, assigning the positive class (1) if it exceeds a threshold (commonly 50%) and the negative class (0) if below (Géron, 2017).

**2. K-NN**

K-NN classifies data based on the k closest points in the dataset. It measures the distance to find the nearest neighbors and assigns the most common class among them. The parameter k represents the number of neighbors considered for classification (Géron, 2017; Marsland, 2014).

**3. Decision Tree**

A Decision Tree uses a series of decision rules to split data at decision nodes, forming a tree-like structure. Each rule creates branches leading to new nodes, with terminal branches called leaves. This model is popular for its interpretability and serves as the foundation for various tree-based extensions (Albon, 2018).

**4. Random Forest**

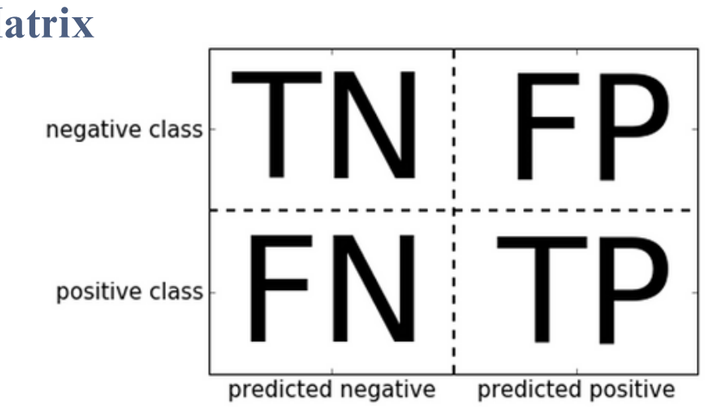
Random Forest is a collection of Decision Trees that work together (Albon, 2018).

**5. Support Vector Machine**

Support Vector Machine (SVM) determines the optimal boundary (hyperplane) that separates classes by creating the widest possible margin. To handle complex, non-linear data, SVM transforms it into a higher-dimensional space using kernel functions. However, it requires high computational power for large datasets (Marsland, 2014).

### MODEL EVALUATION

**Confusion Matrix**




Typical structure of a confusion matrix (Müller & Guido, 2016).

From the confusion matrix, we derive key **Performance metrics**:

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The percentage of predicted positives that are actually correct.
- **Recall (Sensitivity)**: The percentage of actual positives correctly identified.
- **F1 Score**: The harmonic mean of precision and recall, balancing both metrics.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

(Sarkar et al., 2017)

## RESEARCH METHODOLOGY

**DATA SOURCE** →secondary data from the website https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data/data , accessed on February 14, 2025.

### RESEARCH VARIABLES

**Independent Variables (Features)**:
represented by the symbol "X"

numerical features/columns:
- Sepal Length (cm)
- Sepal Width (cm)
- Petal Length (cm)
- Petal Width (cm)

categorical features/columns:
- island
- sex

**Dependent Variable (Target/Label)**:
represented by the symbol "y"
Species (Adelie, Chinstrap Gentoo)

### Analysis Steps

1. Identify the problem and input the dataset.
2. Exploratory Data Analysis
3. Preparation of Train Data and Test Data.
4. Train the models and evaluate their performance.
5. Summarize the results and draw conclusions.

# SUPERVISED MACHINE LEARNING WITH PALMER ARCHIPELAGO (ANTARTICA) PENGUIN DATASET

Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine

by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1EuEOJQf1u2GyGMOKkWkAV6AwdAPpM0Qt?usp=sharing


Three species of Palmer Archipelago Penguin (Pfeifer et al., 2025).

## DISCUSSION

### 1. INPUT THE DATASET

change all column names & categorical features into lowercase

```
df = pd.read_csv('/content/penguins_size.csv')
```

```
df.columns = df.columns.str.lower().str.replace(' ', '_')

categorical_columns = list(df.dtypes[df.dtypes == 'object'].index)

for c in categorical_columns:
  df[c] = df[c].str.lower().str.replace(' ', '_')
```

```
df.head()
```

| | species | island | culmen_length_mm | culmen_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| 0 | adelie | torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | male |
| 1 | adelie | torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | female |
| 2 | adelie | torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | female |
| 3 | adelie | torgersen | NaN | NaN | NaN | NaN | NaN |
| 4 | adelie | torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | female |

### 2. EXPLORATORY DATA ANALYSIS

Handling Missing Value

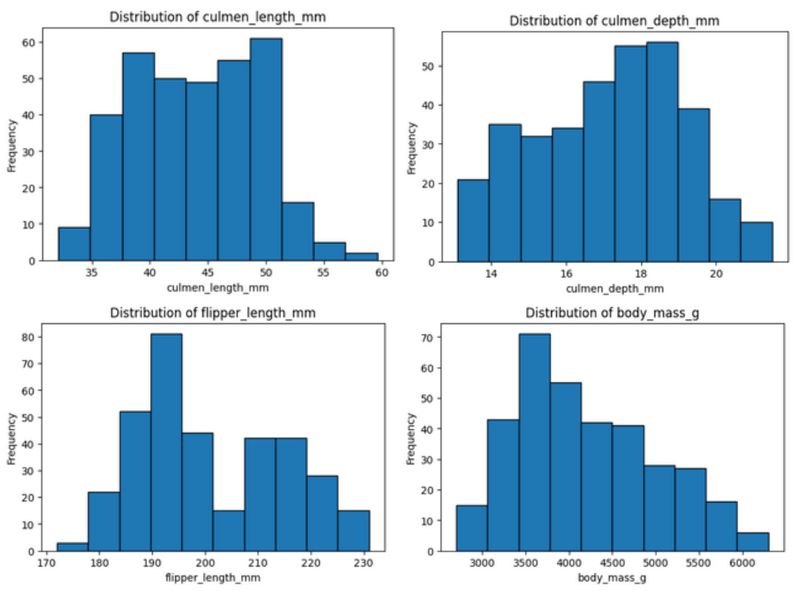
| | Missing Values |
|---|---|
| species | 0 |
| island | 0 |
| culmen_length_mm | 2 |
| culmen_depth_mm | 2 |
| flipper_length_mm | 2 |
| body_mass_g | 2 |
| sex | 10 |

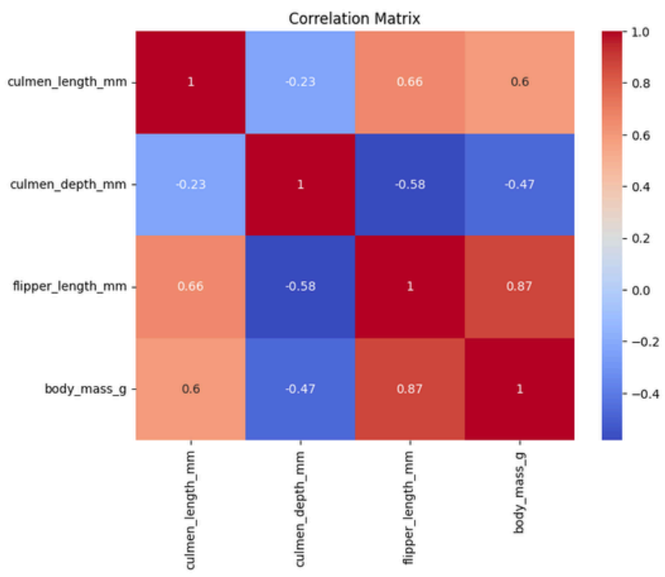
```
#Handling missing values
from sklearn.impute import SimpleImputer

#replace the value '.' in the column 'sex' to nan
df['sex'] = df['sex'].replace('.', np.nan)

#setting strategy to 'most frequent'
imputer = SimpleImputer(strategy='most_frequent')
df.iloc[:,:] = imputer.fit_transform(df)
df.isnull().sum()
```

Fill in the missing values with the most frequent value in each column that has missing data.

## NUMERICAL FEATURES

Data Distribution & Correlation between Features



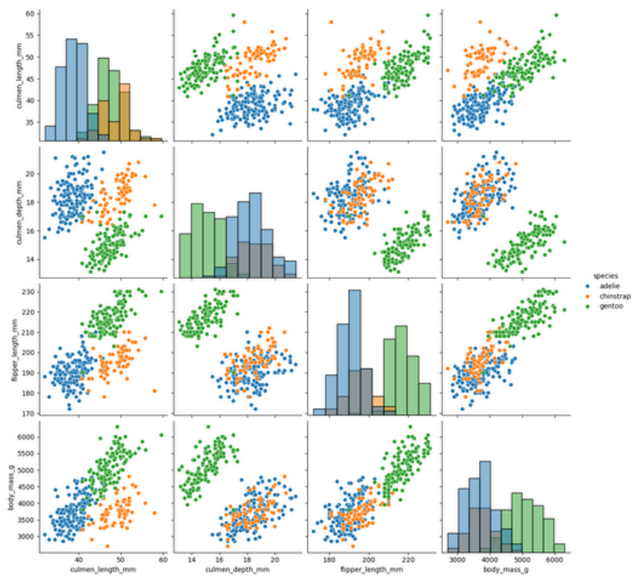

Correlation Matrix shows the strength of the relationship between features.

- a number close to 1 or -1 → have strong correlation
- 0 → don't have a correlation

(Sarkar et al., 2017)

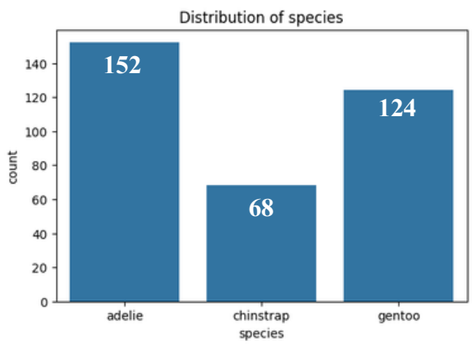
Features that have strong correlation

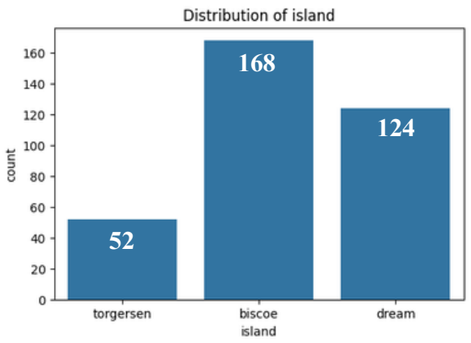
- flipper_length_mm & body_mass_g (0.87)
- culmen_length_mm & flipper_length_mm (0.66)
- culmen_length_mm & body_mass_g (0.6)

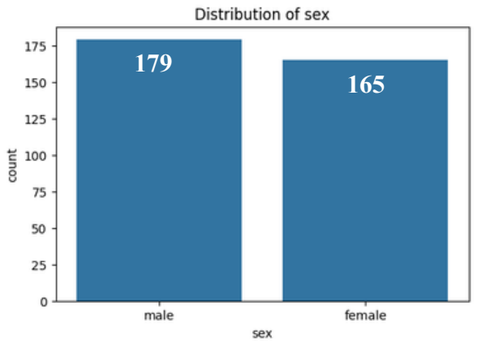


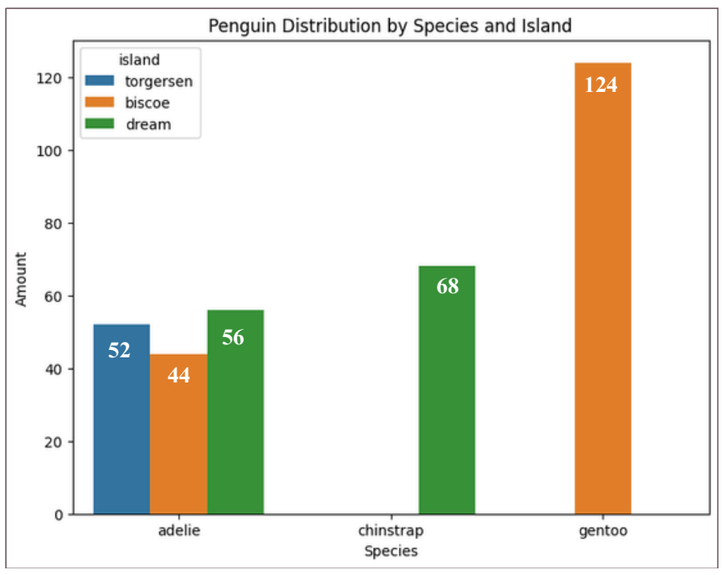

## CATEGORICAL FEATURES

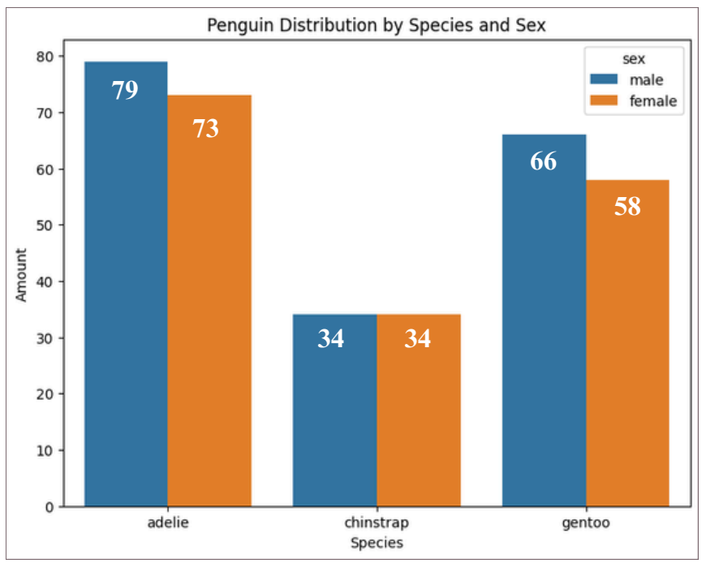
Data Distribution



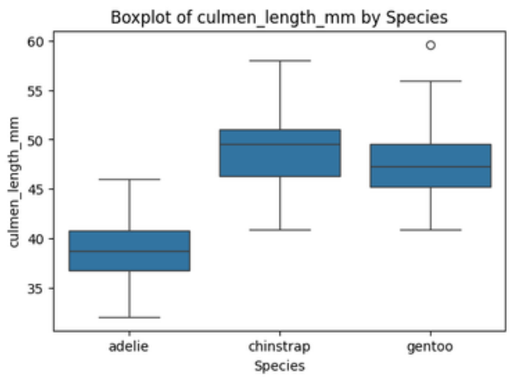



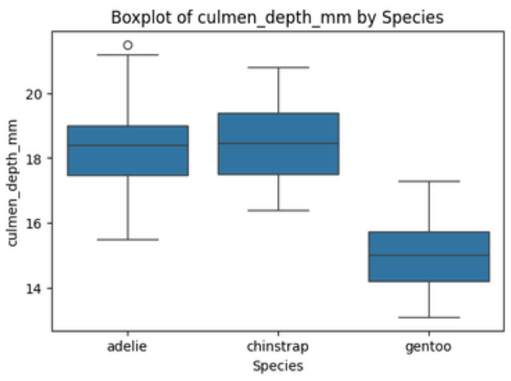



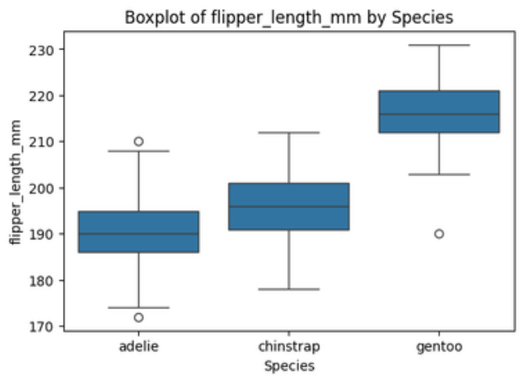



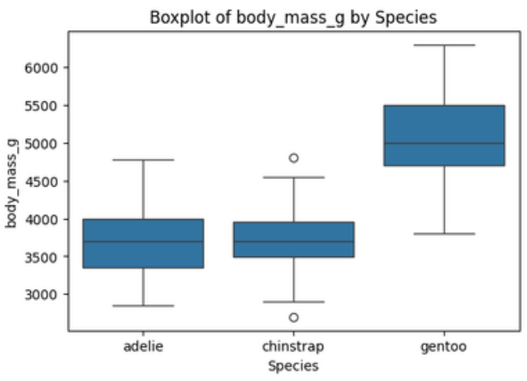

## NUMERICAL & CATEGORICAL FEATURES

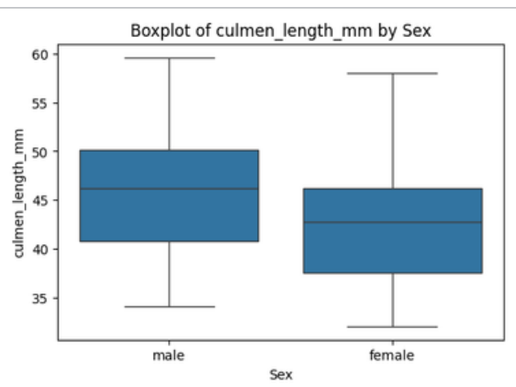
Boxplot

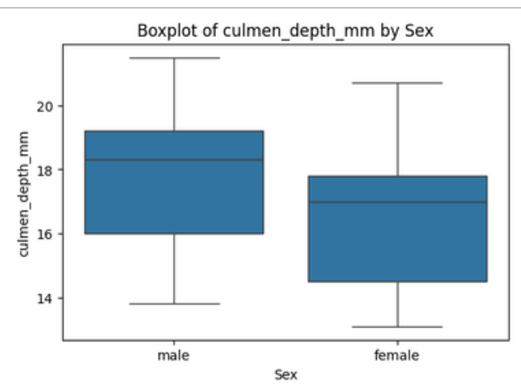


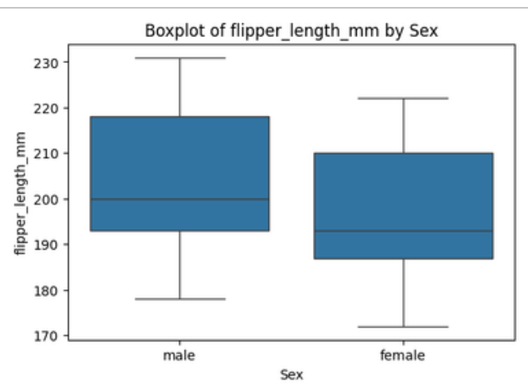



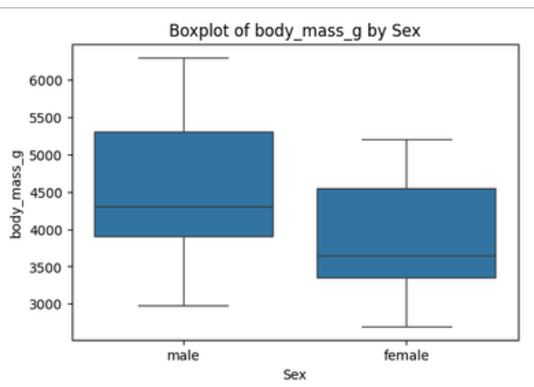



# SUPERVISED MACHINE LEARNING WITH PALMER ARCHIPELAGO (ANTARTICA) PENGUIN DATASET

Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine

by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1EuEOJQf1u2GyGMOKkWkAV6AwdAPpM0Qt?usp=sharing


Three species of Palmer Archipelago Penguin (Pfeifer et al., 2025).

### Convert the categorical columns into numerical value

The *get_dummies( )* function in pandas is used to convert categorical variables into numerical form by creating one-hot encoding (Sarkar et al., 2017). It transforms each unique category of a categorical variable into a separate column (indicator variable) with binary values (0 or 1). If a row belongs to a certain category, the corresponding column will have 1, while all others will have 0.

```
df_encoded = pd.get_dummies(df, columns=categorical_columns, drop_first=False)
df_encoded = df_encoded.replace({True: 1, False: 0})
print(df_encoded.head())
```

| | culmen_length_mm | culmen_depth_mm | flipper_length_mm | body_mass_g | species_adelie | species_chinstrap | species_gentoo | island_biscoe | island_dream | island_torgersen | sex_female | sex_male |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 39.1 | 18.7 | 181.0 | 3750.0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 1 | 39.5 | 17.4 | 186.0 | 3800.0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 40.3 | 18.0 | 195.0 | 3250.0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 3 | 41.1 | 17.0 | 190.0 | 3800.0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 4 | 36.7 | 19.3 | 193.0 | 3450.0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

## 3. PREPARATION OF TRAIN DATA AND TEST DATA

### Split the data into train & test data

80% Train Data → 275 Data
20% Test Data → 69 Data

```
X = df_encoded.drop(columns=['species_adelie', 'species_chinstrap', 'species_gentoo'])
y = df['species']
```

```
# 80% train, 20% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

### Standardization for Numerical Features

- Standardization works by subtracting the mean from each value and then dividing by the standard deviation.
- It is important to fit the scalers to the training data only, not to the full dataset (including the test set). Only then can you use them to transform the training set and the test set (and new data).
- Standardization is a type of feature scaling (along with Min-Max scaling). Feature scaling is important when numerical attributes have very different scales to ensure that the machine learning model performs well.

(Albon, 2018; Geron, 2017).

## 4. MACHINE LEARNING MODELS

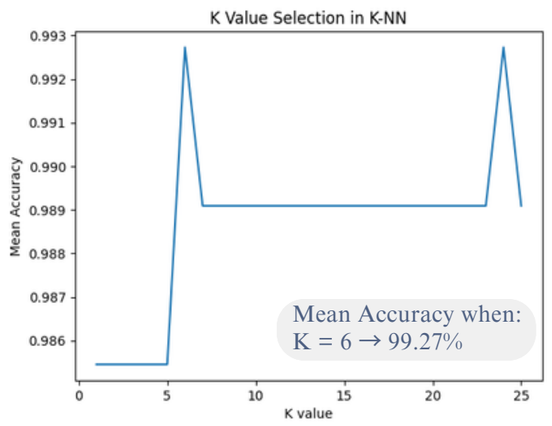
### 4.1 Logistic Regression

```
Accuracy of Logistic Regression Model: 100.00%

Classification Report:
              precision    recall  f1-score   support

      adelie       1.00      1.00      1.00        34
   chinstrap       1.00      1.00      1.00        11
      gentoo       1.00      1.00      1.00        24

    accuracy                           1.00        69
   macro avg       1.00      1.00      1.00        69
weighted avg       1.00      1.00      1.00        69
```

Accuracy = 100%

### 4.2 K-Nearest Neighbor

1. Search the best K



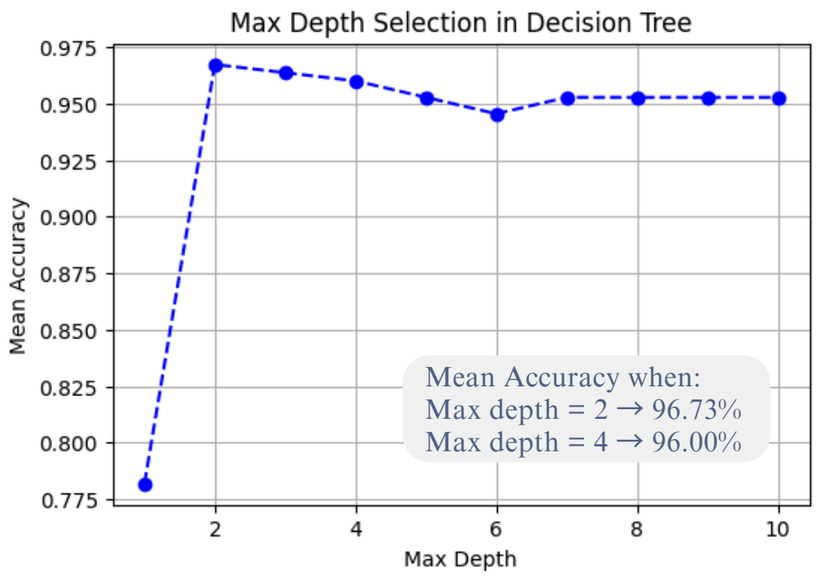

2. Search accuracy of K-NN after applying it to test data

```
Accuracy of KNN model = 100.00%

Classification Report:
              precision    recall  f1-score   support

      adelie       1.00      1.00      1.00        34
   chinstrap       1.00      1.00      1.00        11
      gentoo       1.00      1.00      1.00        24

    accuracy                           1.00        69
   macro avg       1.00      1.00      1.00        69
weighted avg       1.00      1.00      1.00        69
```

Accuracy = 100%

### 4.3 Decision Tree

1. Search the best depth




2. Compare the accuracy of Decision Tree models for both Max depth values after applying them to test data

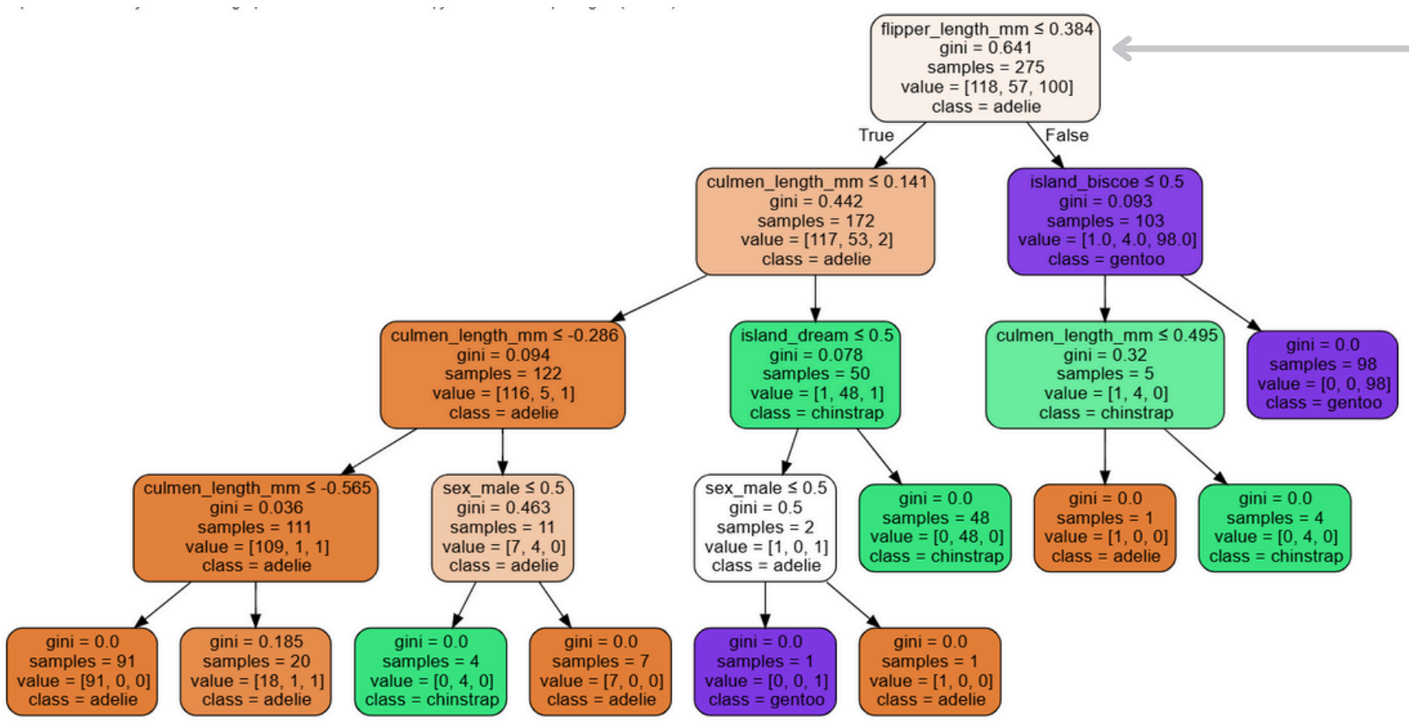
```
When Max Depth = 2, The Accuracy of Decision Tree Model: 94.20%

Classification Report:
              precision    recall  f1-score   support

      adelie       0.97      0.91      0.94        34
   chinstrap       0.77      0.91      0.83        11
      gentoo       1.00      1.00      1.00        24

    accuracy                           0.94        69
   macro avg       0.91      0.94      0.92        69
weighted avg       0.95      0.94      0.94        69
```

```
When Max Depth = 4, The Accuracy of Decision Tree Model: 100.00%

Classification Report:
              precision    recall  f1-score   support

      adelie       1.00      1.00      1.00        34
   chinstrap       1.00      1.00      1.00        11
      gentoo       1.00      1.00      1.00        24

    accuracy                           1.00        69
   macro avg       1.00      1.00      1.00        69
weighted avg       1.00      1.00      1.00        69
```

Max Depth = 2 (94.20% accuracy): The model is too simple (underfitting) and doesn't capture enough details, leading to a lower test accuracy, even though it had the best mean accuracy during training.

Max Depth = 4 (100% accuracy): The model learns more details, improving its ability to generalize to test set.

3. Decision Tree Visualization




How to Read Decision Tree :

1. Start at the root node (the top).
2. Move **left** if the condition is **True**, or **right if False**.
3. Continue down the branches until you reach a leaf node.
4. The class label at the leaf node is the predicted class

# SUPERVISED MACHINE LEARNING WITH PALMER ARCHIPELAGO (ANTARTICA) PENGUIN DATASET

Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine

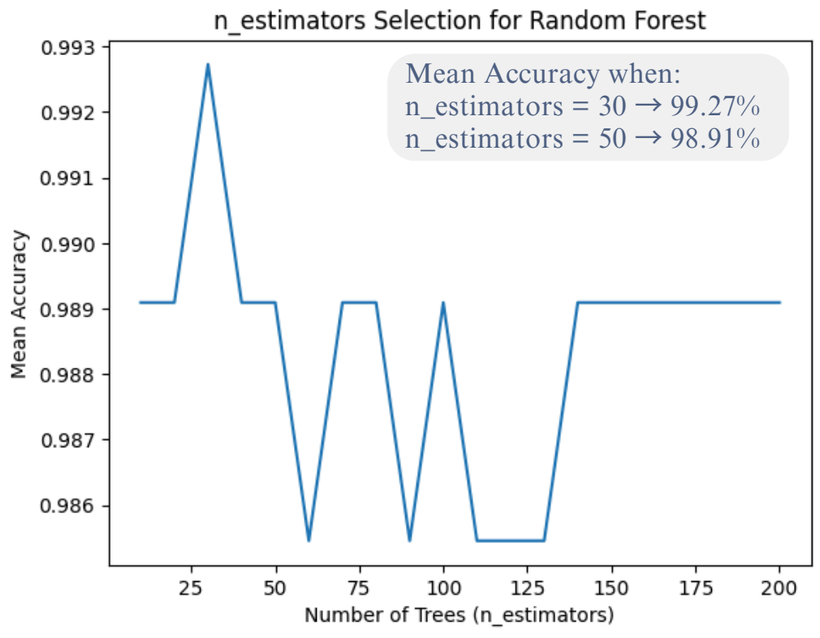
by: Raniah Mufidah Admayana

Code in Google Colab = https://colab.research.google.com/drive/1EuEOJQf1u2GyGMOKkWkAV6AwdAPpM0Qt?usp=sharing


Three species of Palmer Archipelago Penguin (Pfeifer et al., 2025).

## 4.4 Random Forest

### 1. Search the best n_estimators




### 2. Search accuracy of Random Forest after applying them to test data

with n_estimators = 30,

Accuracy of Random Forest Model: 98.55%

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| adelie | 1.00 | 0.97 | 0.99 | 34 |
| chinstrap | 0.92 | 1.00 | 0.96 | 11 |
| gentoo | 1.00 | 1.00 | 1.00 | 24 |
| accuracy | | | 0.99 | 69 |
| macro avg | 0.97 | 0.99 | 0.98 | 69 |
| weighted avg | 0.99 | 0.99 | 0.99 | 69 |

with n_estimators = 50,

Accuracy of Random Forest Model: 100.00%

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| adelie | 1.00 | 1.00 | 1.00 | 34 |
| chinstrap | 1.00 | 1.00 | 1.00 | 11 |
| gentoo | 1.00 | 1.00 | 1.00 | 24 |
| accuracy | | | 1.00 | 69 |
| macro avg | 1.00 | 1.00 | 1.00 | 69 |
| weighted avg | 1.00 | 1.00 | 1.00 | 69 |

n_estimators = 30 (98.55% accuracy): The model performed well during training but slightly worse on X_test because 30 trees might not have captured enough patterns in the data.

n_estimators = 50 (100% accuracy): Adding more trees improved the model's ability to generalize, reducing randomness and making predictions more stable, leading to perfect accuracy on X_test.

## 4.5 Support Vector Machine

### 1. Search the best kernel

Kernel = linear: Mean Accuracy = 0.9818
Kernel = poly: Mean Accuracy = 0.9855
Kernel = rbf: Mean Accuracy = 0.9855
Kernel = sigmoid: Mean Accuracy = 0.9964

### 2. Search accuracy of SVM after applying them to test data

Accuracy = 100% with any of the kernel

the accuracy of SVM Model: 100.00%

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| adelie | 1.00 | 1.00 | 1.00 | 34 |
| chinstrap | 1.00 | 1.00 | 1.00 | 11 |
| gentoo | 1.00 | 1.00 | 1.00 | 24 |
| accuracy | | | 1.00 | 69 |
| macro avg | 1.00 | 1.00 | 1.00 | 69 |
| weighted avg | 1.00 | 1.00 | 1.00 | 69 |

## CONCLUSION

Accuracy of =
- Logistic Regression → 100%
- K-NN → 100%
- Decision Tree → 100%
- Random Forest → 100%
- SVM → 100%

All model is the best machine learning model for Penguin dataset

## CODE IN COLAB

Google Colab
co google.com
https://colab.research.google.com/drive/1EuEOJQf1u2GyGMOKkWkAV6AwdAPpM0Qt?usp=sharing

## REFERENCE

Albon, C. (2018). Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning.

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Techniques and Tools to Build Learning Machines. O'Reilly Media.

Horst, A. M., Hill, A. P., & Gorman, K. B. (2022). Palmer Archipelago Penguins Data in the palmerpenguins R Package-An Alternative to Anderson's Irises. R Journal, 14(1).

Marsland, S. (2014). Machine learning: An Algorithmic Perspective, Second Edition. CRC Press.

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

Pfeifer, C., Knetsch, S., Maercker, J., Mustafa, O., Rümmler, M. C., & Brenning, A. (2025). Exploring the potential of aerial drone imagery to distinguish breeding Adélie (Pygoscelis adeliae), chinstrap (Pygoscelis antarcticus) and gentoo (Pygoscelis papua) penguins in Antarctica. Ecological Indicators, 170, 113011.

Sarkar, D., Bali, R., & Sharma, T. (2017). Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems. Apress.