

ADS Assignment 3

Penguin Classification Analysis

Problem Statement:

The Penguin Classification Analysis problem involves predicting the species of a penguin based on various physical characteristics. The dataset includes information about the body mass, culmen length, culmen depth, flipper length, and sex of different penguin species.

The problem is typically approached as a classification problem, where the target variable is the penguin species, and the features are the physical characteristics of the penguins.

Accurate classification of penguin species can also help researchers understand the effects of climate change and other environmental factors on penguin populations. The problem can also be useful for conservation efforts, as it can help identify and protect endangered penguin species.

Attribute Information:

- Species: penguin species (Chinstrap, Adélie, or Gentoo)
- Island: island name (Dream, Torgersen, or Biscoe) in Antarctica
- culmen_length_mm: culmen length (mm)
- culmen_depth_mm: culmen depth (mm)
- flipper_length_mm: flipper length (mm)
- body_mass_g: body mass (g)
- Sex: penguin sex

What is culmen?

The upper margin of the beak or bill is referred to as the culmen and the measurement is taken using calipers with one jaw at the tip of the upper mandible and the other at base of the skull or the first feathers depending on the standard chosen.

Perform the below Tasks to complete the Assignment:-

Clustering the data and performing classification algorithms

1. Download the dataset: [Dataset](#)
2. Load the dataset into the tool.
3. Perform Below Visualizations.

- Univariate Analysis
- Bi- Variate Analysis
- Multi-Variate Analysis

4. Perform descriptive statistics on the dataset.
5. Check for Missing values and deal with them.
6. Find the outliers and replace them outliers
7. Check for Categorical columns and perform encoding.
8. Scaling the data
9. Perform any of the clustering algorithms
10. Add the cluster data with the primary dataset
11. Split the data into dependent and independent variables.
12. Split the data into training and testing
13. Build the Model
14. Train the Model
15. Test the Model
16. Measure the performance using Evaluation Metrics.