Random Forest Penguin Classification Project

Project Overview

This project aims to build a Random Forest classification model to identify penguin species based on physical measurements. The dataset is sourced from the popular Palmer Penguins dataset.

Contents

Data Description

The dataset includes the following features:

- **species**: Target variable representing the penguin species (Adelie, Gentoo, Chinstrap)
- **island**: The island the penguin inhabits (Biscoe, Dream, Torgersen)
- **culmen_length_mm**: Culmen length in millimeters
- **culmen_depth_mm**: Culmen depth in millimeters
- **flipper_length_mm**: Flipper length in millimeters
- **body_mass_g**: Body mass in grams
- **sex**: Gender of the penguin

Data Preprocessing

- Missing values are removed.
- Categorical variables `island` and `sex` are converted to numerical features using one-hot encoding.
- Original data files remain unchanged; preprocessing is handled in a dedicated function within the `one_hot_encoder.py` module.

Modular Project Structure

- `one_hot_encoder.py`: Contains the `preprocess_penguin_data(filepath)` function which loads, cleans, and encodes the dataset.
- `data_inspector.py`: Loads the processed data using the encoder and provides utilities for data inspection (e.g., displaying the first rows).
- `random_forest.py`: Implements model training, prediction, evaluation, and visualization.

Machine Learning Workflow

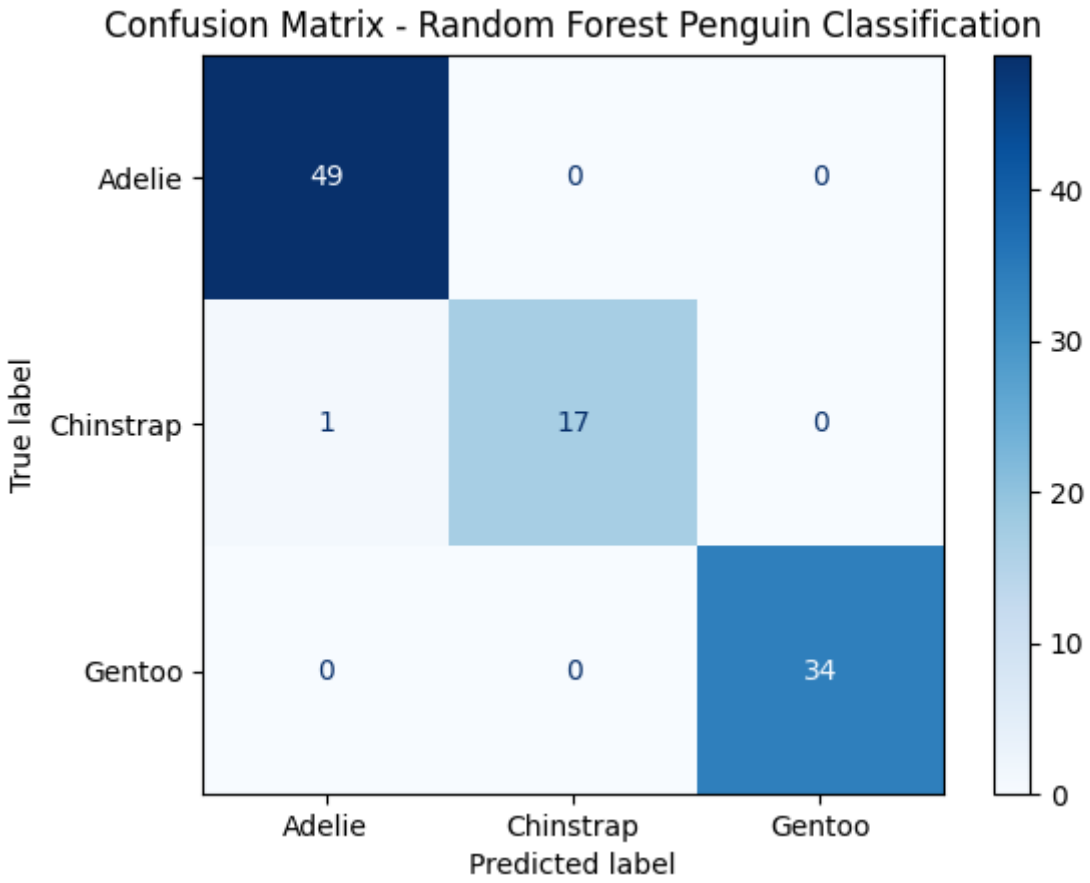
1. Load and preprocess data using `preprocess_penguin_data`.

2. Split data into training and testing sets.
3. Train a Random Forest classifier (`n_estimators=100`, using appropriate number of CPU cores).
4. Predict on test data.
5. Evaluate performance using accuracy, precision, recall, F1-score, and visualize results with a confusion matrix.

Model Performance

- Achieved high accuracy (~99%).
- Balanced classification performance across the penguin species.

Sample Dataset Preview



Visualization

- Confusion matrix visualization using `sklearn.metrics.ConfusionMatrixDisplay` to show true vs. predicted classifications.

Installation and Requirements

- Python 3.12.4 (recommended)

Required Python libraries (installation via pip):

- pandas
- numpy

- scikit-learn
- matplotlib