Assignment 20: Comparison of Machine Learning Algorithms to Classify Penguin Species

55 Points scaled to 20 Points

Introduction

In this assignment you will compare three different machine learning algorithms (Random Forest, Support Vector Machines, and *k*-Nearest Neighbor) for differentiating between three species of penguins found in Antarctica. The predictor variables include bill length, bill depth, flipper length, body mass, and sex. These data ("penguins.csv") were obtained from Kaggle (https://www.kaggle.com/datasets/larsen0966/penguins).

Objectives

- *Prepare data as input to machine learning algorithms*
- *Visualize relationships between variables*
- *Train and assess multiple machine learning algorithms*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and output embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and answered within Markdown cells.*

Questions

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following task and answer the associated questions.

Task 1: Subset out the following columns from the dataset: "species", "bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g", and "sex". In the models, "species" is the dependent variable and should be treated as nominal data. All predictor variables are continuous other than "sex", which should be treated as a nominal variable. Remove any rows that have missing data in any column. (5 Points)

Task 2: Generate grouped box plots to show the distribution of each continuous predictor variable within each species. Also, differentiate the data by sex within each species. Based on these visualizations, do the penguin species appear to have different characteristics? Do characteristics vary by sex, indicating that sex should be included as a predictor variable? What variables seem most predictive? Generally, do you anticipate being able to separate the species with high accuracy? (10 Points)

Task 3. Split the data into separate training and testing sets stratified by the species type. (5 Points)

Task 4. Center and scale all the continuous predictor variables. This is required for using the Support Vector Machine and k-Nearest Neighbor algorithms. Also, create dummy variables from the "sex" variable. (5 Points)

Task 5. Use the training partition to train Random Forest, Support Vector Machine, and *k*-Nearest Neighbor models. For Random Forest, use 100 trees and set the number of predictor variables available for splitting at each node hyperparameter to 3. Support Vector Machines can use the default hyperparameters. For k-Nearest Neighbor, set the number of neighbors to 11. Note that it would be best to tune the hyperparameters to offer a fairer comparison of the algorithms. However, we are not asking you to do so here to save time. (10 Points)

Task 6. Use the train models to predict to the withheld validation data. Use the results to create Confusion Matrices and the Overall Accuracy metrics. Discuss your results. How do the algorithms compare for this specific problem? Were the species generally separated with high or low accuracy? (10 Points)

Task 7. Repeat the algorithm comparison. However, this time only use the "body_mass_g" as a predictor variable. How do the three algorithms compare if only one predictor variable is provided? How does the performance using only one predictor variable compare to that obtained when using the entire set of predictors? (10 Points)