Data Science Test 1 - Penguins classification

October 4, 2024

1 Introduction

This exercise focuses on setting up a classification algorithm using a publicly available dataset. We have chosen the Penguin data set, which was collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long-Term Ecological Research Network.

The classification task does not necessitate complex methodologies; simple machine-learning algorithms, such as Random Forest, provide excellent classification results.

The report is structured as follows. Section 2 presents the dataset along with the trends and features identified during the Exploratory Data Analysis (EDA). Then Sect. 3 explains the methods implemented in this study and present the main results.

2 Data

The Penguins data set is a collation of 333 penguins, collected as part of research conducted at Palmer Station, Antarctica. For each penguin, the data set provides the following featues:

- **ID**: A unique identifier for each penguin.
- **Species**: The species of each penguin, which is the target variable for prediction. The possible species are Adelie, Gentoo, and Chinstrap.
- **Island**: The island where the penguins were observed, with three possible values: Torgersen, Biscoe, and Dream.
- **Bill Length**: The length of the penguin's bill.
- **Flipper Length**: The length of the penguin's flipper.
- **Body Mass**: The mass of the penguin's body.
- **Sex**: The gender of the penguin.
- **Year**: The year the data was collected, which includes 2007, 2008, and 2009.

To set up a classifier for distinguishing among penguin species, we will first explore the dataset to assess the quality and importance of the features provided for this task. Given the relatively small size of the dataset, we will focus on simpler machine learning algorithms, such as logistic regression, or random forest, which are more suitable for this scale and complexity.

Figure 1 shows the correlation matrix of the features. We are interested in the correlation between the species and other features. We note that several features are highly correlated with the species, with the bill length and the body mass being the most correlated ones. This will ease the classification, since correlations indicate that there are information available in our data.

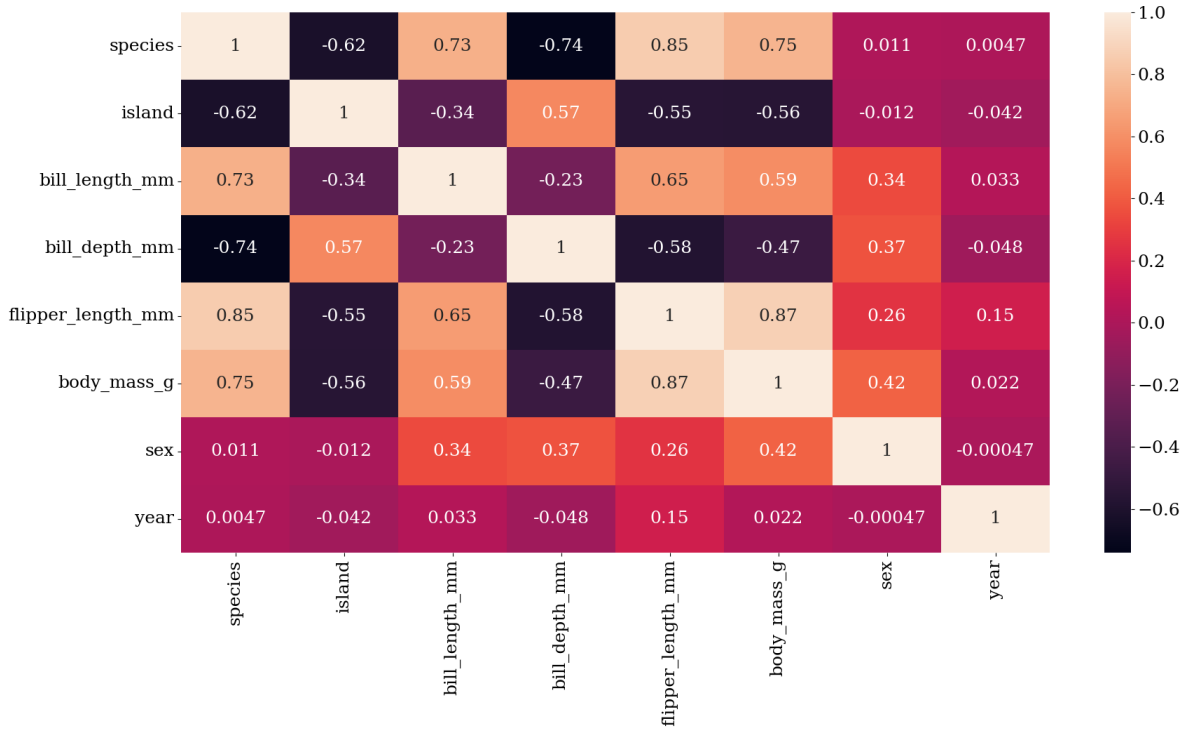

Figure 1: Correlation matrix of the penguin dataset features. The correlation matrix indicates that strong correlation of all features but 'Sex' and 'Year'.

The plot also exhibits no significant correlation between species and sex, suggesting a gender balance across the species. Additionally, there is no observable correlation between species and year, which implies that the population of penguins did not undergo significant changes during the recorded years.

We test these hypothesis inn Fig. 2, which shows a histogram of the species population splitted in gender (left) and the population evolution of each species over time (right). The plots support both statetements: there is a gender balance among all species, and the population has remained relatively stable over the years. Based on this results, we will exclude the gender and year from our classification algorithm, as they do not contribute meaningful information to the prediction.

The left panel of Fig. 2 already shows a data imbalace among species. To see this clearly, Fig. 3 illustrates the distribution of penguin species (left) without splitting by geneder we note that while Adelie and Gentoo penguis are represented with 44% and 36% of the datas, Chinstrap penguins only correspond to 20%. The underrepresentation of Chinstrap penguins compared to other species may lead to biased model performance, where the model could be skewed toward predicting the more frequent classes. It will be important to assess whether this imbalance affects the model's ability to generalize effectively across all species.

The panel on the right shows the penguin populations across the islands. In this case, it is evident that island distribution is imbalanced. For instance, the Gentoo penguin is found exclusively on Biscoe Island, suggesting that the island feature could serve as a unique identifier for certain species, making it a highly relevant feature for species classification.

While this observation is useful, it also raises a potential concern: if the dataset is not fully representative, the absence or underrepresentation of certain species on specific islands could introduce bias into our model. For example, if Gentoo penguins are underrepresented or missing from data for Biscoe Island, the model might struggle to accurately predict their presence, leading to skewed results and reduced

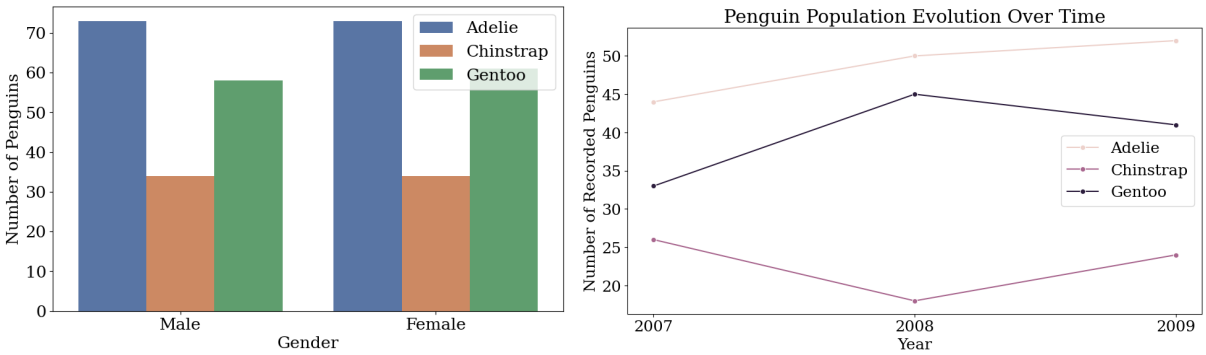

Figure 2: *Left*: Distribution of penguin species split by gender. The similar distribution across all species for both genders suggests no correlation between species and gender, consistent with the findings of the correlation matrix. *Right*: Temporal evolution of the species population, showing no significant changes over the recorded years.

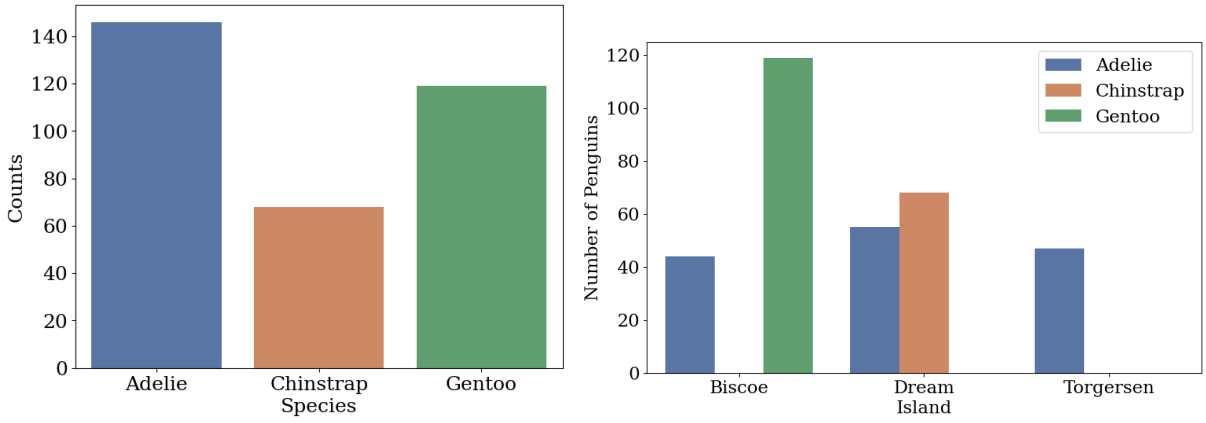

Figure 3: *Left*: Distribution of penguin species, showing an imbalanced dataset, with Chinstrap penguins being underrepresented. *Right*: Population of penguins across different islands, highlighting significant differences, including the exclusive presence of Gentoo penguins on Biscoe Island.

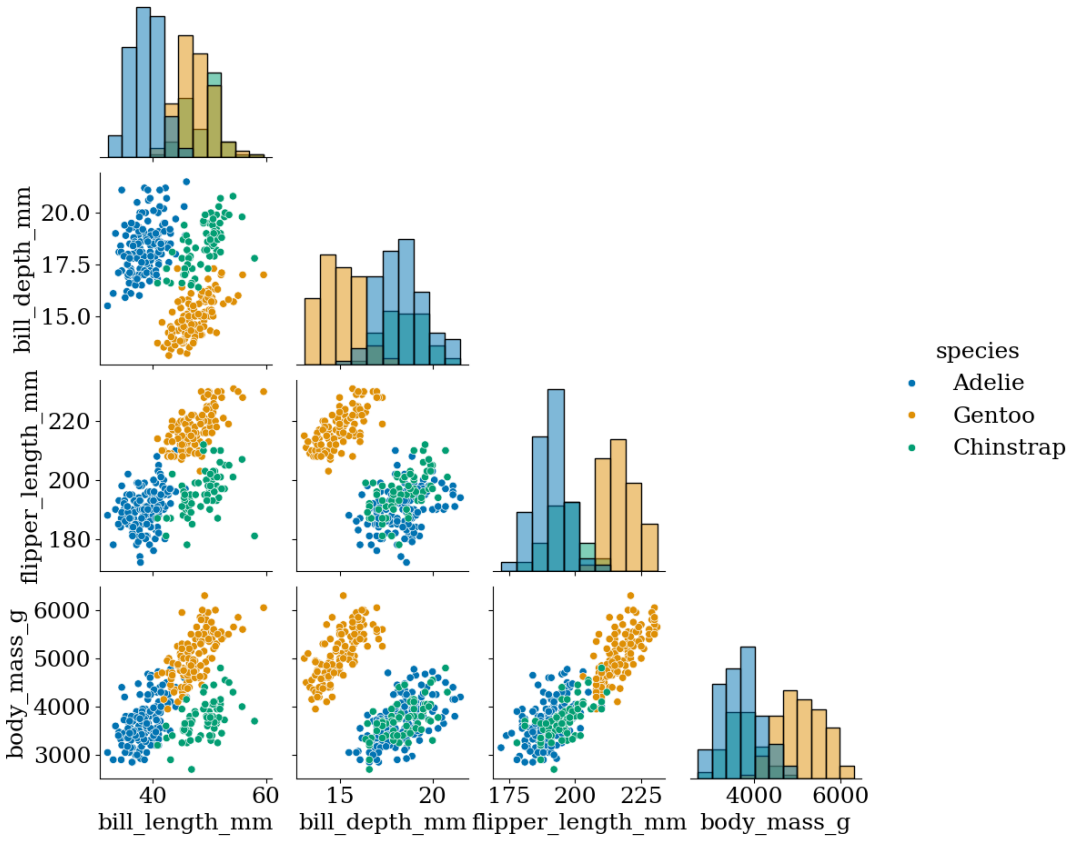

Figure 4: Relationship among features for each species.

generalizability.

Finally, Fig.4 illustrates the relationships among variables for each species. Gentoo penguins show distinct clustering in all plots, standing out clearly from the other species. Conversely, Adelie and Chinstrap penguins exhibit significant overlap, particularly in feature spaces such as flipper length vs. body mass, suggesting these species may be more challenging to differentiate based on these variables alone.

In plots involving bill length combined with other features, the separation among species becomes much clearer, especially in the bill length vs. flipper length space. This strong separation indicates that classification could be straightforward; even a simple classification approach using this plane could yield high accuracy, although with some overlap between species. These findings suggest that the chosen features carry discriminatory information for building an effective classifier.

3 Methods and results

As discussed in Sect.2 with the EDA, the simplicity and structure of the input data suggest that basic machine learning algorithms are well-suited for this classification task. The clear distinctions observed in certain feature spaces, such as those shown in Fig.4, further support this approach. Given the limited number of structured features and the apparent separability of the species in some dimensions, even low-complexity models are likely to achieve robust classification performance

The algorithms that were tested are

- Gradient Boosting (GB)
- Random Forest (RF)
- Logistic Regression (LR)
- Support Vector Machine (SVM)

, for which we use their SK-LEARN implementation.

Before implementing the classification algorithms, we preprocessed the raw input data. The 333 samples were first randomly split into a training set and a test set to ensure there was no cross-contamination between them. The two sets of data are found in the data directory. The ID, Year and Sex columns were excluded from the dataset. We encoded the Island and Species categorical features into integer representations suitable for ingestion by the algorithms. Subsequently, we scaled the remaining features. This scaling process was performed separately on the training and test sets to ensure that the scaling of the test set was not influenced by the characteristics of the training set.

The four methods were trained and subsequently validated on the test split, with each model being registered using MLFLOW. The performance of the models could then be analyzed through a notebook connected to the MLFlow registry.

All methods achieve an accuracy of 1.0 (see left panel on Fig.5), indicating perfect classification with no errors. This result suggests that predicting penguin species using this dataset is straightforward, and even basic machine-learning algorithms can reach perfect precision. Therefore, there is no need to complicate the task with more advanced models such as neural networks.

However, as noted earlier, model performance may be heavily influenced by a few dominant features, such as island. For instance, Gentoo penguins are exclusively recorded on Biscoe Island, introducing a potential sampling bias. Such a reliance on specific categorical features could negatively affect model generalization if the feature distribution differs in an external test set.

To further explore this, we assessed feature importance using the Gradient Boosting (GB) algorithm and found that the most influential features are indeed bill and flipper lengths (see right panel of Fig.5). This outcome aligns with the scatter patterns in Fig.4, where species distinctions are clearest within this feature space. This finding is promising, as these key features are continuous variables, making them less susceptible to the biases associated with categorical features, such as island.

3.1 Model deployment

For simplicity and given that the four algorithms yield the same perfornace, we only deploy the GB model. This is deployed to production in two different ways: via the Weights and Biases (WaB) platform and through a Docker container. The WaB platform provides a user-friendly interface for making predictions from the trained model, but it is limited by the inability to automate the process and its restriction to handling only one prediction at a time.

In contrast, the Docker container offers a more practical solution. It includes a notebook that facilitates predictions allowing for more flexible and scalable usage.

4 Conclusions

In conclusion, our analysis of the Palmer Penguins dataset demonstrated that simple machine-learning algorithms allowed us to achieve perfect accuracy in classifying penguin species. This performance is driven by the informative nature of the data but underscores the effectiveness of low-complexity machine learning techniques in scenarios where clear patterns exist.

However, it is crucial to recognize potential biases that could arise from reliance on specific features, such as the island variable. Our examination of feature importance revealed that bill and flipper length

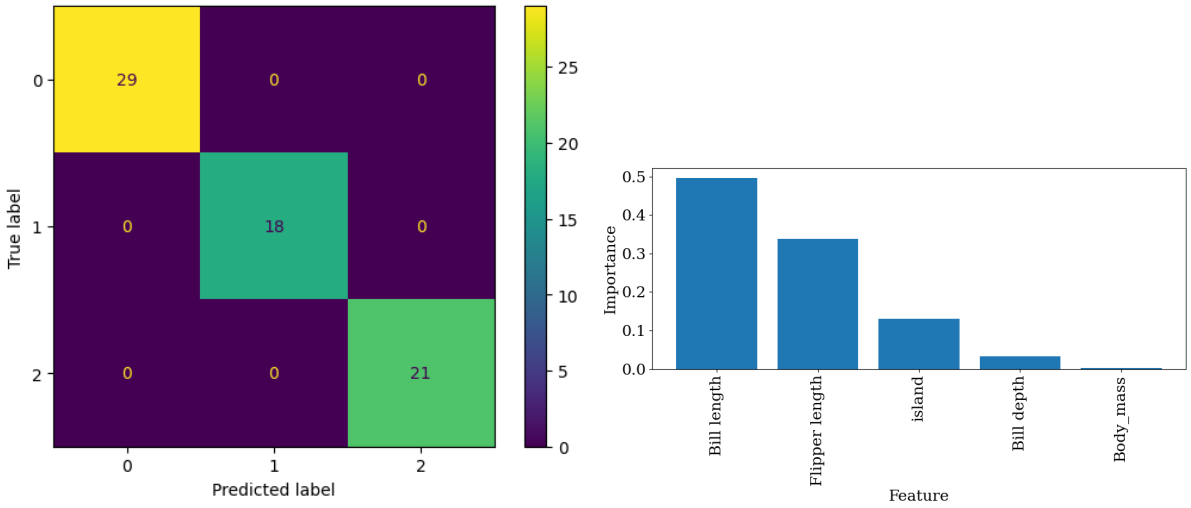

Figure 5: *Left*: Confussion matrix of the classification usingGB. *Right*: Feature importance with the GB model.

were the most significant predictors, suggesting a deeper understanding of the data is necessary to ensure robust predictions across diverse conditions.