Data Science Test 1 - Penguins classification

October 3, 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 finally Sect. **??** discusses the performance of the methods and analyzes the results.

2 Data

The dataset contains features of penguins collected as part of research conducted at Palmer Station, Antarctica. The available features are as follows:

- **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.

Figure 1 shows the correlation matrix of the features. We are particularly interested in the correlation between the species and other features. The plot indicates 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.

Figure 2 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 hypotheses: there is a gender balance among all species, and the population has remained relatively stable over the years.

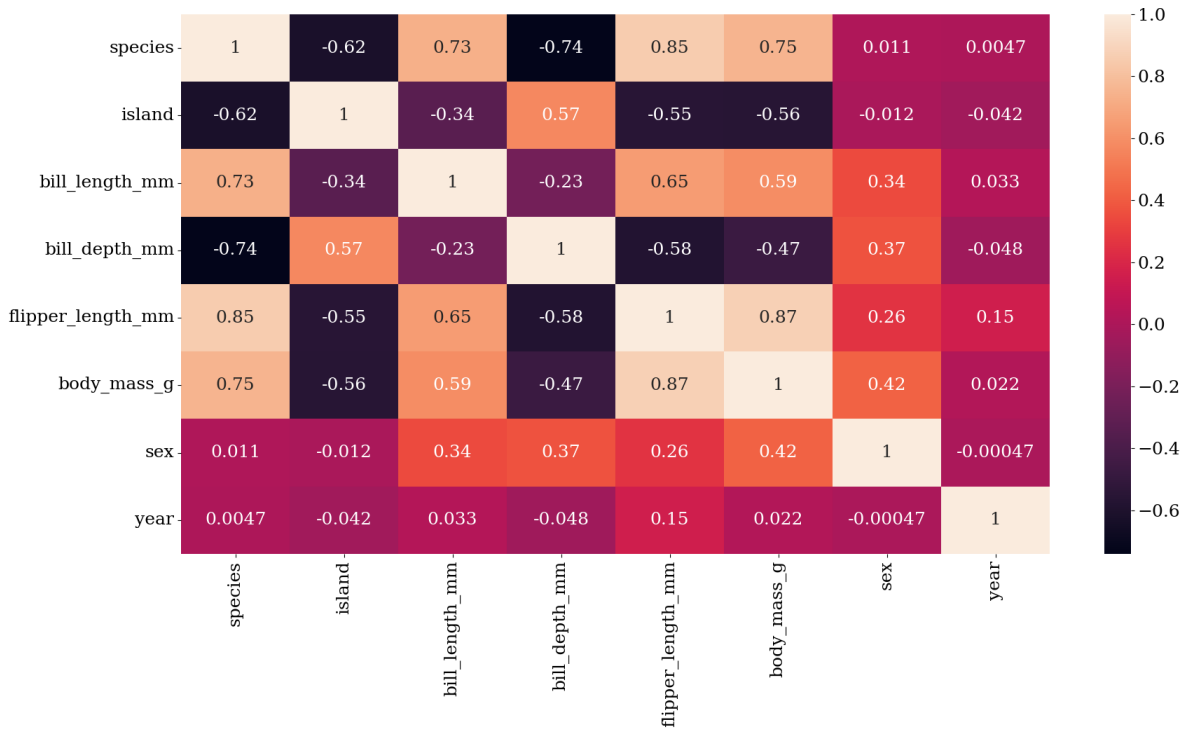

Figure 1: Correlation matrix of the penguin dataset features. The correlation matrix indicates that strong correlation of all features but 'Sex' and 'Year'.

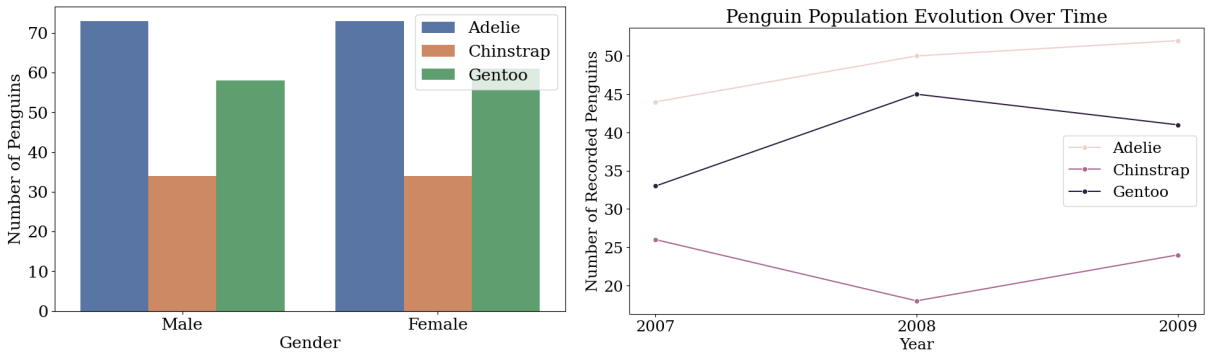

Figure 2: *Left*: Distribution of penguin species split by gender. The similar distribution across all species for both genders suggests no correlation between species and gender, consistent with the findings of the correlation matrix. *Right*: Temporal evolution of the species population, showing no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

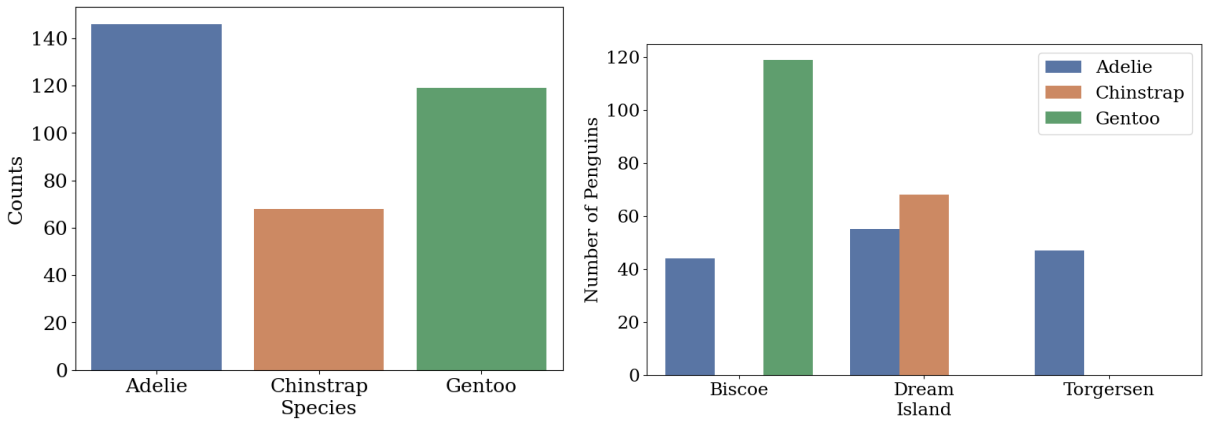

Figure 3: *Left*: Distribution of penguin species, showing an imbalanced dataset, with Chinstrap penguins being underrepresented. *Right*: Population of penguins across different islands, highlighting significant differences, including the exclusive presence of Gentoo penguins on Biscoe Island.

Based on the correlation analysis and additional tests performed, we will exclude the gender and year features from our classification algorithm, as they do not contribute meaningful information to the prediction.

Figure **??** illustrates the distribution of penguin species (left) and the population of each species distributed across islands (right). The species distribution chart reveals that the dataset is imbalanced, with some species underrepresented. This class imbalance can pose challenges for predictive modeling, especially in classification tasks. In the penguin dataset, 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 plot on the right shows the penguin populations across the islands. In this case, it is evident that island location plays a significant role in distinguishing penguin species. 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 generalizability.

3 Method, results, and discussion

After the EDA presented in Sect. 2, we considered that simple machine-learning algorithms could already achieve good classification performance. This is motivated both by the type of input data (few structured data), and by having clear correlation of these features with the species, as e.g. the Island.

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 split into a training set and a test set to ensure there was no cross-contamination between them. The ID, Year and Sex columns were expcluded 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 achieved an accuracy of 1.0 (see left panel on Fig.4), 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 attain perfect precision. Therefore, there is no need to complicate the task with more advanced models such as neural networks.

However, as previously noted, the models may heavily rely on a few key features, such as island in this case. Any sample bias in such features could significantly impact the models' performance when applied to an external test set, where the feature distribution might differ from that of the training set.

To further investigate this, we evaluated the feature importance of the Gradient Boosting (GB) algorithm and found that the most important features are, in fact, the bill and flipper lengths (see right panel on Fig.4). This is a positive outcome, as these are continuous variables that are less likely to be subject to the same biases that could affect categorical features, such as island.

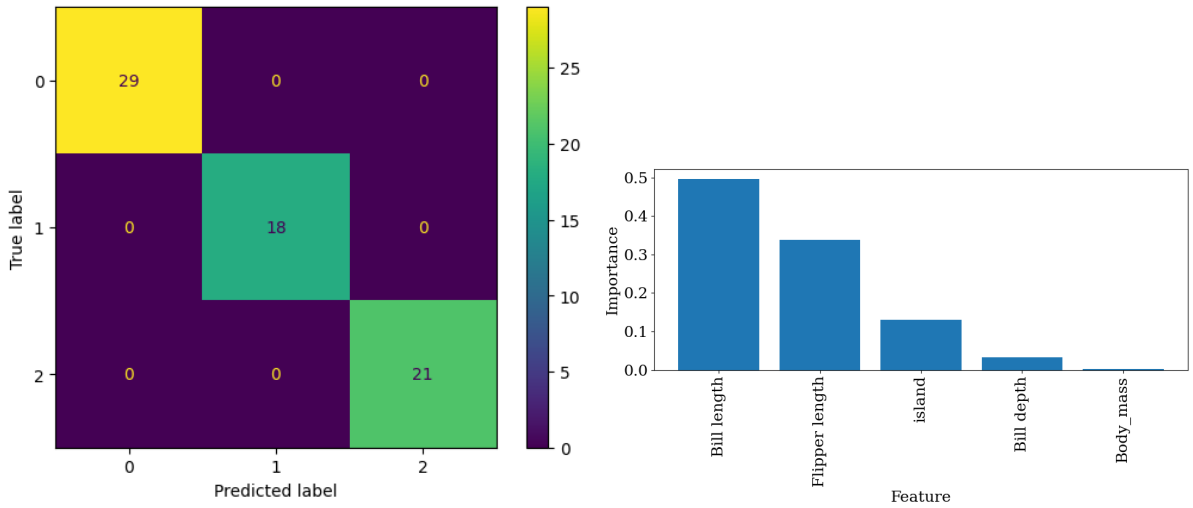

Figure 4: *Left*: Confussion matrix of the classification usingGB. *Right*: Feature importance with the GB model.

The GB algorithm was 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 from the deployed model, allowing for more flexible and scalable usage.

4 Conclusions

In conclusion, our analysis of the Palmer Penguins dataset demonstrated that this relatively simple dataset allowed us to achieve perfect accuracy in classifying penguin species using the Gradient Boosting (GB) algorithm. This impressive performance 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 were the most significant predictors, suggesting a more nuanced understanding of the data is necessary to ensure robust predictions across diverse conditions.

We successfully deployed the GB algorithm through two distinct methods: the user-friendly WaB platform and a more flexible Docker container. While WaB facilitates easy predictions, it is limited in automation and scalability. In contrast, the Docker container provides a versatile solution for making multiple predictions, integrating seamlessly into broader workflows.