Penguin Species Classification Technical Report

Junbo Dong(12432995)

December 17, 2024

Abstract

For the penguin classification problem, this assignment first used the random forest model to rank the importance of each feature. By conducting ablation experiments, we selected the top 5 features as the key features for the next step of training. Before starting training, we filled the missing parts of the selected feature data with the mean and then trained it using the random forest method. By evaluating on the validation set, we obtained an amazing result of 100% accuracy.

Contents

1 Introduction

Penguin species classification is of great significance in ecological research, enabling us to know about the quantity and distribution of penguin species. This task typically necessitates utilizing various features, including culmen_length_mm, culmen_depth_mm, and flipper_length_mm, for predicting penguin species based on observed attributes. The goal of this report is to apply machine learning techniques to classify penguin species with the data from tran_data. We focus on selecting the most informative features, training a random forest classifier, and evaluating its performance, aiming to achieve high classification accuracy by optimizing the feature set and model parameters.

2 Methodology

We used the dataset provided by Blackboard, which includes physical characteristics such as flipper length, culmen length, body mass, and culmen depth. The dataset contains both training and test data, where the training set includes species labels, and the test set does not.
We first performed One-Hot Encoding to convert categorical variables (such as studyName) into numeric features. Missing values were handled by imputing the mean for numerical features.

2.1 Select Features

During the feature selection process, we used the feature importance scores generated by the random forest classifier. After evaluation, the top six features were selected based on feature importance for subsequent training. The reason is that starting from the seventh feature, the data type changes to one-shot encoding. This type of data has limited effect on improving the accuracy of classification problems, but it will cause a significant increase in data dimensions, which not only takes up more memory space, but also slows down training. Therefore, we finally decided to select only the first six important features for training, including:

Table 1: Feature importance ranking

	Feature	**Importance**
3	Flipper Length (mm)	0.185727
1	Culmen Length (mm)	0.131595
6	Delta 13 C (o/oo)	0.125321
2	Culmen Depth (mm)	0.112716
4	Body Mass (g)	0.101891

2.2 Model

In light of its proficiency in managing both categorical and numerical data, along with its resilience against overfitting, we opted to employ a Random Forest[1] Classifier. The model was trained by leveraging the pre-selected features, and the hyperparameters were fine-tuned through the application of cross-validation techniques. This approach enabled us to enhance the model's performance and generalization capabilities, ensuring its reliability and accuracy in handling diverse datasets.

2.3 Evaluation

In this task, we used F1-score to evaluate the performance of the classification model. F1-score is the harmonic mean of precision and recall, and is particularly suitable for cases with imbalanced classes. It comprehensively considers the classifier's misclassification (false positives) and omission (false negatives), thus providing a comprehensive evaluation criterion for multi-classification problems.

3 Results

3.1 Training Results

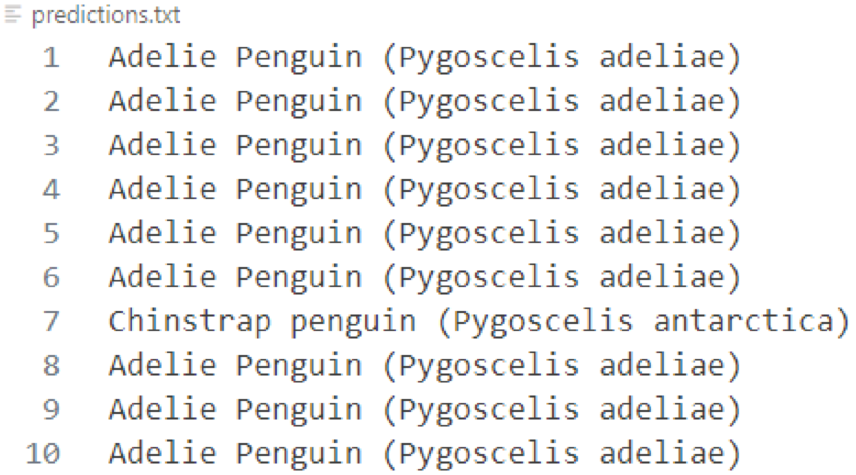
The Random Forest classifier achieved an amazing accuracy of 100% on the test set. The F1-scores for each class (species) were as follows:

Table 2: results

	precision	**recall**	**f1-score**	**support**
Adelie Penguin (Pygoscelis adeliae)	1.00	1.00	1.00	21
Chinstrap penguin (Pygoscelis antarctica)	1.00	1.00	1.00	12
Gentoo penguin (Pygoscelis papua)	1.00	1.00	1.00	13
accuracy			1.00	46
macro avg	1.00	1.00	1.00	46
weighted avg	1.00	1.00	1.00	46

3.2 Prediction results

Some prediction results on test_data.csv are as follows:



```
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Chinstrap penguin (Pygoscelis antarctica)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
Adelie Penguin (Pygoscelis adeliae)
```

Figure 1: Prediction results(the order matches the given .csv file)

3.3 Ablations

We conducted ablation experiments on the number of features selected by the model. When we used the first four features of importance, the f1-score of the model was reduced to 0.92, and the accuracy dropped significantly, with the lowest accuracy being 88%. When we added the fifth feature: Baby mass (g), the accuracy of the model increased to 100%. This shows that reducing the number of features has a significant impact on the prediction results of the model, and when we continue to increase the features, the accuracy of the model remains unchanged at 100%. Based on the experimental results and analysis, we then selected five features to train the model. These features include the first four most important features and the additional Baby mass (g) feature. When these five features are selected, the accuracy of the model reaches 100%, and the model performance is stable without adding too much complexity and can avoid overfitting.

4 Conclusion

This report demonstrates the effectiveness of using a random forest approach to classify penguin species. By selecting the 5 most important features through ablation experiments and applying appropriate data preprocessing, a classification accuracy of 100% can be achieved. Future work can continue to explore the limits of the model, using cross-validation and regularization to help better control the selection of features and further optimize the model.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.