Report
ML-classification project

Subject: 4IT439 Data-X – applied data analytics models in real world tasks

Team 4:
- Marek Styblik,
- Ondřej Šesták,
- Petr Hollmann,
- Sabína Rimarčíková,
- Tatiana Kliueva

Prague University of Economics and Business, 2023

Table of contents

1. Problem definition

The goal of this project is penguin species prediction. It is based on other data about penguins, which were noted about the penguins, like body parts sizes, sex, geography etc.

Data for the project come from an article "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)" by K. B. Gorman, T. D. Williams and W. R. Fraser [1].

Researchers gathered data on three penguin species: Adélie, Chinstrap and Gentoo penguins, which are displayed on the image 1.

		
Adélie penguin [2]	Chinstrap penguin [3]	Gentoo penguin [4]

Image 1. Penguin species photos

The way this dataset can be used: researcher measures bill and flipper sizes, body mass of the penguin and generates models to predict the “species” variable. In that case we can measure the success of the model with metrics like accuracy, precision, recall, F1 score.

2. Data Understanding

Dataset includes information about *Pygoscelis penguins* nesting on several islands within the Palmer Archipelago west of the Antarctic Peninsula near Anvers Island monitored during the austral summers of 2007/08, 2008/09, and 2009/10.

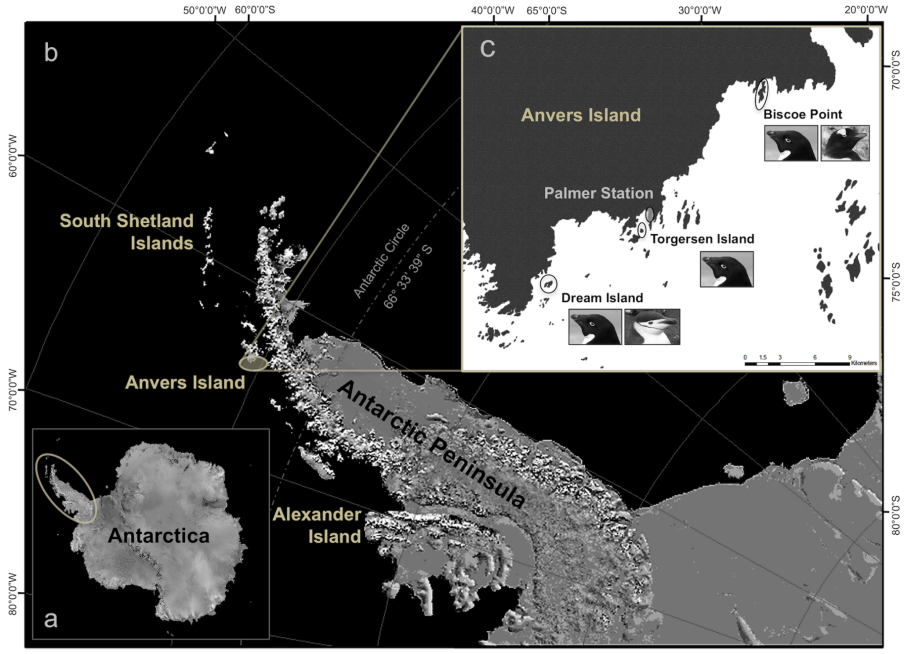


Image 2. Territory of research [1]

Dataset includes 8 columns. Column description is in a table below.

Table 1. Variable description

Variable	Description	Type
species	target variable, the species of a penguin	object
island	island of nesting	object
bill_length_mm	length of bill in millimeters	float64
bill_depth_mm	depth of bill in millimeters	float64
flipper_length_mm	length of flipper in millimeters	float64
body_mass_g	body mass in grams	float64
sex	sex of a penguin	object
year	year of monitoring	int64

Original dataset includes a total of 363 rows.

2.1. Identification of missing values, type of values, unique values

Variables "*species*", "*island*", "*year*" have no missing values.

Other variables include missing values: "*bill_length_mm*", "*bill_depth_mm*", "*flipper_length_mm*", "*body_mass_g*" - 5-6 missing values each, "*sex*" includes 15 missing values.

Sex column includes only "male" and "female" values. Since this is a binary variable we can recode using one-hot encoding it to binary variable "is_female" with values 0 and 1.

Species column includes categorical variables. As it is easier for the models to work with the numerical values, we have decided to give all three species their numerical values: Adelie = 0, Chinstrap = 1, Gentoo = 2.

Island column includes categorical variables. We can use one-hot encoding to get three binary columns with each island instead: island_Biscoe, island_Dream and island_Torgersen.

5 rows don't have information about any measures of the penguin, so we can remove it from model data, because they won't be helpful.

1 row has missing information in "flipper_length_mm", we can try to reconstruct it using information from other rows.

9 rows miss information in "sex" column. Penguins have significant sex dimorphism, so we can predict the sex based on sizes of penguins (more information in data preparation).

2.2. Descriptive statistics

The average bill length is about 44 mm, average bill depth is about 17 mm, average flipper length is about 200 mm. Average body mass is about 4200 g.

The data look reasonable: there are no significant outliers, minimums and maximums of all variables are quite close to the mean and median characteristics.

Let's look at unique values in categorical variables: species, island, sex and year - it is numeric, but we can treat it as a categorical for this purpose.

Table 2. Unique values of categorical variables

Species	
Adelie	159
Gentoo	123
Chinstrap	76

Island	
Biscoe	169
Dream	132
Torgersen	57

Sex	
Male	174
Female	175

Year	
2007	125
2008	114
2009	119

As we can see, information about island will make prediction model far better: for Torgensen island it always will be Adelie penguins, 2 other islands have information about 2 species each, so we would make 2 datasets: with and without island variable to see how the quality of prediction differs in those two cases.

The initial research was focused on sexual dimorphism. According to the table with the “sex” variable, this variable is balanced in the dataset.

2.3. Correlations analysis

We used Pearson and Phic correlation coefficients[7] to look at the correlations between features.

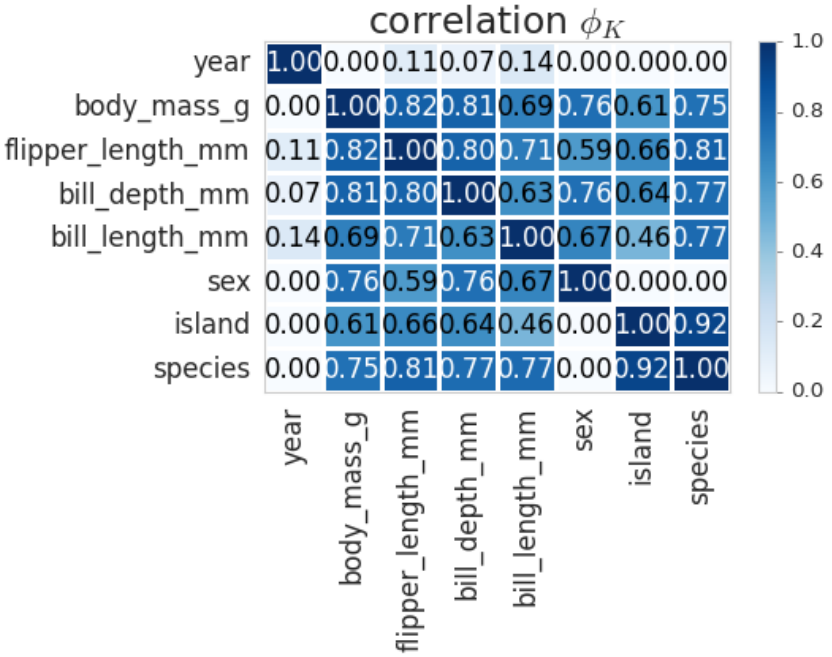


Image 3. Phic correlation matrix.

As we can see, “year” variable does not correlate to any other variable. We can thus remove it from our datasets as it will not affect the results much. Also, we think it should not be in the dataset as we should be able to predict the penguins based on their parameters and possibly where they have been found.

2.4. Duplicates

We have also done a duplicate check, as we did not want to count with possibly duplicated penguins in the dataset. We have learned that the dataset contained 11 duplicates, so we have removed them.

The final shape of the dataset is 347 rows and 13 columns.

3. Data Preparation

We use “species” as a target variable (y). We exclude `year` variable as it doesn’t correlate to the size and species of the penguin.

During data preparation we have excluded rows with more than 3 missing values. For rows with missing only sex and flipper length, we have added data based on the other characteristics of the penguin (average predicted value for the same group).

Our main explanatory variables are:

- bill_length_mm
- bill_depth_mm
- flipper_length_mm
- body_mass_g

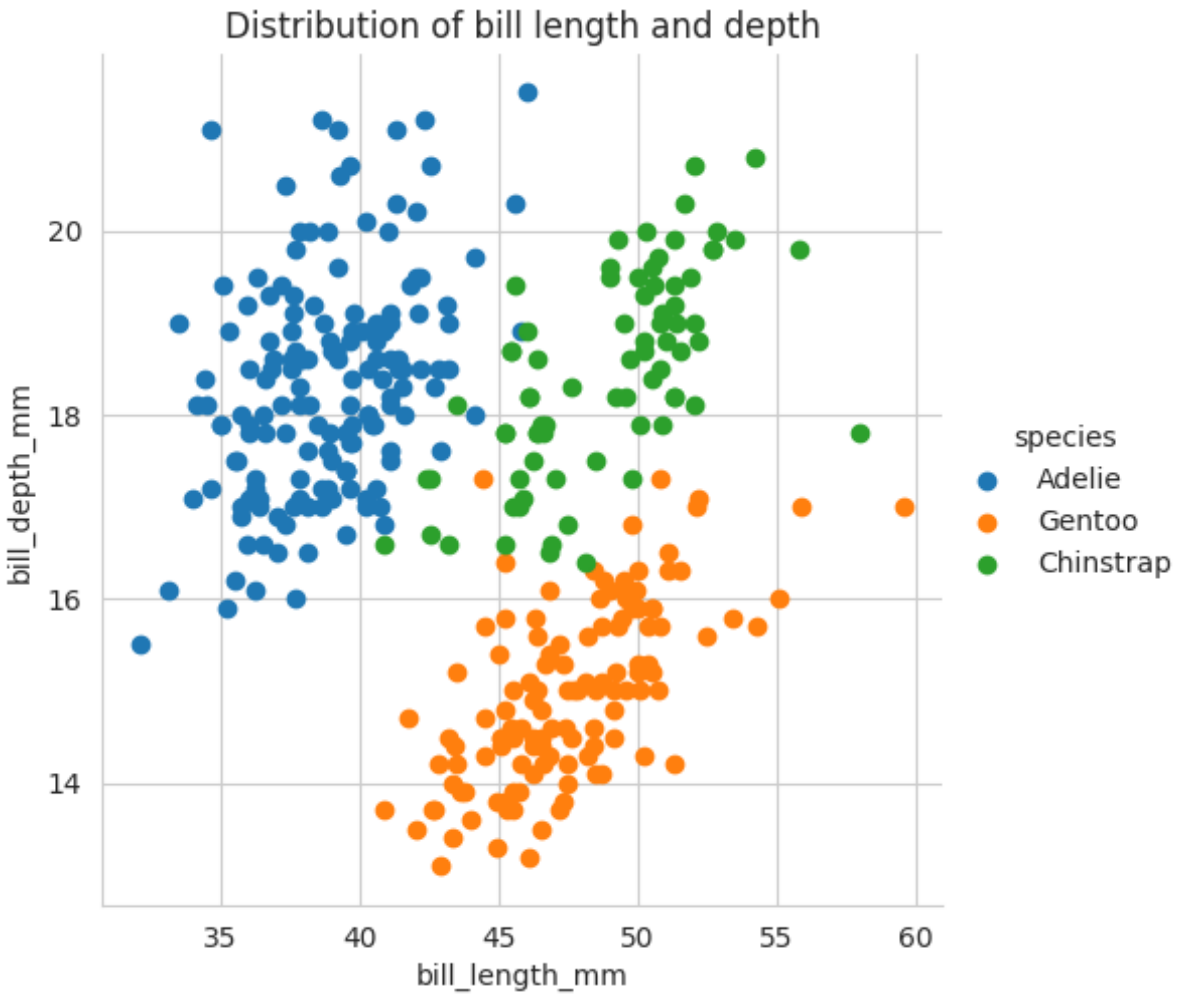
We have decided to make one more dataset, which includes each island variable, and compare results of prediction with and without this variable.

We worked the whole time only with seed 444. We have used it for the train-test split and also for all of the models where it was appropriate.

We split dataset into training values - 80% and testing values - 20%, and saved it as a binary pickle file.

4. Data Visualization

Before the actual models, we have visualized some graphs in order to better understand the data and their characteristics:

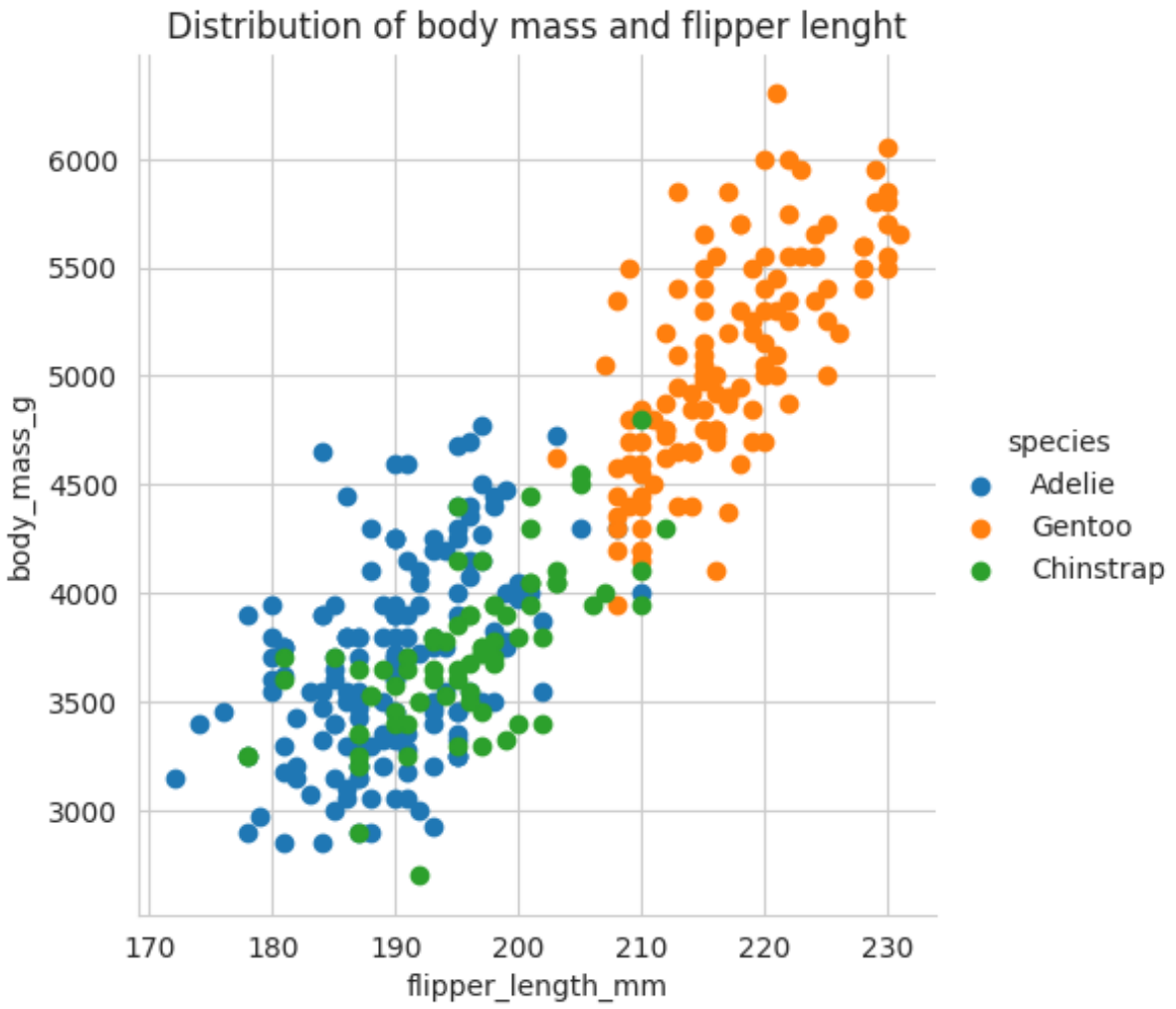


Graph 1. Distribution of bill length and depth

This graph displays the distribution of the length and depth of bill across several species of penguins. The data is divided by the species and each species has its own color. The points on the graph represent individual samples with values for bill length and depth for each penguin.

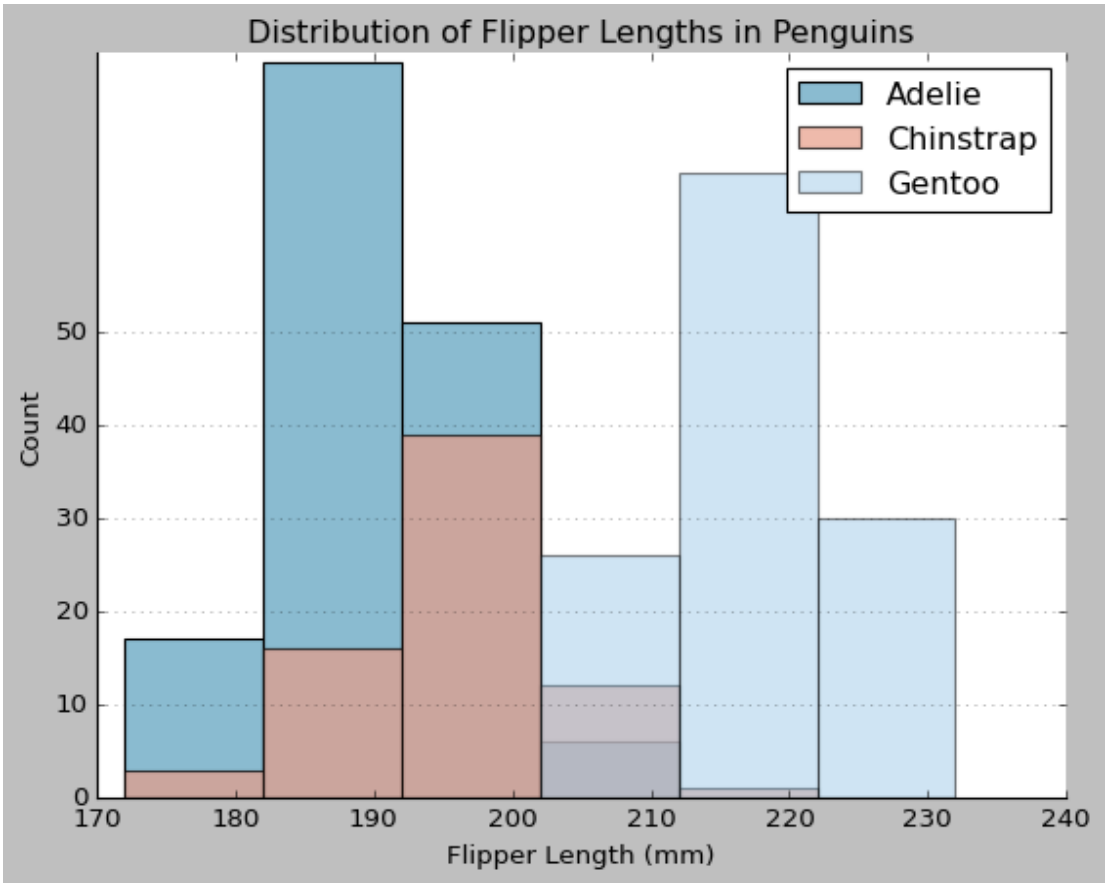
Graph 2. Multiple Regression Plot

The graph shows the relationship between the length and depth of the bill of several species of penguins. For each penguin species, a linear regression line is created, allowing for a comparison of the trend between the length and depth of the beaks among the different species.



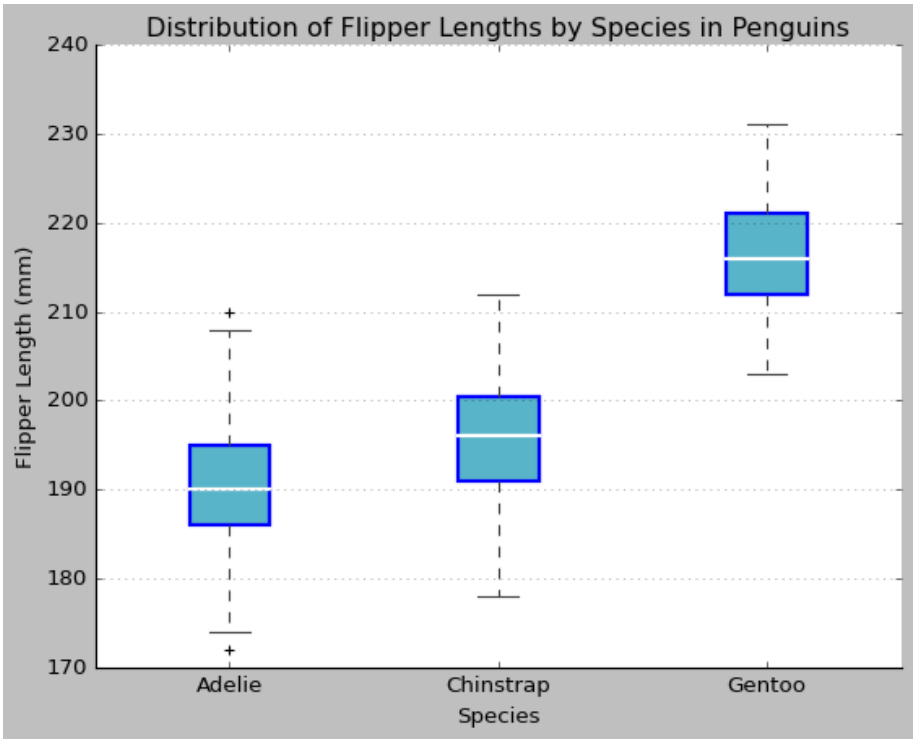
Graph 3. Distribution of body mass and flipper length

This graph shows the relationship between flipper length and body mass among several species of penguins. The data is divided by the species category, and each species has its own color. The dots on the graph represent individual samples with values of flipper length and body mass for each penguin.



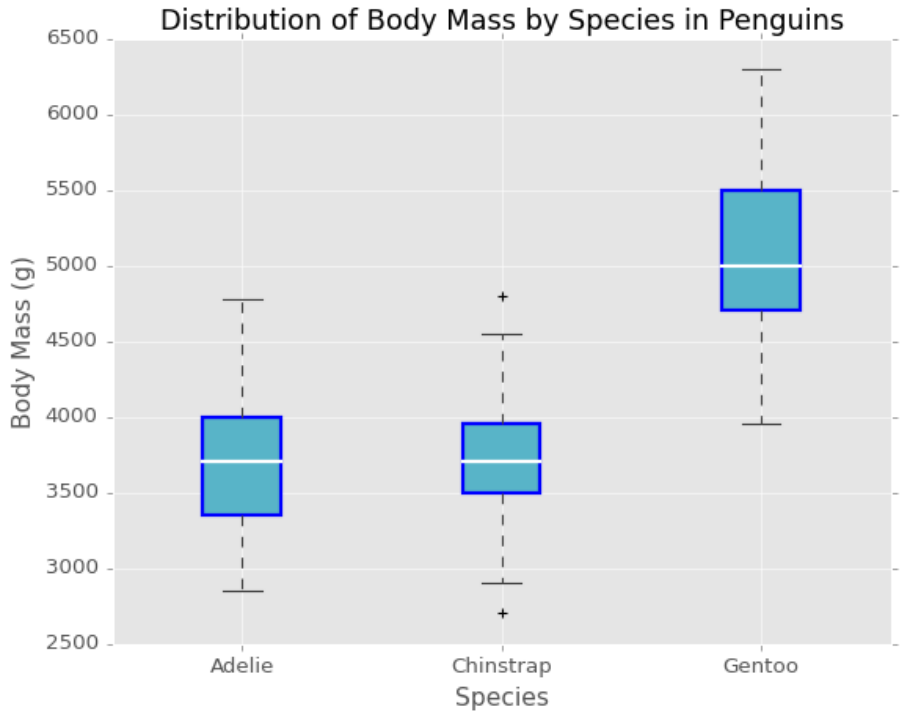
Graph 5. Distribution of Flipper Lengths in Penguins

In this graph, we can see the distribution of penguins based on flipper lengths. For the Adelie penguins, the most typical flipper length is around 182-192mm. For the Gentoo, the most typical flipper length is 212-222mm, and for the Chinstrap, the typical flipper length is 192-202mm.



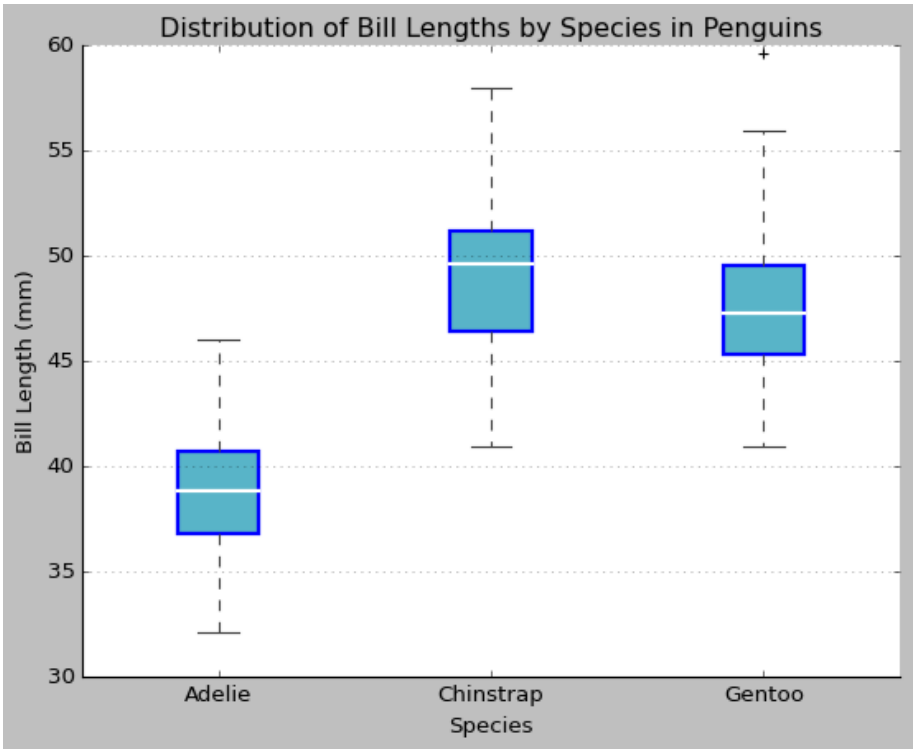
Graph 6. Distribution of Flipper lengths by Species in Penguins

Here we have another graph (boxplot) of Flipper Length, where we can more clearly see that the species with the longest flipper length is Gentoo, followed by Chinstrap, and the species with the shortest flipper length is Adelie.



Graph 7. Distribution of Body Mass by Species in Penguins

In this graph, we can see that when comparing body mass, the heaviest penguins are the Gentoo. The Adelie and Chinstrap species have similar weights, while Adelie having a greater range of body mass variability.



Graph 8. Distribution of Bill lengths by Species in Penguins

In this boxplot we can see the distribution based on bill length. We can observe that the Chinstrap species has the largest bill length, followed by Gentoo, and the Adelie species has the smallest bill length.

5. Modeling

We have chosen classification models, because they can be used for deciding and predicting a single value, in our case it was a species of penguins. At first, we also tried regression models like regression trees and XGBoosting, but in the end, we discarded them as decided that we want just models where the result is obvious - just one penguin, not some number between them.

Our models we chose in the end are:

1) Single classification tree

Single classification tree is one of the easiest classification models. They perform quite well on classification problems, the decisional path is relatively easy to interpret, and the algorithm to build (train) them is fast and simple.
We have created single classification tree models using both datasets with and without the variable “island” on the same seed, while both ended with the same results.

Model limitations and considerations:

- Overfitting: decision trees have a tendency to overfit to the training data. It usually happens when the hyper-parameters are set to consider any details (nodes with less leaves and big maximum depth of the model).
- Bias and variance trade-off: decision trees can suffer from the bias-variance trade-off, where increasing the complexity of the tree can reduce bias but increase variance. This can be addressed by selecting an appropriate tree depth and using techniques like cross-validation to evaluate model performance.
- Complexity and interpretability: while decision trees are relatively easy to interpret compared to other machine learning models, they can still be complex and difficult to understand.
- Sensitivity to class imbalance: decision trees can be biased towards the majority class in imbalanced datasets, which can lead to inaccurate predictions for the minority class.

Ideas to improve the model:

- We have added cross-validation in our model, so our model should prevent bias and variance trade-off. We have also discussed the target class (species) imbalance, but from our point of view it was not that much severe to affect our results.

Explain how did you choose the values for the hyper-parameters of your model:

- We have taken a closer look on how the model behaves and chosen hyper-parameters in order to prevent both overfitting and bias and variance trade-off. We used a hyper grid with these final parameters, which we consider as optimal:
    - max_depth: [2,3,4,5] - Using these parameters, the model should not be high in complexity and easily interpretable. Also, we reduce the risk of overfitting

and bias. Our dataset has not quite that much data, so we think this grid contains the optimal numbers of maximal depth.
    - min_samples_leaf: [0.02, 0.05, 0.1] - by defining minimal samples of the trees, we avoid getting overfitted trees. The minimal number (0.02) means that we can have only 6 penguins in one leaf.

2) Logistic regression

We created logistic regression models for both datasets (with and without variable Island). The model trained on a dataset without variable island have slightly better results.

Model limitations and considerations:

- Logistic regression is good for multiclass classification when the classes are well-separated and there is a clear linear boundary between them. However, it usually does not work well when the classes overlap a lot or when the relationship between the input features and the target classes is complex. So, it really depends on the specific dataset.

Limitations:

- Linearity between input features and target class
- Overfitting (as for every model), however, we used cross-validation, when creating the model so the risk of overfitting is lowered
- If the target classes in the dataset are imbalanced, the model is usually biased towards the most populated class and may have low recall for the minority class

Considerations:

- Before we created the model we considered following:
    - How to choose optimal hyper-parameters
    - Use cross validation to reduce the chance of overfitting

Ideas to improve the model:

- We used an approach using the softmax function, however, there is also a one-vs-all (OvA) approach, where binary classifiers are trained for each class to distinguish it from all other classes. Comparison of both logistic regression models with various hyper-parameters may be useful.

Explain how did you choose the values for the hyper-parameters of your model:

- To select the optimal hyperparameters we used GridSeachCV class from scikit-learn, where we created a dictionary of parameters (penalty and “C” (inverse of the regularization strength)). The GridSearchCV then tries every possible combination of the parameters to find the best hyperparameters setup.

3) Bagged random forest

Model limitations and considerations:

- Interpretability: Bagged random forest models are often considered less interpretable compared to other models, as the ensemble of decision trees can result in complex and non-linear relationships between features and predictions. This can make it challenging to explain and interpret the model's predictions, especially in scenarios where interpretability is crucial.
- Computational Complexity: Bagged random forest models can be computationally expensive, especially if the number of trees in the ensemble or the size of the dataset is large. This can impact the training and prediction speed of the model, making it less suitable for real-time or resource-constrained environments.

Ideas to improve the model:

- Regularization Techniques: Incorporating regularization techniques, such as feature bagging or tree pruning, can help reduce overfitting in the model and improve its generalization performance.

Explain how did you choose the values for the hyper-parameters of your model:

- Grid Search: Trying different values of hyper-parameters in a predefined range and selecting the one that performs best on a validation set (hyper grid and best parameters can be found in the evaluation section of this report).

6. Evaluation

We have created a table with evaluation for all described classification models below. The best performing model on our dataset based on Accuracy, Precision, Recall, F1 Score is random forest.

Table 3. Metrics of model performance, %

	Accuracy	Precision	Recall	F1 score
Single tree (without island)	97.14	97.20	97.14	97.12
Single tree (with island)	97.14	97.20	97.14	97.12
Logistic Regression (without island)	95.71	95.68	95.71	95.66
Logistic Regression (with island)	97.14	97.14	97.14	97.14
Random Forest (without island)	98.57	98.67	98.57	98.58
Random Forest (with island)	100.00	100.00	100.00	100.00

Bagged random forest with variable island

Based on performance metrics of all our models, we have chosen the best one for final use and evaluation. Best performing model is bagged random forest, that is using dataset dataset extended by the island variable.

Bagged Random Forest is an ensemble learning technique that combines multiple individual decision tree models, trained on different subsets of the training data with replacement (bootstrap samples). These individual trees are then combined by averaging or voting to create a more robust and accurate prediction model that can reduce overfitting and improve overall performance.

Used parameters

At first we defined the hyper-parameter grid for tuning. Tuning algorithm will then choose the optimal value for each parameter.
"n_estimators": [5, 10, 20], specifies the number of decision
"max_depth": [None, 10,20,30], maximum depth of each decision tree
"min_samples_split": [6, 8, 10], minimum number of samples required to split an internal
"min_samples_leaf": [3, 4, 5], minimum number of samples required to be at a leaf
"bootstrap": [True, False], determines whether bootstrap samples should be used
"random_state": [444] sets the random seed for reproducibility

In this case, optimal parameters are as follows:
'n_estimators': 20,
'max_depth': None,
'min_samples_split': 6,
'min_samples_leaf': 3,
{'bootstrap': True,
'random_state': 444}

Example decision tree (image on the next page):

In this example, we can describe one of the decision trees from the victorious random forest. At first, this model decides whether the penguin lives on the island of Biscoe. If we take the right path, we can see that from this particular training dataset, 48 Adelie and 90 Gentoo penguins were from the island of Biscoe. Then, these penguins are separated by their bill depth value - smaller bills than 16.4 depth are only Gentoo, while equal or larger bill depth contains 48 Adelie and 1 Gentoo. It is then finally sorted by the body mass. Using this logic, we can similarly classify the penguins on the left side of the tree.

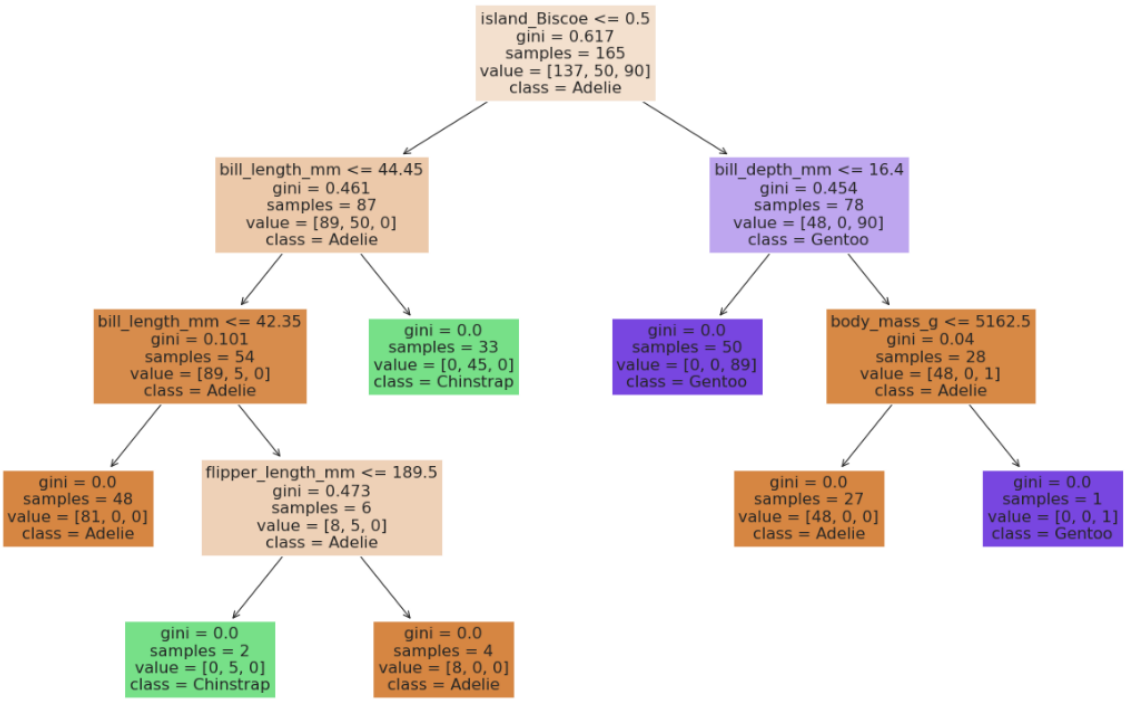


Image 4. Example decision tree

Confusion Matrix

In the confusion matrix below, we can see that our best model (Bagged Random Forest with the variable “island”) correctly classified every penguin. We can observe that all 70 examples are on the diagonal, indicating that the model achieved 100% of all chosen metrics.

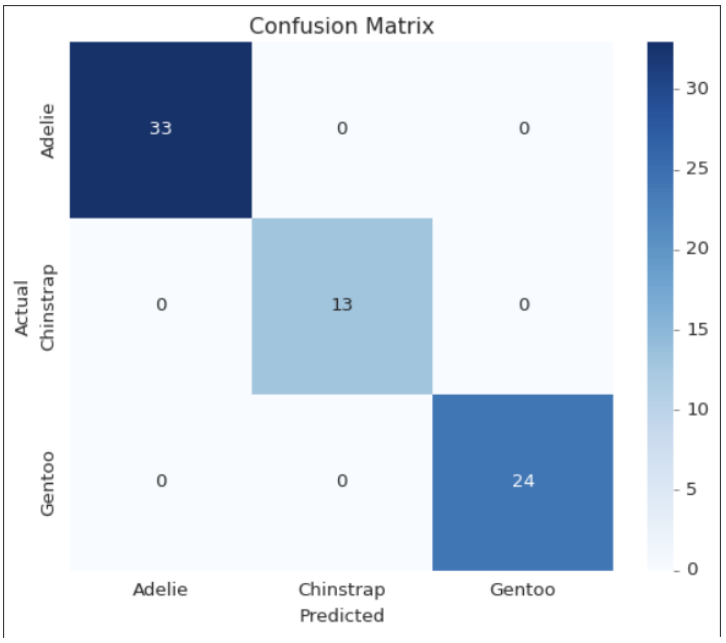


Image 5. Confusion Matrix

Feature importance

Feature importance in random forest is a technique used in machine learning to determine the importance of features (or variables) in a random forest model. It is calculated by measuring the decrease in model performance when a particular feature is randomly shuffled or removed. The higher the decrease in performance, the more important the feature is considered to be, as it has a larger impact on the model's predictive accuracy.

For example, "bill_length_mm" and "bill_depth_mm" are identified as the two most important variables, with an importance of over 20%, indicating a significant impact on the model. In contrast, the variable "is_female" is considered the least important."

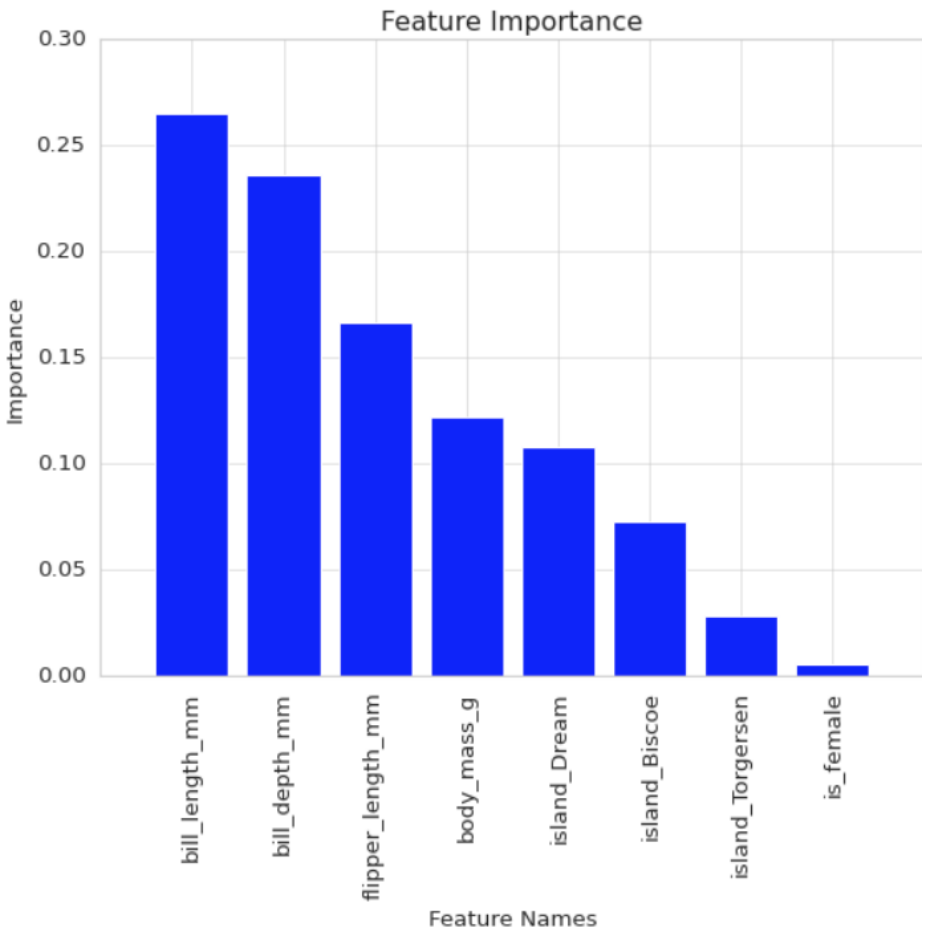


Image 6. Feature importance

7. Development Environment Characteristics

Packages utilized and its respective version and python version: Requirements.txt

References

1. Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081
2. Adelie Penguin Picture https://commons.wikimedia.org/wiki/File:Adeliepinguin-01.jpg
3. Chinstrap Penguin Picture https://commons.wikimedia.org/wiki/File:A_chinstrap_penguin_(Pygoscelis_antarcticus)_on_Deception_Island_in_Antarctica.jpg
4. Gentoo Penguin Picture https://en.wikipedia.org/wiki/Gentoo_penguin#/media/File:Brown_Bluff-2016-Tabarin_Peninsula%E2%80%93Gentoo_penguin_(Pygoscelis_papua)_03.jpg
5. Vidiyala R. (2020) Performance Metrics for classification Machine Learning. https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007
6. Silipo R. (2019) From a Single Decision Tree to a Random Forest. https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147
7. Lewinson E. (2021) Phik (ϕk) — get familiar with the latest correlation coefficient https://towardsdatascience.com/phik-k-get-familiar-with-the-latest-correlation-coefficient-9ba0032b37e7