Prague University of Economics and Business

Faculty of Informatics and Statistics

CLASSIFICATION OF PENGUINS

4IT439 - DATA X

Authors: Michal Svoboda, Eva Schwarzová, Ondřej Vévoda, Martina Kunešová,
Veronika Struhárová, Veronika Nováková

Prague, April 2023

Contents

1 Problem definition I

2 Data understanding II

2.1 Missing values and descriptive Statistics II

2.2 Correlation IV

3 Data visualisation VI

4 Data preparation...X

5 Models XI

5.1 Decision tree Classifier XI

5.2 KNeighbors Classifier.. XI

5.3 Random Forrest Classifier XI

5.4 XGBoost Classifier XII

5.5 Summary XII

Conclusions.. XIII

1 Problem definition

The goal of this classification task is to train and validate a model that is able to recognise species of penguins. The model has to be able to predict this fact based on features contained within penguins.csv dataset. Among the provided variables contained within the dataset are: species, island, bill size, etc. After the initial data preprocessing and EDA (Exploratory Data Analysis) is performed to determine the best features to include into a model. After that, various models are trained and evaluated to determine the best one. Models are tested and evaluated by various methods and metrics such as accuracy, precision, recall or f1 score.

2 Data understanding

The data set contains information on penguins that were observed on three different islands in the Southern Ocean: Biscoe, Dream, and Torgersen. The penguins belong to three different species: Adelie, Gentoo, and Chinstrap.

For each penguin in the data set, the following variables are recorded:

- Island: The island on which the penguin was observed. This variable is categorical with three possible values: "Biscoe", "Dream", or "Torgersen".
- Species: The species of the penguin. This variable is categorical with three possible values: "Adelie", "Gentoo", or "Chinstrap".
- Bill length: The length of the penguin's bill in millimeters. This variable is numeric.
- Bill depth: The depth of the penguin's bill in millimeters. This variable is numeric.
- Flipper length: The length of the penguin's flipper in millimeters. This variable is numeric.
- Mass: The mass of the penguin in grams. This variable is numeric.
- Sex: The sex of the penguin. This variable is categorical with two possible values: "male" or "female".
- Year

The data set contains a total of 363 observations. The goal of the classification task is to predict the species of a penguin based on its island, bill length, bill depth, flipper length, mass, and sex.

2.1 Missing values and descriptive Statistics

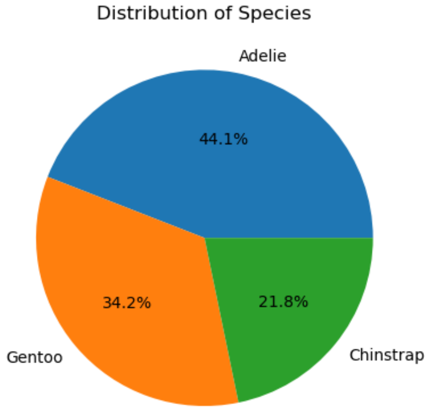
Missing values can be a common problem in data sets, and the penguin classification data set is no exception. In this data set, missing values occur for the following variables: sex, bill length, bill depth, flipper length, and body mass.

Species	Island	Bill lenghth	Bill depth	Flipper length	Body mass	Sex	Year
0	0	5	5	6	5	14	0

The following table describes all variables in more detail:

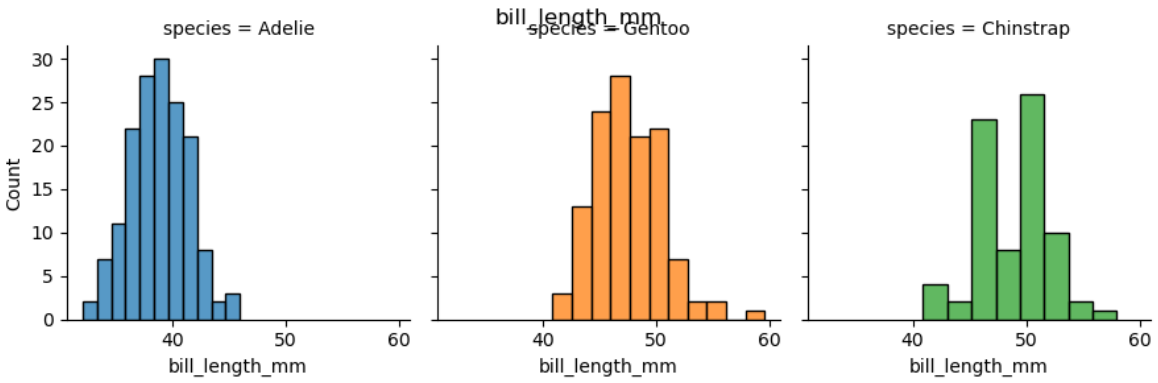
	species	Island	Sex	Bill length	Bill depth	Flipper length	Body mass	Year
count	363	363	349	358	358	357	358	363
missing_val	0	0	14	5	5	6	5	0
unique	3	3	2	-	-	-	-	-
top	Adelie	Biscoe	female	-	-	-	-	-
freq	160	170	175	-	-	-	-	-
mean	-	-	-	43,93	17,21	200,45	4173,74	2007,99
std	-	-	-	5,44	1,95	14,00	796,40	0,83
min	-	-	-	32,10	13,10	172,00	2700,00	2007,00
0,25	-	-	-	39,35	15,70	190,00	3550,00	2007,00
0,50	-	-	-	44,45	17,50	197,00	3950,00	2008,00
0,75	-	-	-	48,50	18,70	213,00	4743,75	2009,00
max	-	-	-	59,6	21,5	231	6300	2009

We also show here a classified variable, species, where the percentage of each species may influence the choice of the appropriate model and method.

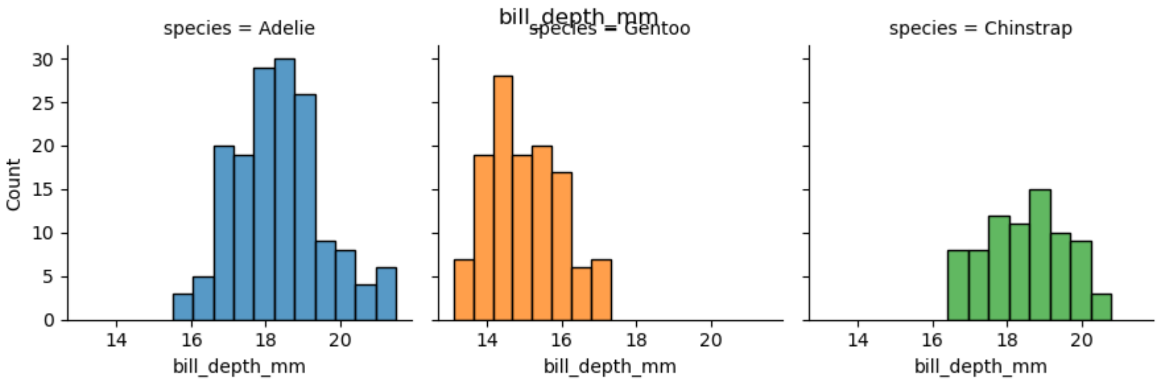


We decided to visualize some numeric variables by penguin category to get a first idea of the overall categorization. We can see some significant differences between the species in bill size and body mass.

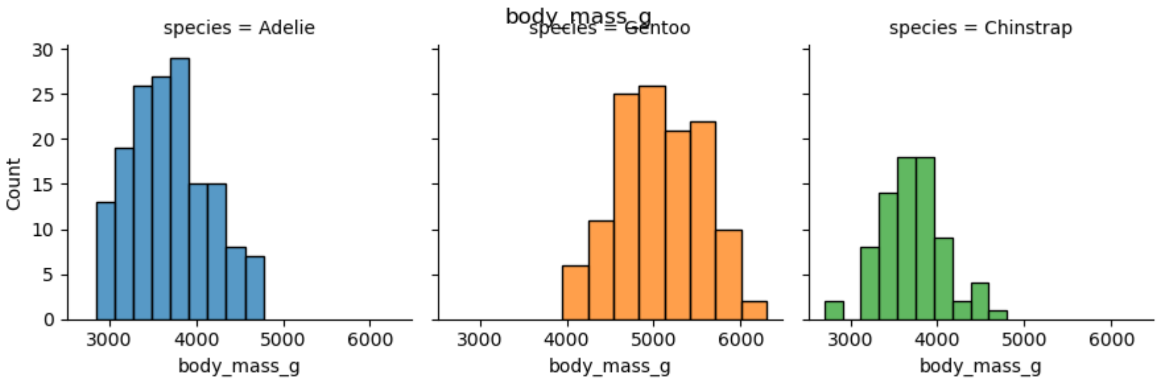
a, bill length



b, bill depth

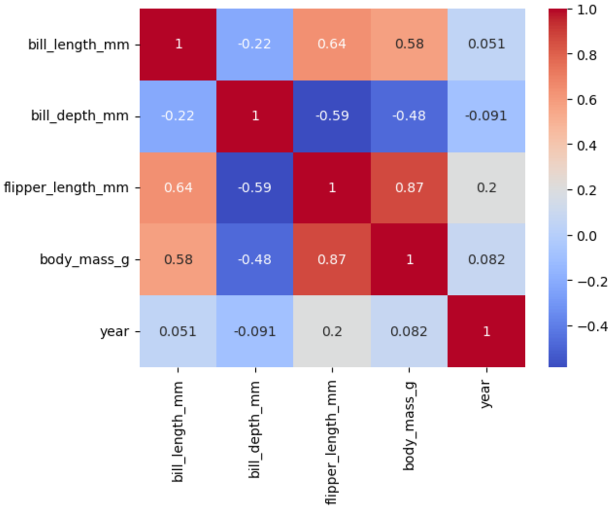


c, body mass



2.2 Correlation

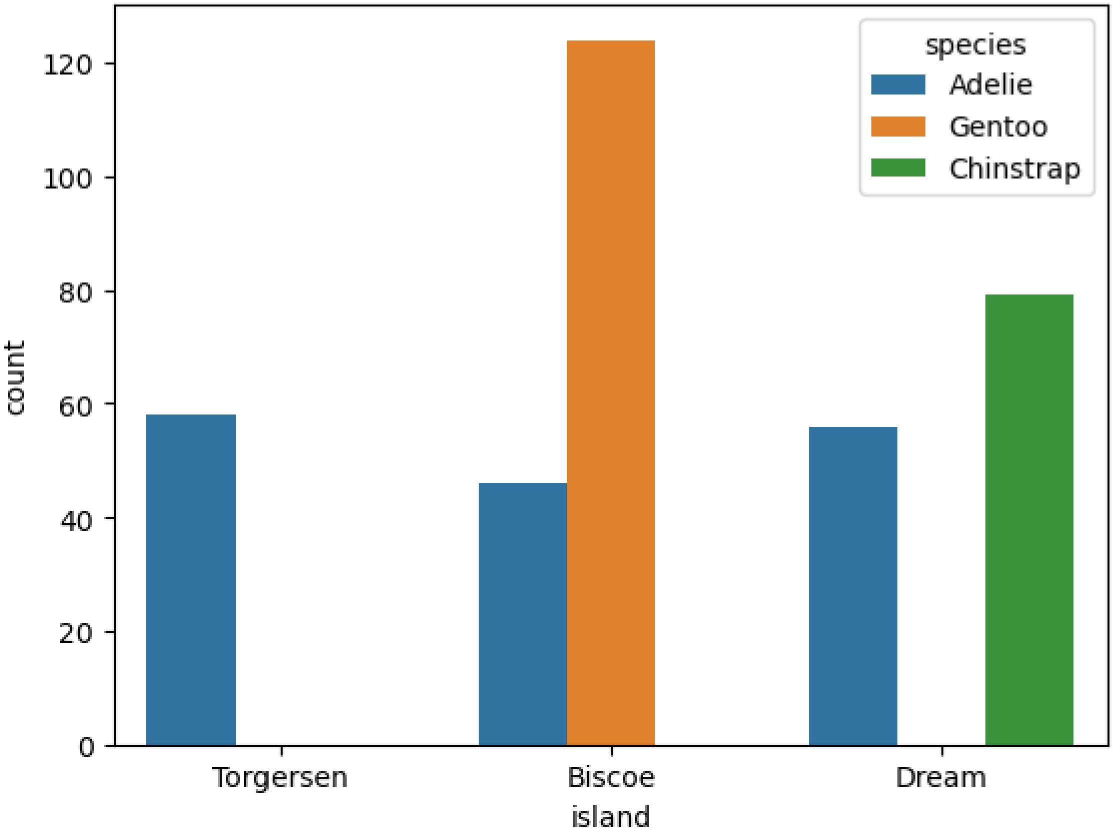
Correlation is a statistical measure that describes the relationship between two variables. In the penguin classification data set, there is a significant correlation between flipper length and body mass. In case the model does not work, we could consider removing one of the mentioned variables.



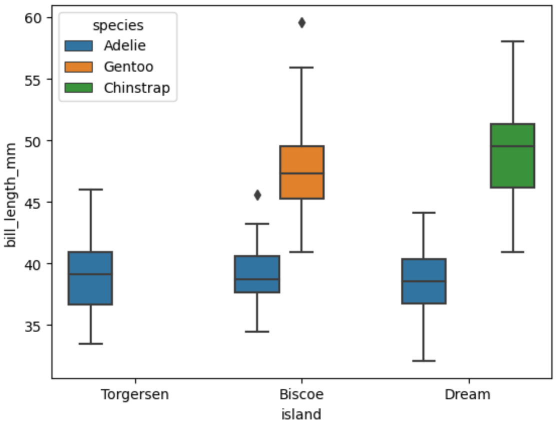
3 Data visualisation

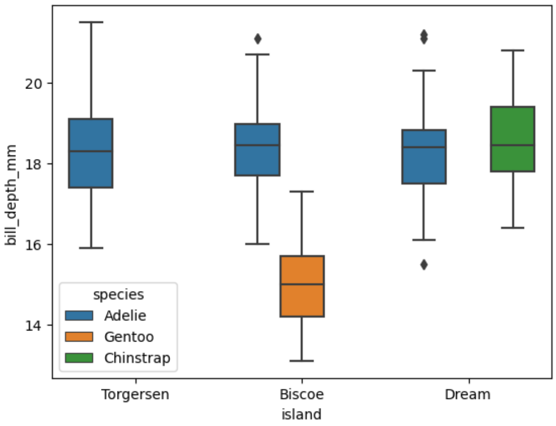
In the previous chapter, it was already indicated that differences between penguin species can be observed in some variables. This section will take a closer look at what can be conclude about the data based on the visualization.

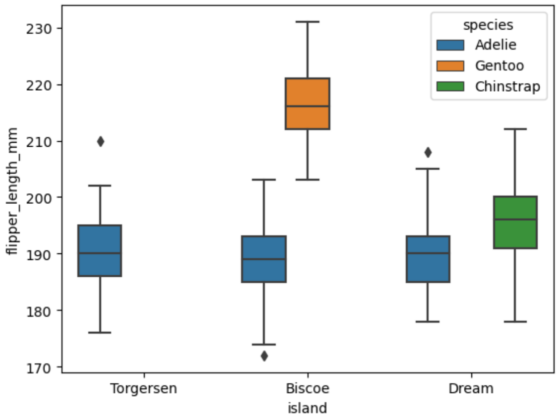
The graph below shows the frequency of species on individual islands. It can be seen that the Adelie species occurs on all three islands. At the same time, it occurs similarly frequently on these islands. The Gentoo species was observed only on Biscoe Island, the Chinstrap species was observed only on Dream Island. If we come across a penguin on Torgersen Island, it can be assumed that it will be an Adelie species. For penguins on other islands, we will need more information to identify the species.

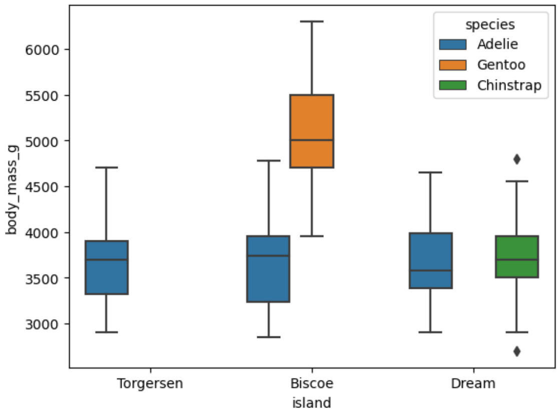


It can be expected that it will be appropriate to use body measures for further analysis. In the boxplots below, you can see that the dimensions take on different values for the species. It can also be noticed that the dimensions of Adelia are the same on different islands.



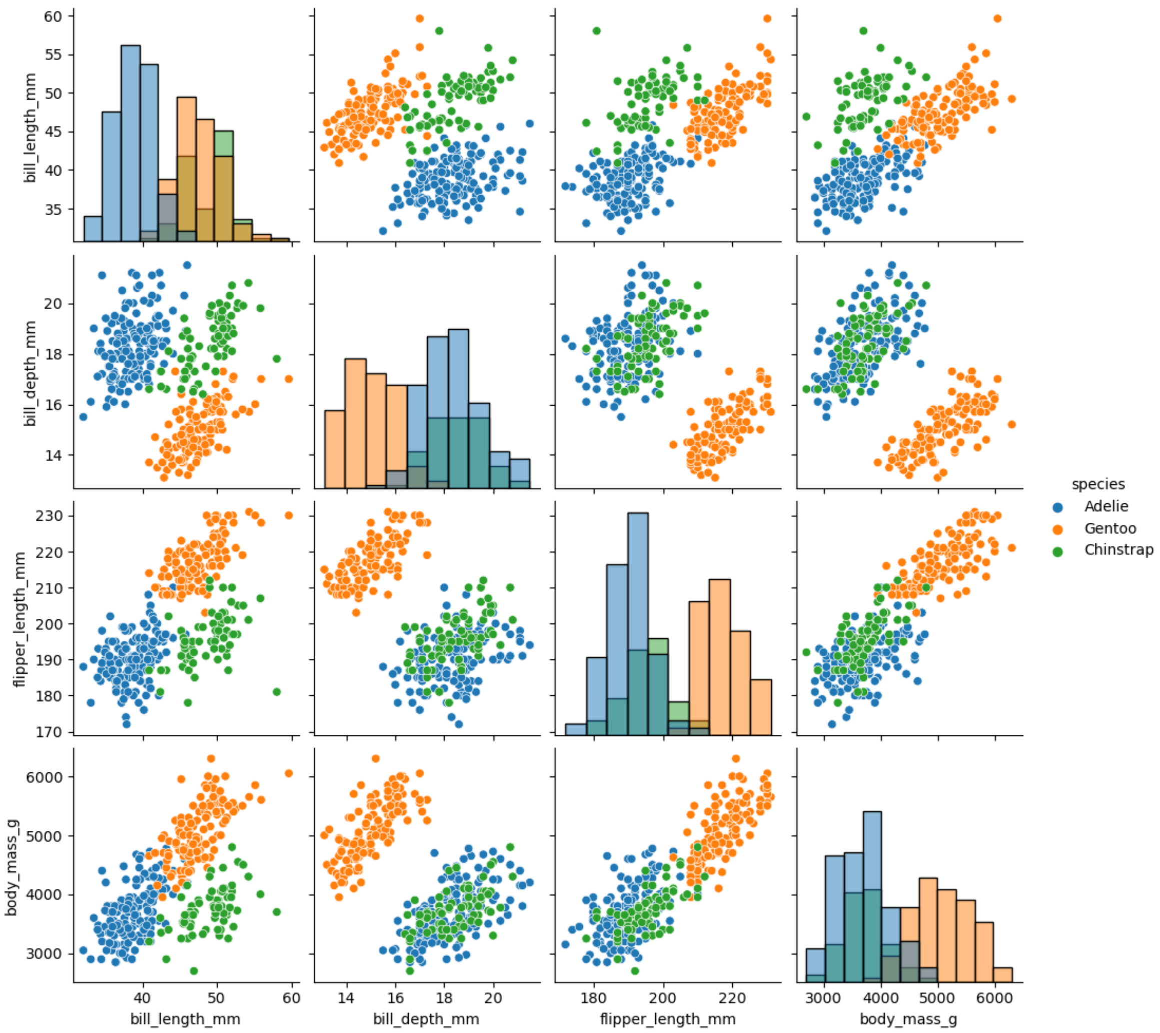




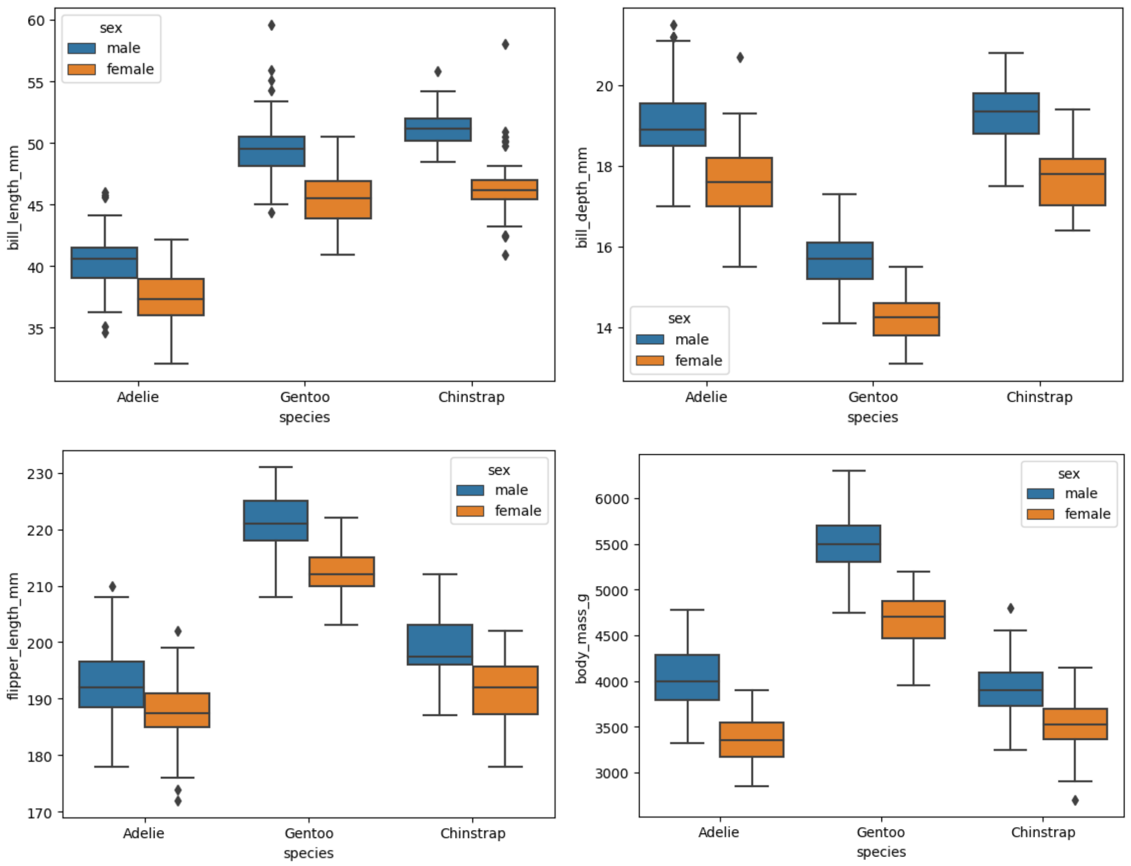


Now we will focus on the relationships of individual variables with body measures. It is clear from the scatterplots that there are positive linear relationships between the variables. In some cases, clusters corresponding to species can be identified in the graphs. These variables can be expected to be very helpful in distinguishing species.

From the histograms and scatterplots, it can be seen that the Gentoo will be the largest species of penguin, as they are the heaviest and have the longest flipper. The other two species are similar in size. We will look for the differences between Adelie and Chinstrap in the bill measures. Adelies have a shorter bill compared to other species. They have a greater bill depth. Chinstraps have a longer bill with greater depth. Gentoo have a longer bill, but compared to other species, their bill have smaller depth.



For all variables with body measures, it can be observed that females reach lower values. The differences between each species are very similar for both sexes. Based on these judgments, we do not expect a large contribution of variable sex in recognition of species.


bill_length_mm
60
55
50
45
40
35
sex
male
female
Adelie
Gentoo
Chinstrap
species
bill_depth_mm
20
18
16
14
sex
male
female
Adelie
Gentoo
Chinstrap
species
flipper_length_mm
230
220
210
200
190
180
170
sex
male
female
Adelie
Gentoo
Chinstrap
species
body_mass_g
6000
5500
5000
4500
4000
3500
3000
sex
male
female
Adelie
Gentoo
Chinstrap
species

4 Data preparation

Based on data exploration there were number of issues that needed to be addressed. In the initial steps of data preparation year column was dropped as it is not relevant to the classification problem. Remaining categorical variables were converted to category type.

Then we continued with data splits, a standard 0.7, 0.15, 0.15 split was used to divide the dataset into training, validation, and testing subsets. A splitter function was defined to generate balanced data splits. Stratified sampling was used to ensure that each subset had an equal distribution of the target variable, species. Since the penguin dataset was imbalanced, with some species having more samples than others, oversampling was performed on the training set. Random oversampling was used to increase the number of samples for each species to ensure that the model could learn from a more balanced dataset. Class weights were also calculated to ensure the correctness of the methos chosen.

After the splits were done, missing values and encoding was taken care of. There is a one difference between train and val(test) sets processing and it overcomes data leak. While in train set we compute the mean and modus (for numerical and categorical variables), in validation and train set we impute those calculated values from train set. Sex column was encoded into a binary value and the island column was one-hot encoded to create separate binary columns. For case of model KNeighborsClassifier we also used MinMaxScaler for numerical columns.

There were 2 saving points. One for intermediate data, right after train-test split. And another one for processed data after imputing and encoding. Both of them are to be found in pickle format.

5 Models

In the modelling process we focused on characteristics of dataset as its nature in tabular data, small dataset with few features. We focused on basic models as we believed that the dataset will behave comparably good with lower computational expenses than using more complex models. Even though, we added XGBoost for comparison and better overall view.

When introducing hyperparameters, the ones not mentioned are set up by default.

5.1 Decision tree Classifier

As a first model we used Decision Tree Classifier with grid search. Best hyper parameters were max_depth = 5, min_samples_leaf = 2, min_samples_split = 2.

The accuracy for this model was 0.945.

Precision with the ‘macro’ average setting was used, the result was 0.955.

Recall with the ‘macro’ average setting was used, the result was 0.945.

F1 score for this model was 0.956.

5.2 KNeighbors Classifier

As mention before, in comparison with tree models MinMaxScaler and simple pipeline was used.

The accuracy for this model was 0.982.

Precision with the ‘macro’ average setting was used, the result was 0.974.

Recall with the ‘macro’ average setting was used, the result was 0.982.

F1 score for this model was 0.980.

5.3 Random Forrest Classifier

Comparable steps as with Decision tree were taken. Using grid search we chose the best hyperparameters which were max_depth = 5, min_samples_leaf = 2,
min_samples_split = 2, n_estimators = 100.

The accuracy for this model was 0.982.

Precision with the ‘macro’ average setting was used, the result was 0.974.

Recall with the ‘macro’ average setting was used, the result was 0.982.

F1 score for this model was 0.980.

5.4 XGBoost Classifier

Lastly we tried XGBC, again with grid search and best hyperparameters learning _rate = 0.2, max_depth = 5, n_estimators = 50, tree_metod = auto.

The accuracy for this model was 1.

Precision with the ‘macro’ average setting was used, the result was 0.974.

Recall with the ‘macro’ average setting was used, the result was 1.

F1 score for this model was 0.980.

5.5 Summary

Method	Accuracy	Precision	Recall	F1
XBoost	1	0.974	1	0.98
KNeigbours	0.982	0.974	0.982	0.98
Random Forest	0.982	0.94	0.982	0.98
Decision Tree	0.945	0.955	0.945	0.956

Conclusions

As we can see from the validation of the models, every model has a very high accuracy.

Based on the evaluation metrics provided for the different methods, it seems that XGBoost and KNeigbours are the best methods for classification of penguin species. Both methods have high accuracy, precision, recall, and F1 scores, which indicates that they are able to correctly classify the different penguin species with a high degree of confidence.

Random Forest also has high scores across all the metrics except precision, where it has a slightly lower score. Decision Tree, on the other hand, has lower scores across all the metrics compared to the other methods.

For further usage we would recommend either the KNeighborsClassifier or Random Forrest Classifier which both have slightly higher accuracy than a simple decision tree and have lower computation expenses than the XGBoost model.