Classification of Penguin Species

A Data Analysis Project using Python to Compare Four Classification Models

Jessie Lenarz

May 1, 2022

Introduction

From 2007 – 2009, Dr. Kristen Gorman collected data for three penguin species on three islands of the Palmer Archipelago, Antarctica. Her work was part of the Palmer Station Long Term Ecological Research Program, a part of the US Long Term Ecological Research Network. This data set is gaining popularity as an alternative to the traditional Iris data set in academic settings. In this project, we will analyze four different machine learning algorithms (k nearest neighbor, decision trees, support vector machine, and logistic regression) to determine which is best suited to classify penguins into species.

Methodology

Data Sources and Collection

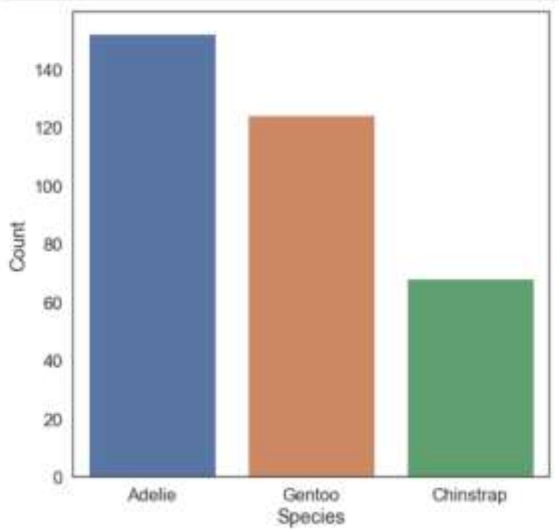
The data were collected by Dr. Kristen Gorman. The data set is gaining popularity in classrooms, available from multiple online sources, and are available for use by CC0 license ("No Rights Reserved") in accordance with the Palmer Station Data Policy. I exported the data as a CSV file from GitHub user Slopp. The data set contains information on three hundred forty-four subjects including the island (location), year collected, length of bill, depth of the bill, flipper length, body mass, sex, and species of the subject.

Data Cleaning

Eleven of the three hundred forty-four records were missing information. Nine of those were missing only the sex of the penguin; the other two were missing the four quantitative measurements and the sex of the penguin. The two records missing the quantitative data were dropped. I also decided to drop the other nine records with missing sex information, resulting in a data set with three hundred thirty-three complete entries.

Exploratory Data Analysis with Data Visualization

I first looked at frequency for the categorical variables of species, island, sex, and year. The bar chart and frequency table for species is in Figure 1. The bar charts and frequency tables for island, sex, and year are available in Appendix A (Figures A.1, A.2, and A.3).



Species	**Count**
Adelie	152
Gentoo	124
Chinstrap	68

Figure 1: Bar chart and frequency table for species

A look at the variables island, species, and year shows a remarkable feature (Figure 2). In each year, the researchers collected data from Adelie penguins on all three islands, Gentoo penguins on Biscoe island only, and Chinstrap penguins on Dream island only. Another way to frame this is the information from Torgersen island is from Adelie penguins only, Biscoe island is from Gentoo and Adelie penguins, and Dream island is from Chinstrap and Adelie penguins. One logical question to ask: are Gentoo penguins restricted to Biscoe island and Chinstrap penguins restricted to Dream island? That is, are there Gentoo penguins on Torgersen and Dream islands and the researchers chose not to collect data from them?

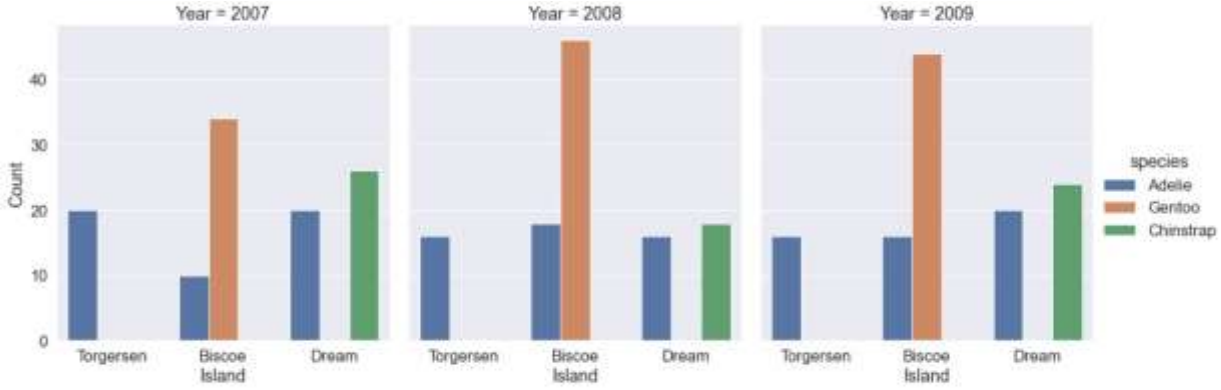


Figure 2: Frequency of the three species on each island in the years 2007-2009

Moving on to the quantitative variables, I looked at the summary statistics for bill length, bill depth, flipper length, and body mass.

	Bill Length (mm)	Bill Depth (mm)	Flipper Length (mm)	Body Mass (g)
Count	342	342	342	342
Mean	43.922	17.151	200.915	4201.754
Standard Deviation	5.460	1.975	14.062	801.955
Minimum	32.1	13.1	172	2700
25th percentile	39.225	15.6	190	3550
Median	44.45	17.3	197	4050
75th percentile	48.5	18.7	213	4750
Maximum	59.6	21.5	231	6300

I used the seaborn library in Python to create histograms for each of the four quantitative features (Figure 3).

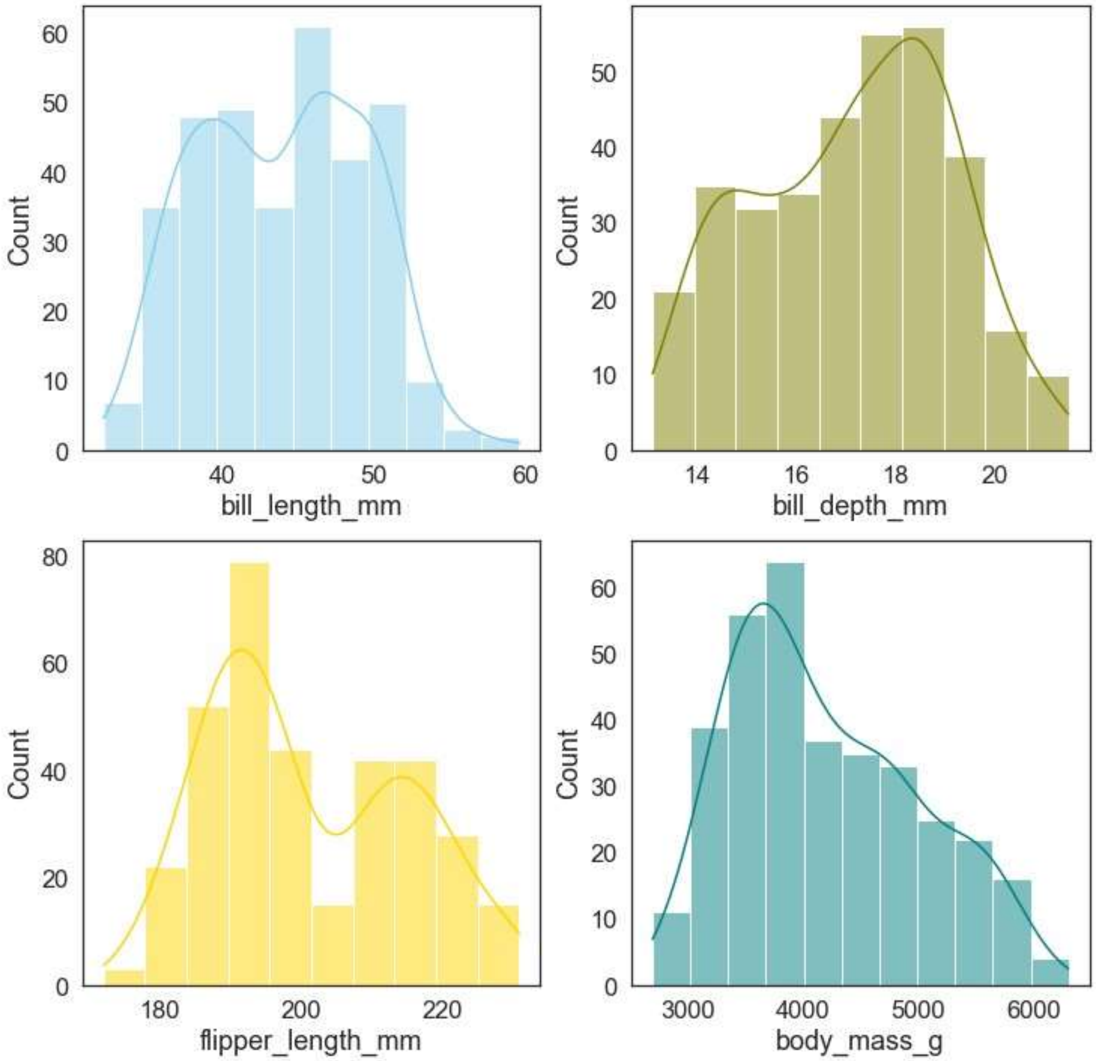


Figure 3: Histograms for the four quantitative features in the dataset

Body mass appears skewed to the right; Bill length, bill depth, and flipper length appear roughly bimodal. None of the plots show the presence of outliers. I dug deeper and looked at the summary statistics grouped by species. Figure 4 shows side-by-side boxplots comparing the four features based on the species of penguin. The boxplots make clear that for all four features, one species of penguin has noticeably different values than the other two. In the case of bill depth, flipper length, and body mass the Gentoo penguins appear different. For the bill length, the Adelie penguins appear different. These observations may explain the bimodal distributions we see in Figure 3. Also note that when grouped by species, we do see some outliers.

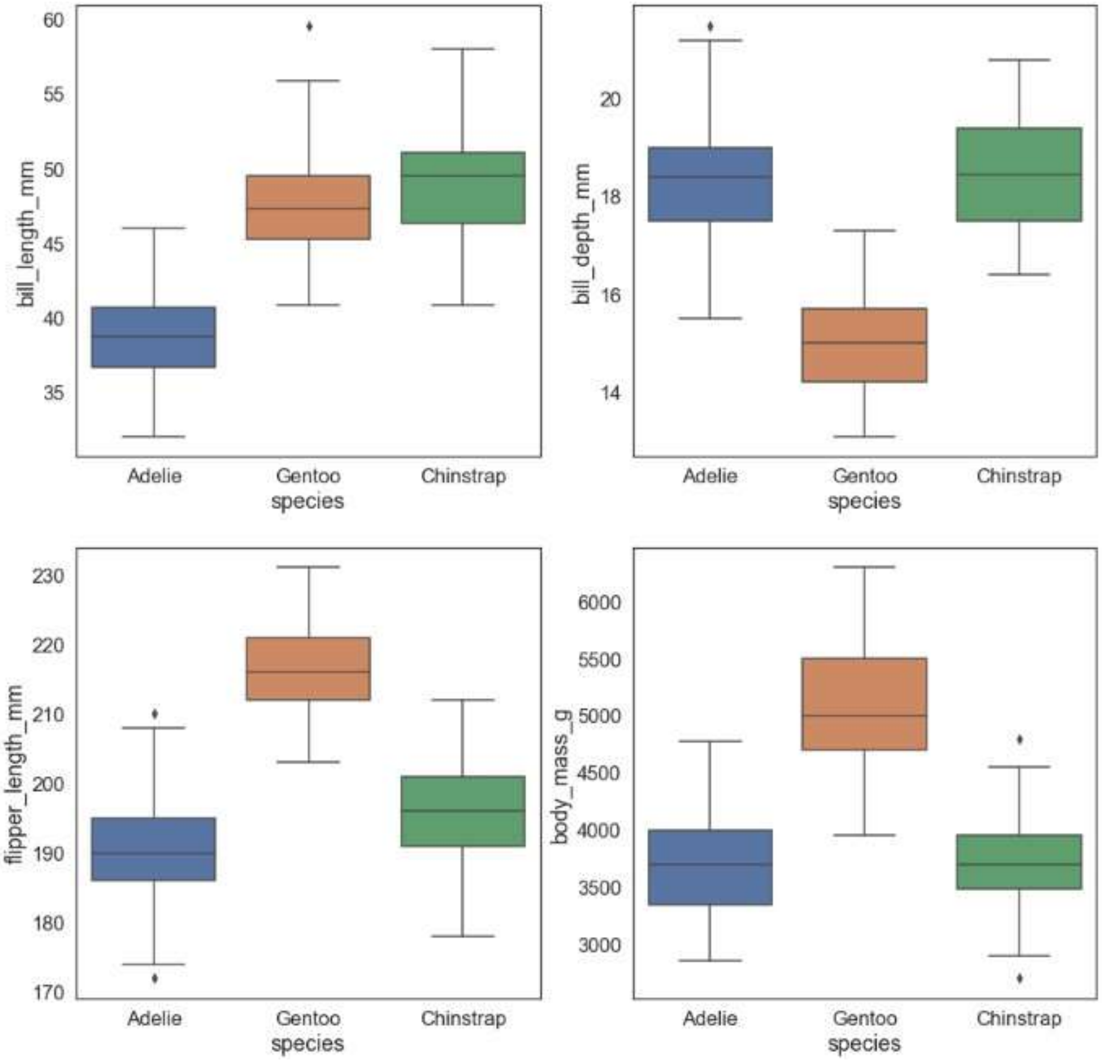


Figure 4: Side-by-side boxplots for each feature grouped by species

I also wanted to see if sex had an impact within the species, so I looked at boxplots detailing the four features grouped by both species and sex. Figure 5 shows there is a noticeable difference in the four features within a species based on gender. Again, I noticed outliers when the penguins were grouped by both species and gender.

I also looked at side-by-side boxplots detailing the four features grouped by both species and island (see Appendix A.4) and both species and year (see Appendix A.5). The features did not seem to change much within the species based on the island or year.

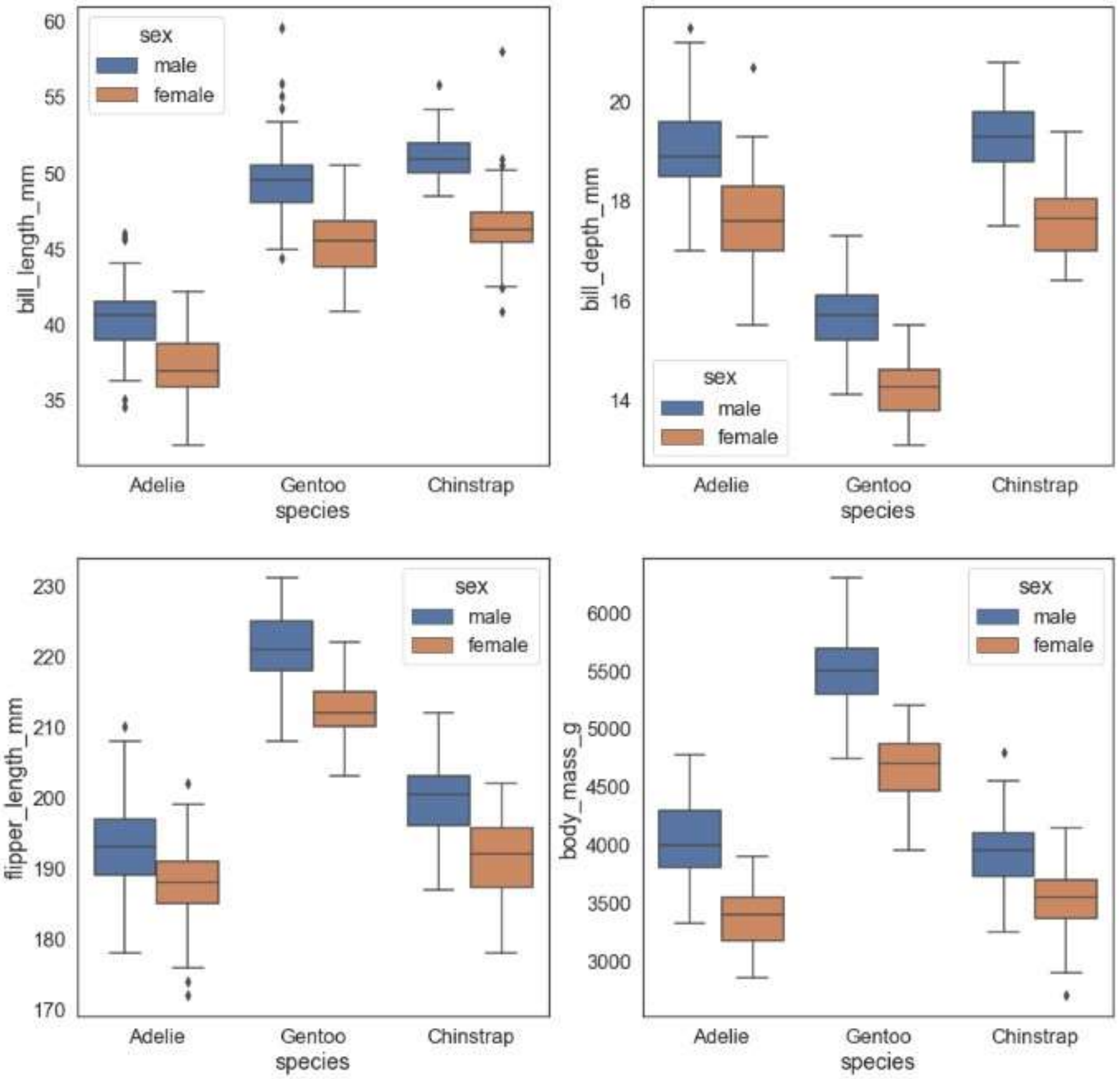


Figure 5: Side-by-side boxplots for each feature grouped by both species and gender within species

I then took a look at charts comparing each pair of features (e.g. bill length vs bill depth, flipper length vs bill length, and so on). The charts, shown in Figure 6, identify each species using different colors.

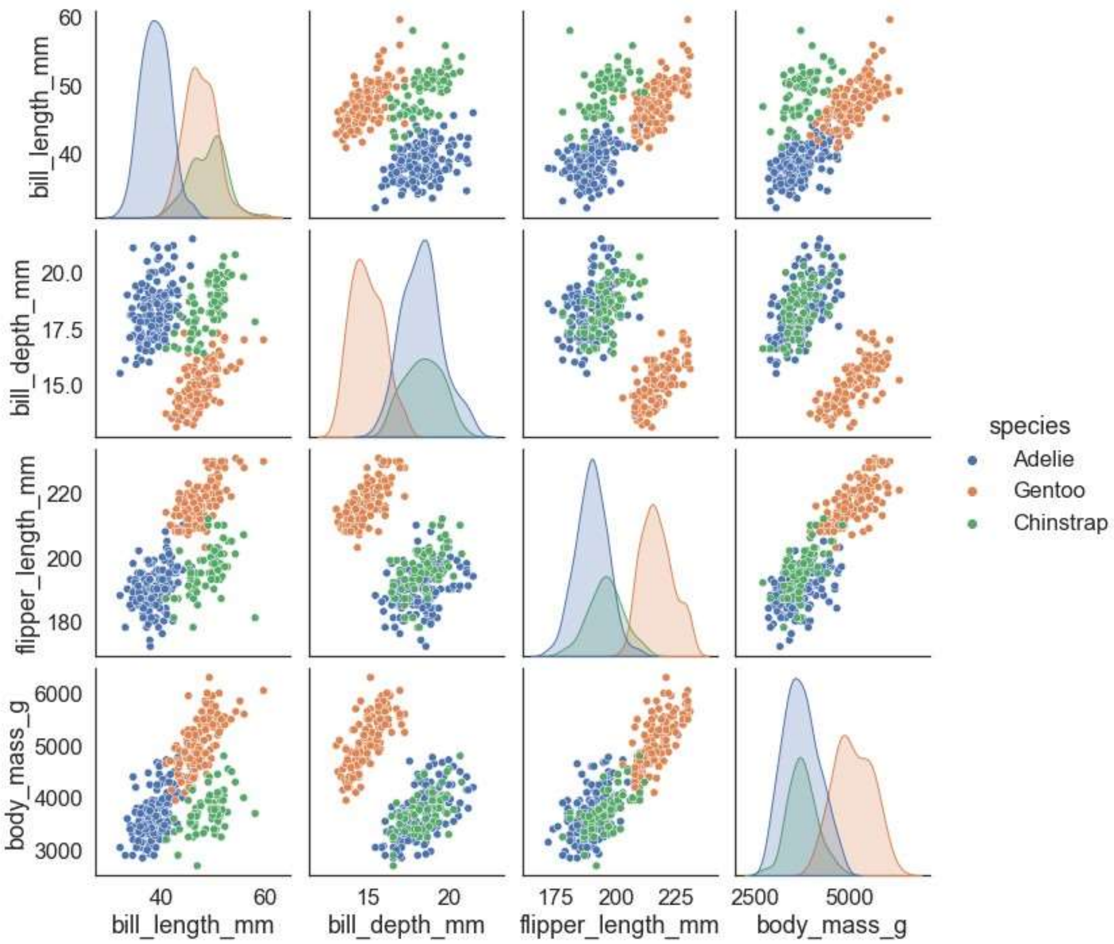


Figure 6: Pairwise analysis of features with color indicating species

The charts indicate there are distinct groupings of species in many of the pairwise comparisions, so these features are appropriate to look at. I also noticed that some of the plots (particularly flipper length vs bill depth, body mass vs flipper length, and body mass vs bill depth) did not do a great job separating Adelie and Chinstrap penguins. I wonder if there is a feature (other than bill measurements, flipper length, and body mass) that might do a better job distinguishing between those two species.

Based on the observations made above, I elected to remove year from the analysis. That variable did not seem to provide any distinguishing information. Since the data set is relatively small, I kept the remaining features to use in the models with a target variable of species. If we had a larger data set, I would be inclined to reduce the number of features by backward elimination.

Feature	**Description**
bill_length_mm	Bill length (in millimeters)
bill_depth_Mm	Bill depth (in millimeters)
flipper_length	Flipper length (in millimeters)
body_mass_g	Body mass (in grams)
sex	Sex (male or female)
island	Island (Torgersen, Biscoe, or Dream)

Data Pre-processing

In order to run the classification models, all features need to be converted to numeric values. I converted sex by assigning male the value 0 and female the value 1. I used one-hot encoding to convert the island column into numerical data.

Before running any of our four models, I used Scikit Learn to split the data for validation (70% of the data to train the model, 30% of the data to test the model for accuracy). I also scaled the data using the StandardScaler, which transforms each data value to its z-score (so that each feature has a mean of 0 and a standard deviation of 1). The scaled data was used in the k nearest neighbors (KNN), support vector machine (SVM) , and logistic regression models. The non-scaled data was used for the decision tree model.

I then fit each of the four models to the training data set using the classifiers in Scikit Learn. For the KNN model, I further split the training data set to help select the correct value for the parameter k. I ran the model for values of k ranging from 1 to 50 and looked at the error. Figure 7 gives a visual representation of the error for different values of k.

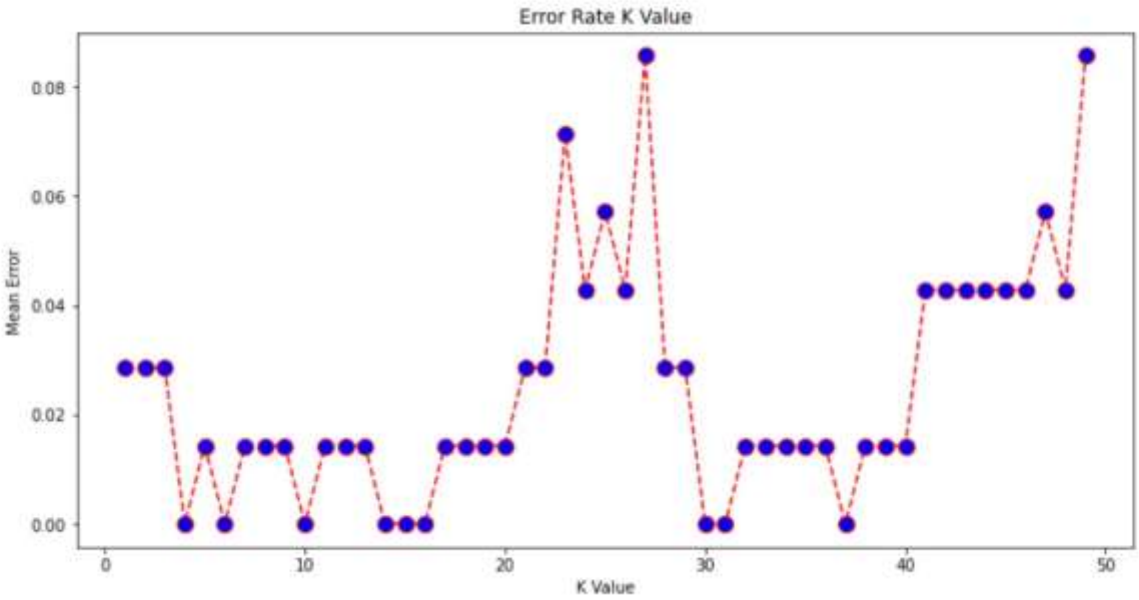


Figure 7: Error rates in the k nearest neighbor algorithm for different values of k

The error rate is lowest for several values of k. I chose to use k=4 since it is the smallest value of k (and hence most efficient for the model runtime) with the lowest error rate.

Results

I fit four machine learning models to the data: k nearest neighbor (KNN), decision tree, support vector machine (SVM), and logistic regression. After fitting each model, I used the testing portion of the data set to check the accuracy of each model. The confusion matrices for each model are shown in Figure 8.

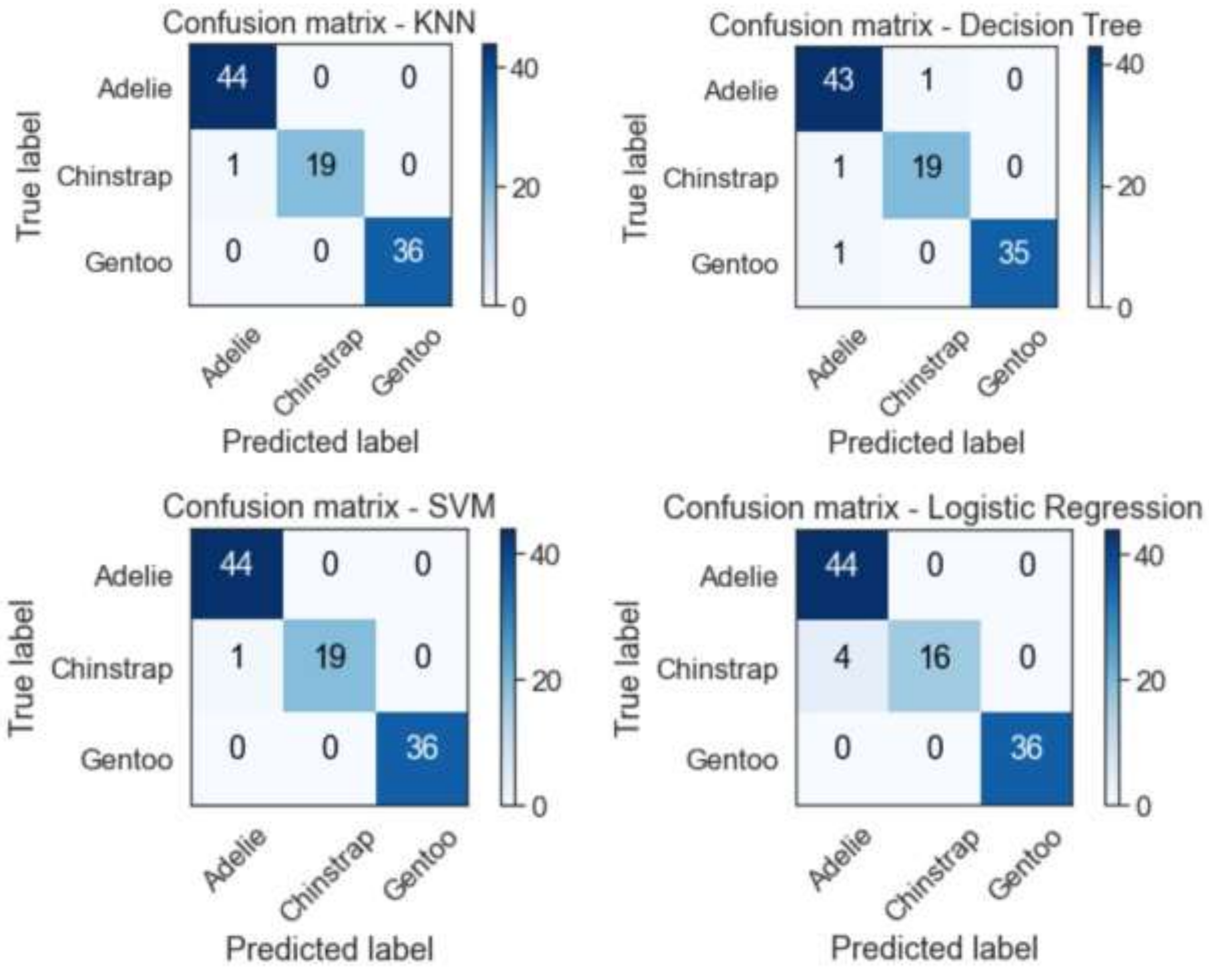


Figure 8: Confusion matrices for each machine learning model

Using the results in the confusion matrices, we can compute the precision, recall (sensitivity), and F1 score for each species and then find the average precision, recall, and F1 score across the three species.

Model	**Average Precision**	**Average Recall**	**Average F1 Score**
K nearest neighbor	0.99	0.99	0.99
Decision tree	0.97	0.97	0.97
SVM	0.99	0.99	0.99
Logistic regression	0.96	0.96	0.96

Based on the results in the table, either k nearest neighbor or support vector machine appear to tie as best model of the four to use for classifying penguins based on their bill length, bill depth, flipper length, body mass, sex, and island.

Discussion

There are five items of concern or places for improvement in this analysis.

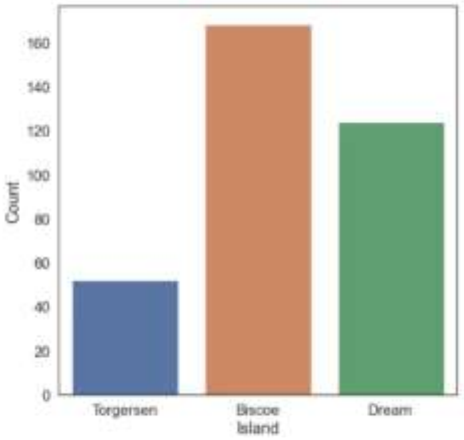
1. It is important to note that the data set is not very large. I would rather have a much larger data set.
2. The plots comparing two features at a time (Figure 6) show that Adelie and Chinstrap penguins are more difficult to distinguish based on these features. I would like to consult someone with expertise regarding penguins to determine if there is another feature that might help distinguish between those two particular species.
3. The amount of data set aside for the testing set has an impact on the accuracy – more data in the training set (and less in the testing set) could result in overfitting.
4. I would prefer to implement k-fold cross validation, rather than a single train-test split.
5. The method used for scaling the data has an impact on the model fit. One improvement would be to try multiple scaling methods and choose the combination of scaling method and model that results in the highest accuracy.

Conclusion

After fitting KNN, decision tree, SVM, and logistic regression models to the Palmer penguin data set, I found that either KNN or SVM tied for the most accurate of the four at predicting the species of a penguin based on bill length, bill depth, flipper length, body mass, sex, and island.

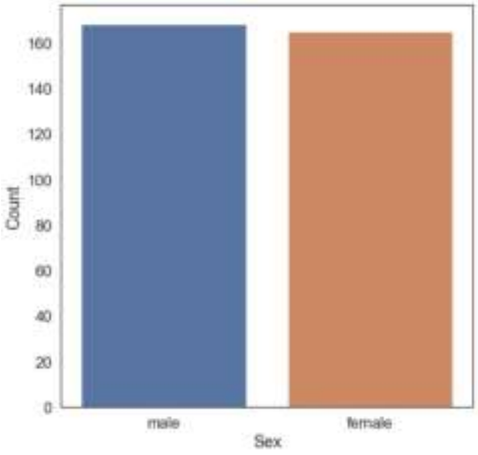
Appendix

Code and data files are available at https://github.com/jklenarz/penguins



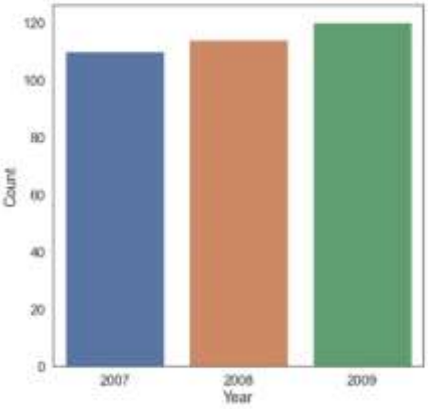
Island	Count
Torgersen	52
Biscoe	168
Dream	124

Figure A.1: Bar Chart and Frequency Table for Island



Sex	Count
Male	168
Female	165

Figure A.2: Bar Chart and Frequency Table for Sex



Year	Count
2007	110
2008	114
2009	120

Figure A.3: Bar Chart and Frequency Table for Year

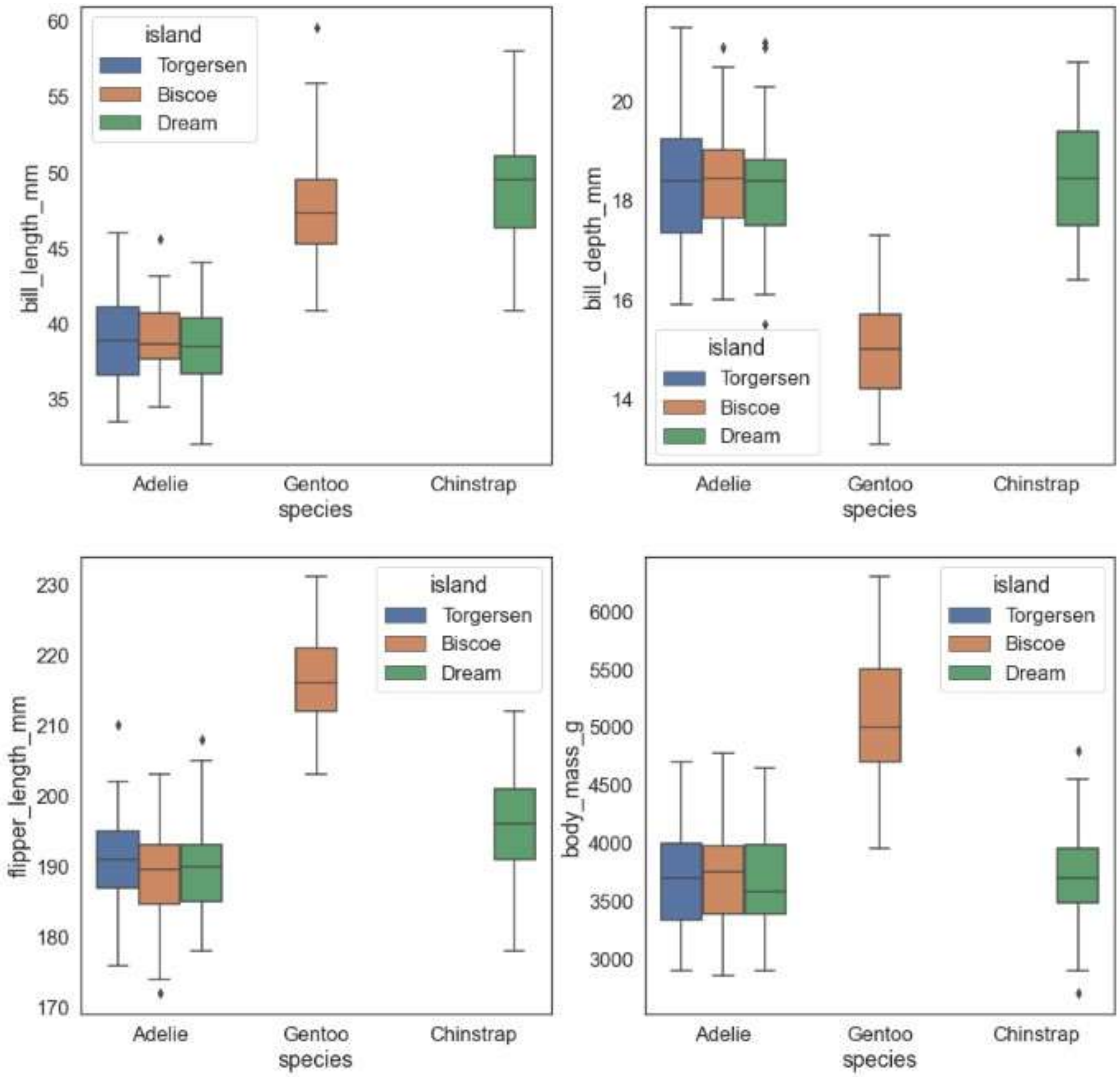


Figure A.4: Four features broken down by species and island

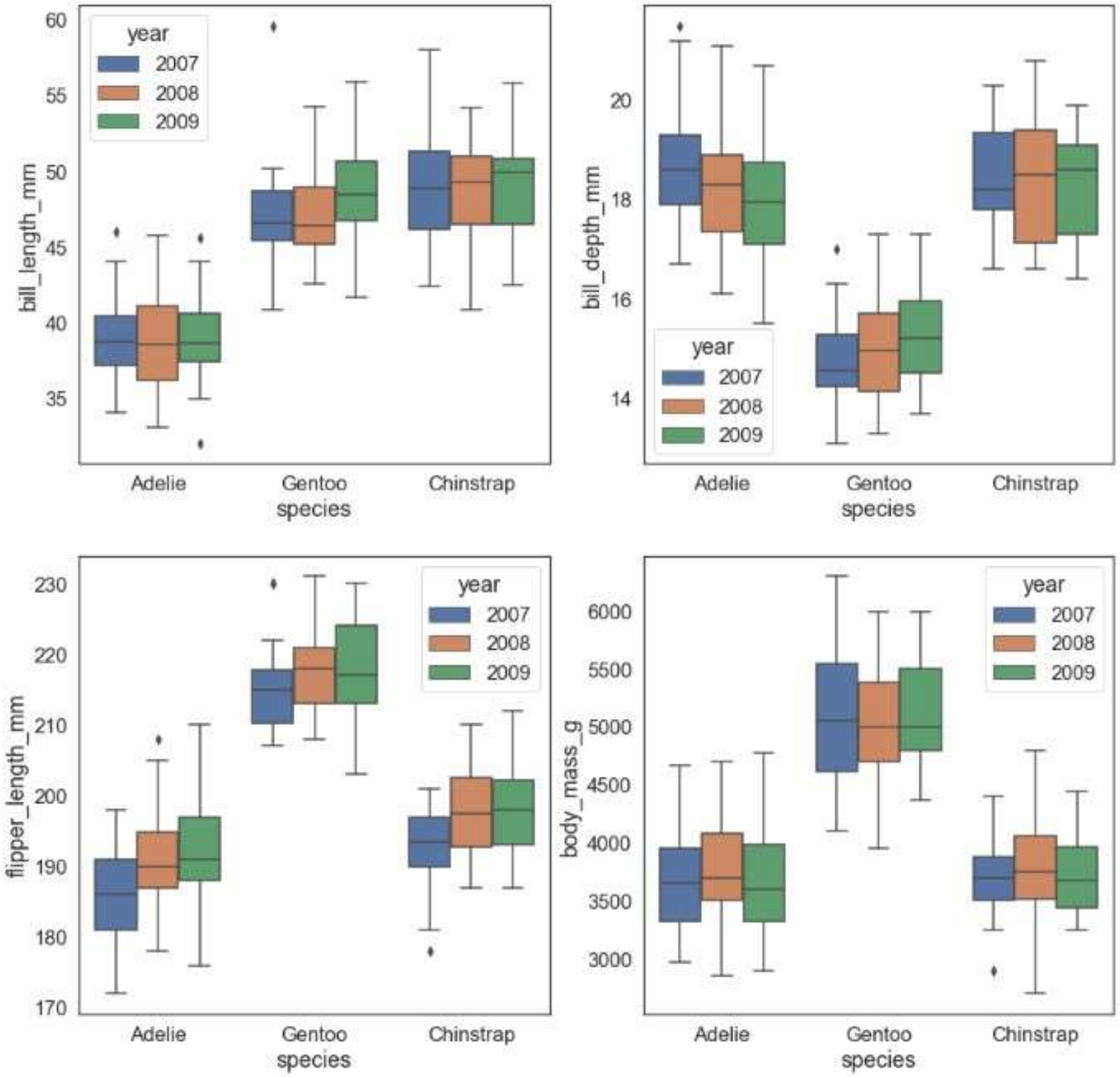


Figure A.5: Four features broken down by species and year