Coursework Template

MSc in Statistics 2025/26, Imperial College London

Andrew Duncan, Oliver Ratmann, 06042960

#Analysis of Penguin Features

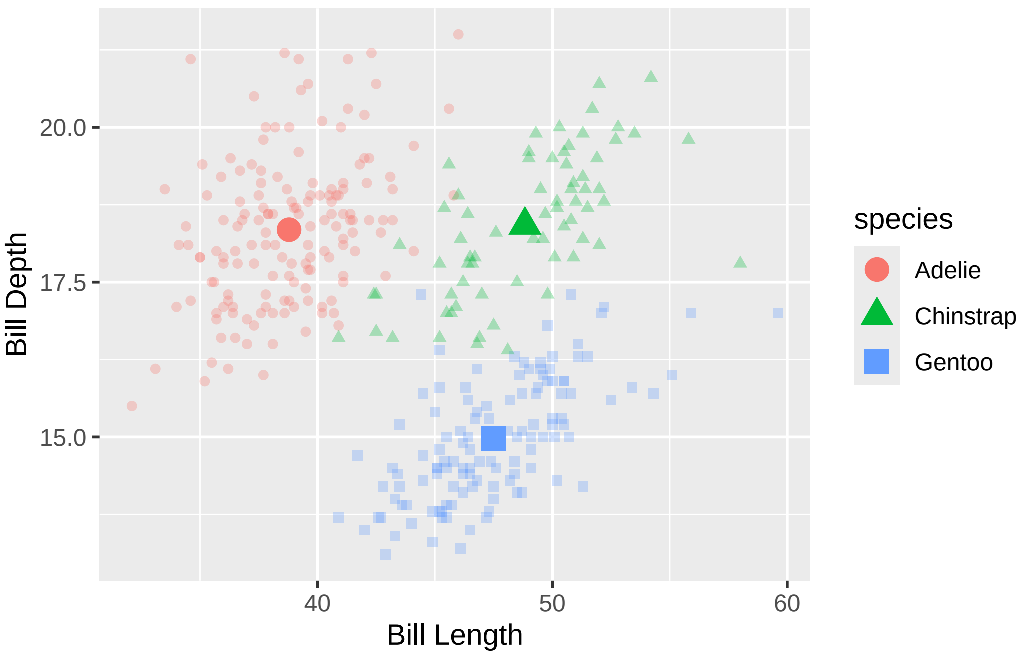
In this analysis, we will be summarising the main features of the different species of penguins. Looking at the raw data:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

To simplify our analysis, we just ignored any missing values and the summary and plots are as follows:

species	avg_bill_length	avg_bill_depth	avg_flipper_length_mm	avg_body_mass
Adelie	38.79139	18.34636	189.9536	3700.662
Chinstrap	48.83382	18.42059	195.8235	3733.088
Gentoo	47.50488	14.98211	217.1870	5076.016

```
#> Warning: Removed 2 rows containing missing values or values outside the scale range
#> (`geom_point()`).
```

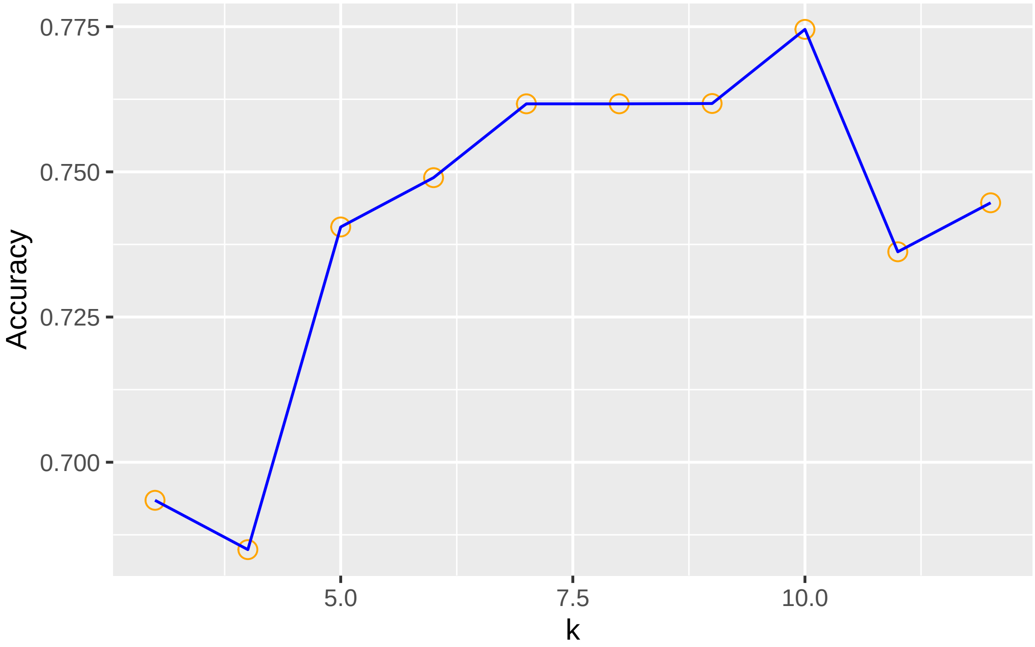


We will now create a k Nearest Neighbour Algorithm to try and make predictions for each penguin groups. To simplify, we will just split the dataset into the training set and test set with 70/30 proportions.

bill_length_mm	bill_depth_mm	body_mass_g	flipper_length	islands	target
39.1	18.7	3750	181	3	Adelie
39.5	17.4	3800	186	3	Adelie
40.3	18.0	3250	195	3	Adelie
36.7	19.3	3450	193	3	Adelie
39.3	20.6	3650	190	3	Adelie
38.9	17.8	3625	181	3	Adelie

To choose the best k, we will choose the best k, i.e.the one with highest

```
#> k-Nearest Neighbors
#>
#> 235 samples
#>   5 predictor
#>   3 classes: 'Adelie', 'Chinstrap', 'Gentoo'
#>
#> No pre-processing
#> Resampling: Cross-Validated (3 fold)
#> Summary of sample sizes: 157, 157, 156
#> Resampling results across tuning parameters:
#>
#>    k   Accuracy   Kappa
#>     3  0.6934437  0.5124645
#>     4  0.6849508  0.4973725
#>     5  0.7405063  0.5820266
#>     6  0.7489992  0.5978511
#>     7  0.7617116  0.6182014
#>     8  0.7617116  0.6139426
#>     9  0.7617657  0.6139734
#>    10  0.7745321  0.6331005
#>    11  0.7362328  0.5682052
#>    12  0.7446716  0.5815168
#>
#> Accuracy was used to select the optimal model using the largest value.
#> The final value used for the model was k = 10.
```



```
#> <ggplot2::labels> List of 2
#>  $ x: chr "k"
#>  $ y: chr "Accuracy"
```