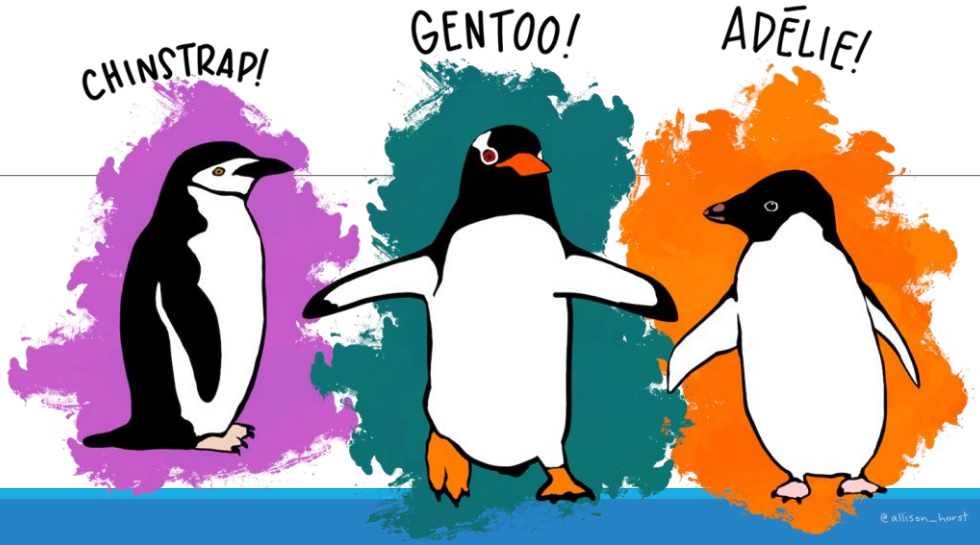
這隻企鵝是哪誰呢?

統計110馬靖宇、統計111翁瑋廷、統計111許祐誠



資料介紹

資料來自KAGGLE網站

此資料有七種特徵，三種類別型(SPECIES、ISLAND、SEX)，

四種連續型(CULMEN LENGTH&DEPTH、FLIPPER LENGTH、BODY MASS)，

共343筆資料

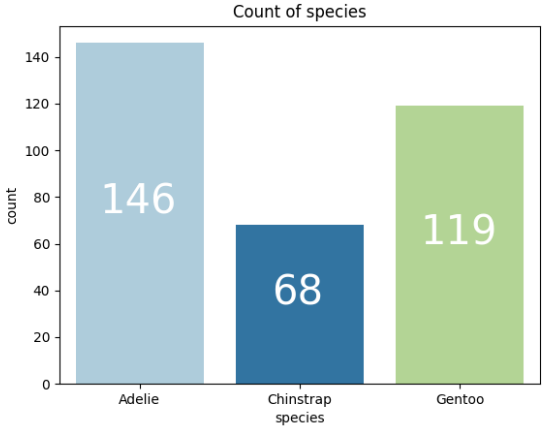
species	island	culmen_length	culmen_depth	flipper_length	body_mass	sex
Adelie	Torgersen	39.1	18.7	181	3750	MALE

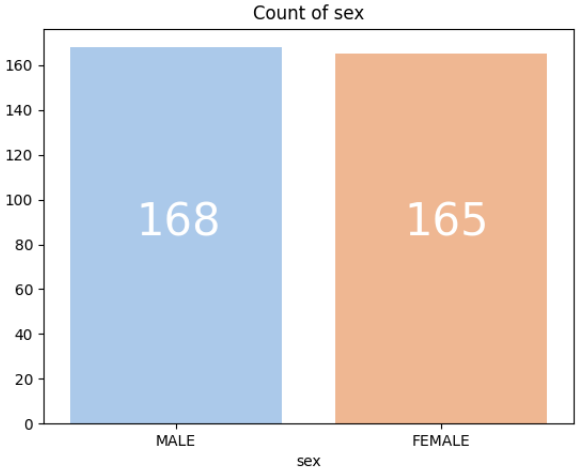
資料前處理

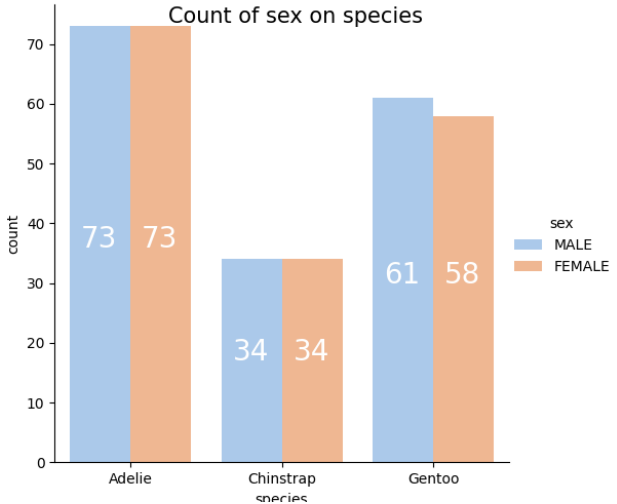
刪除任一欄位有NA的整筆資料，另外第338筆的SEX = “.” ，也將該筆刪除。一共刪除10筆，資料剩下333筆，做後續的分析。

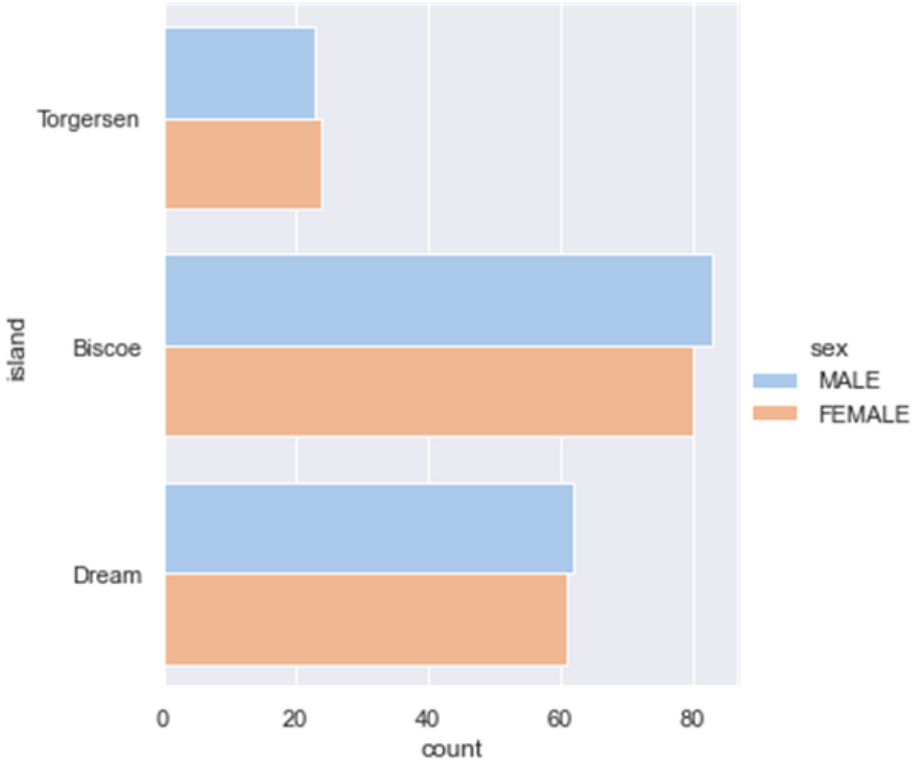
species	island	culmen_length	culmen_depth	flipper_length	body_mass	sex
Gentoo	Biscoe	44.5	15.7	217	4875	.

敘述統計量

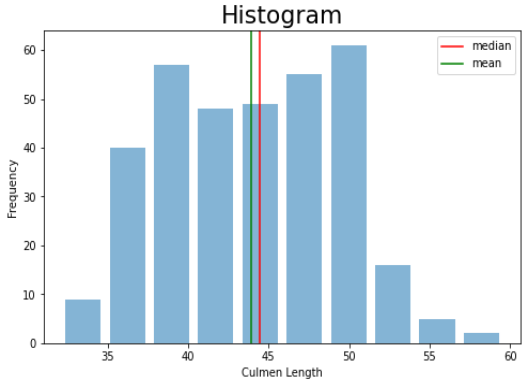

Count of species
count
146
68
119
Adelie
Chinstrap
Gentoo
species

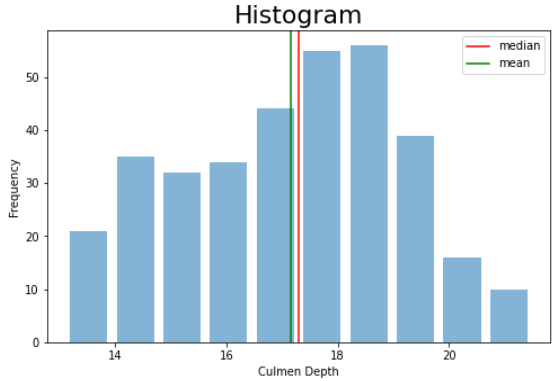

Count of sex
168
165
MALE
FEMALE
sex

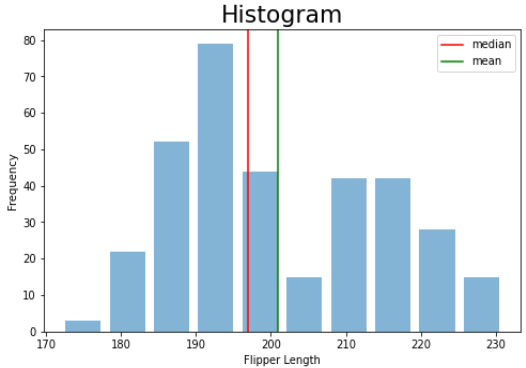

Count of sex on species
count
73
73
34
34
61
58
Adelie
Chinstrap
Gentoo
species
sex
MALE
FEMALE

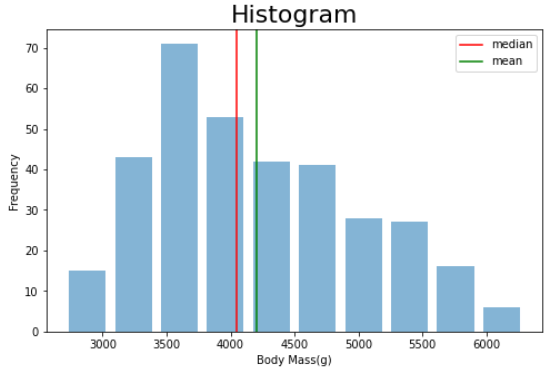

Torgersen
Biscoe
Dream
island
0
20
40
60
80
count
sex
MALE
FEMALE

Histogram

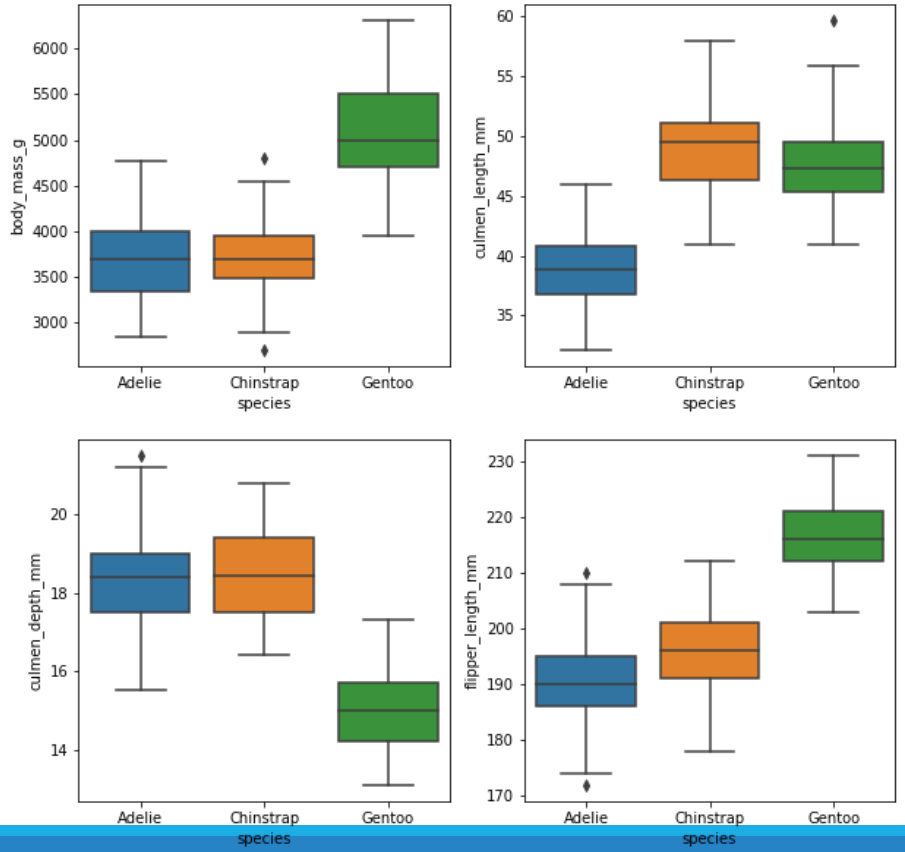

Histogram
median
mean
Frequency
Culmen Length


Histogram
median
mean
Frequency
Culmen Depth

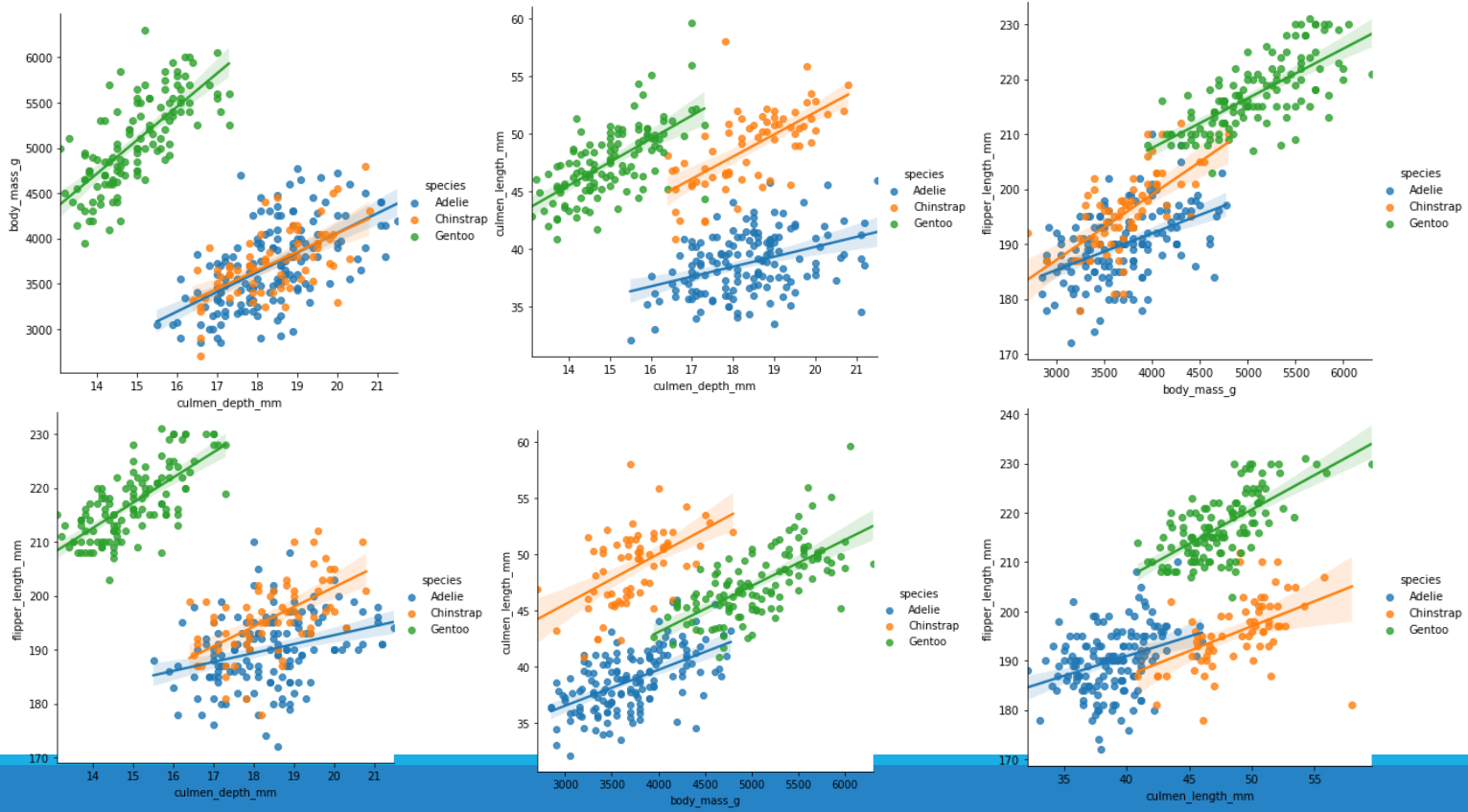

Histogram
median
mean
Frequency
Flipper Length

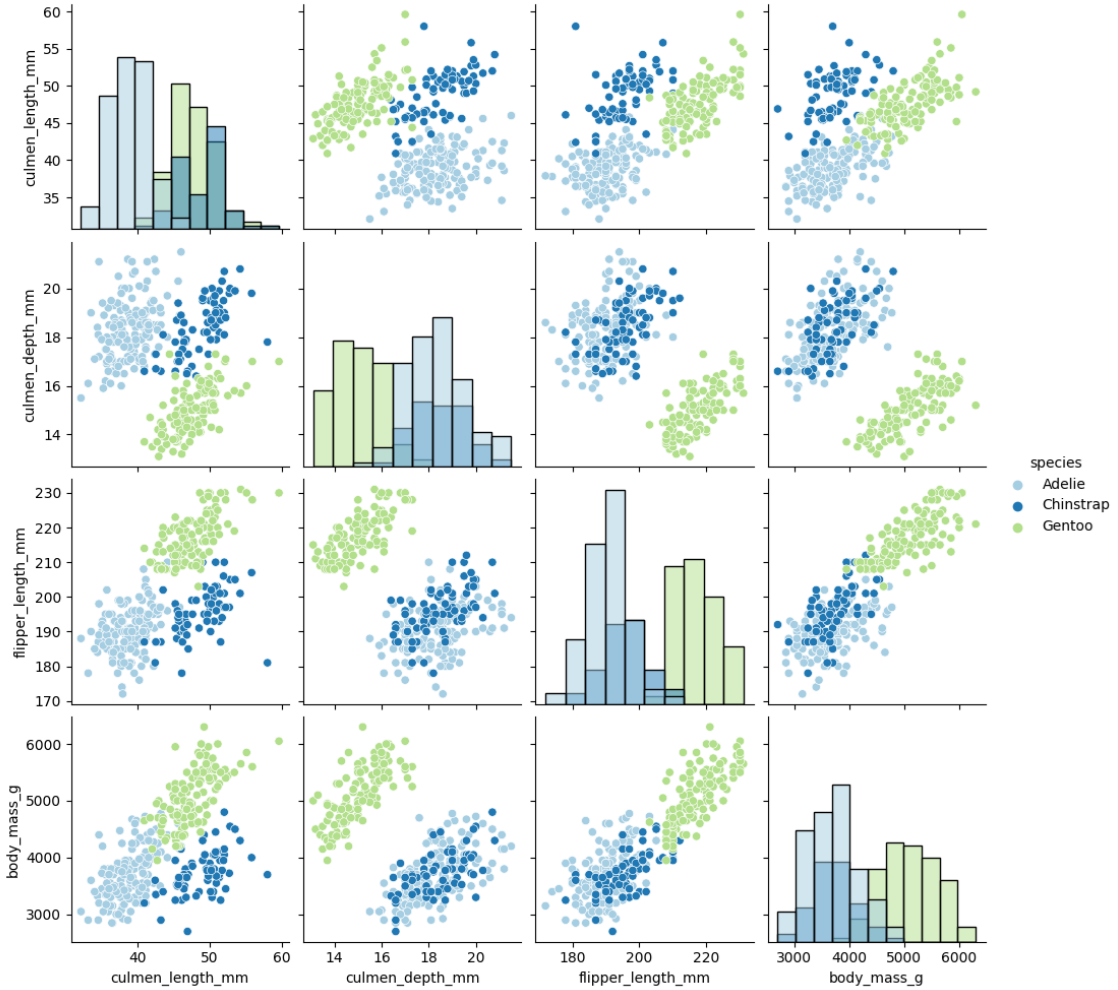

Histogram
median
mean
Frequency
Body Mass(g)

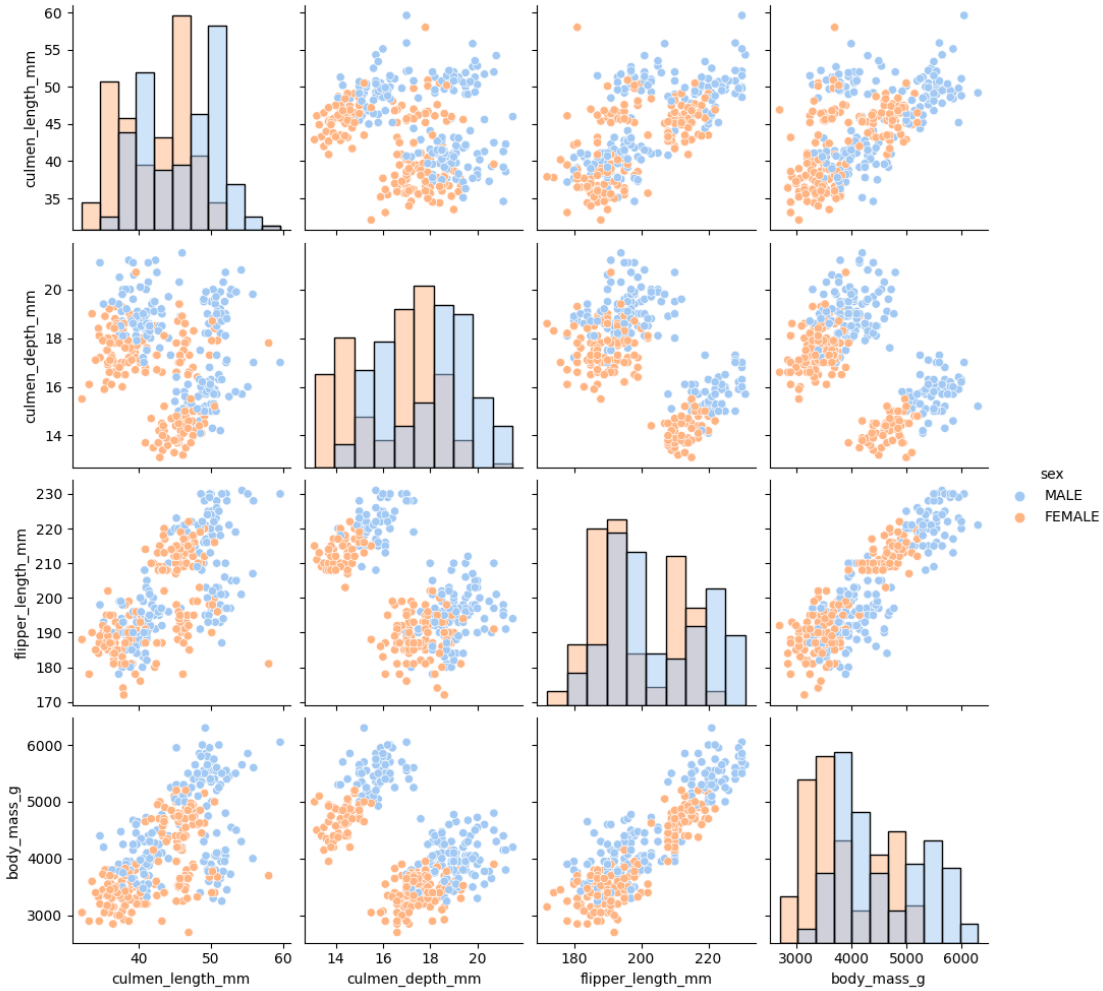
Boxplot


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

Linear Regression




culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
species
Adelie
Chinstrap
Gentoo


culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
sex
MALE
FEMALE

切割訓練、測試資料

1. train：test = 2：1

   將index的0,3,6…和1,4,7…設為訓練資料

   將index的2,5,8…設為測試資料

2. train：test = 3：1

   將index的0,4,8…、1,5,9…、2,6,10…..設為訓練資料

   將index的3,7,11...設為測試資料

模型配適和預測

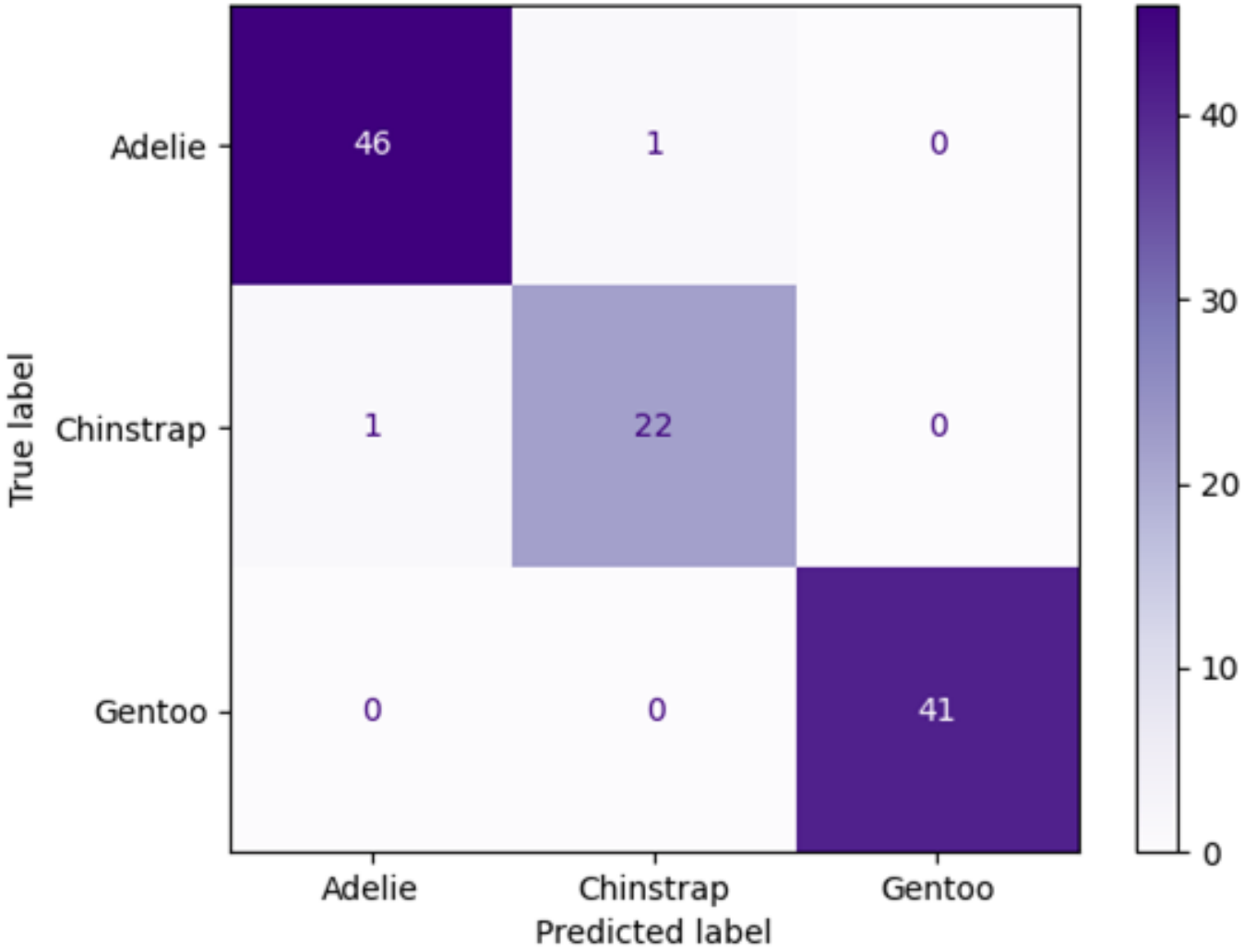
1.DECISION TREE

	Adelie	Chinstrap	Gentoo	All
accuracy	0.978723	0.956522	1	0.981982
Precision	0.978723	0.956522	1	0.981982
Recall	0.978723	0.956522	1	0.981982
F-score	0.978723	0.956522	1	0.981982
Support	47	23	41	

train：test = 2 ： 1

decicision tree model 0

tree.DecisionTreeClassifier (random_state=2)

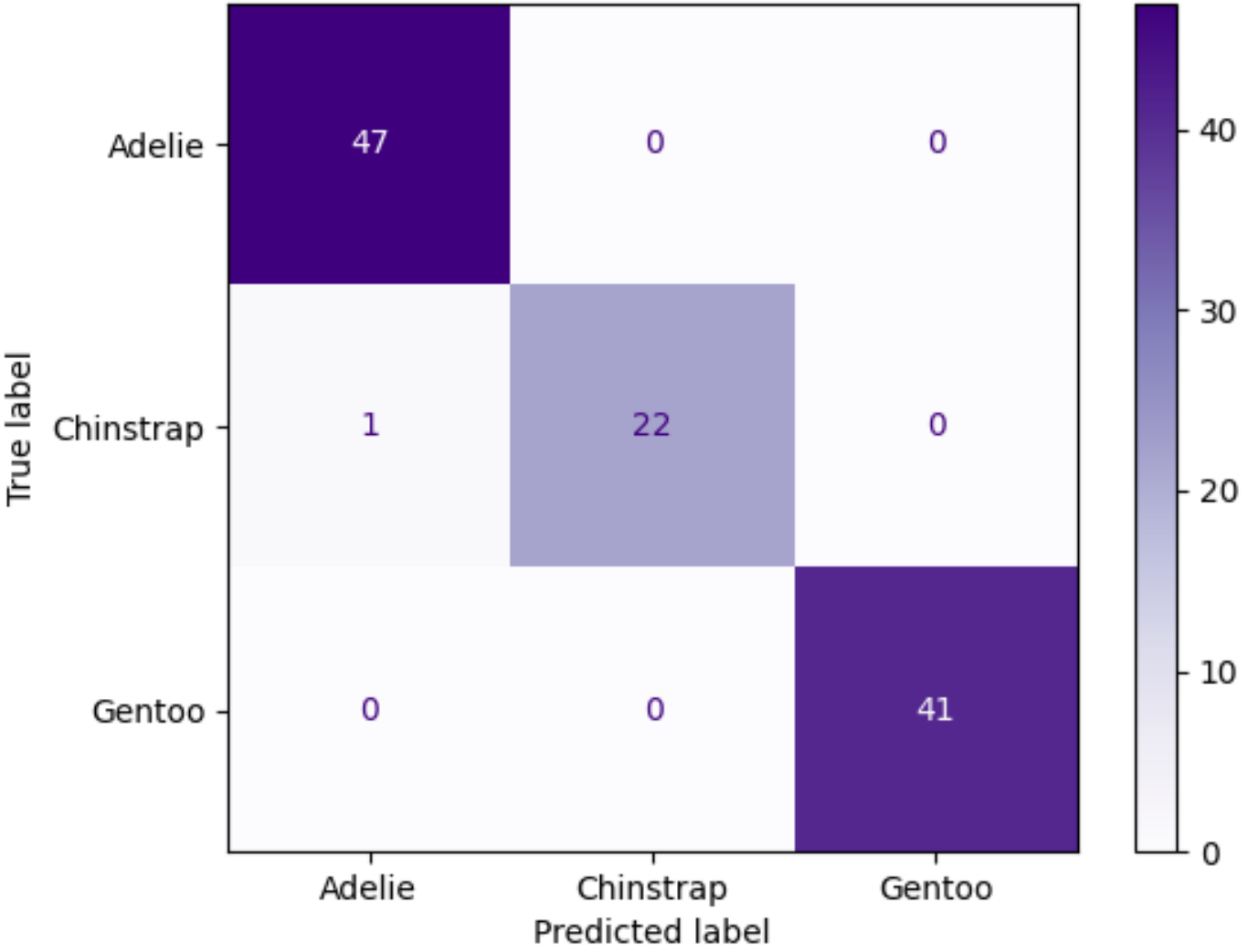

Adelie
Chinstrap
Gentoo
True label
46
1
0
1
22
0
0
0
41
Adelie
Chinstrap
Gentoo
Predicted label
40
30
20
10
0

train：test = 2 ： 1

decicision tree model 1

tree.DecisionTreeClassifier(criterion = "entropy",random_state=2)

	Adelie	Chinstrap	Gentoo	All
accuracy	0.978723	0.956522	1	0.990991
Precision	0.979167	1	1	0.990991
Recall	1	0.956522	1	0.990991
F-score	0.989474	0.977778	1	0.990991
Support	47	23	41	

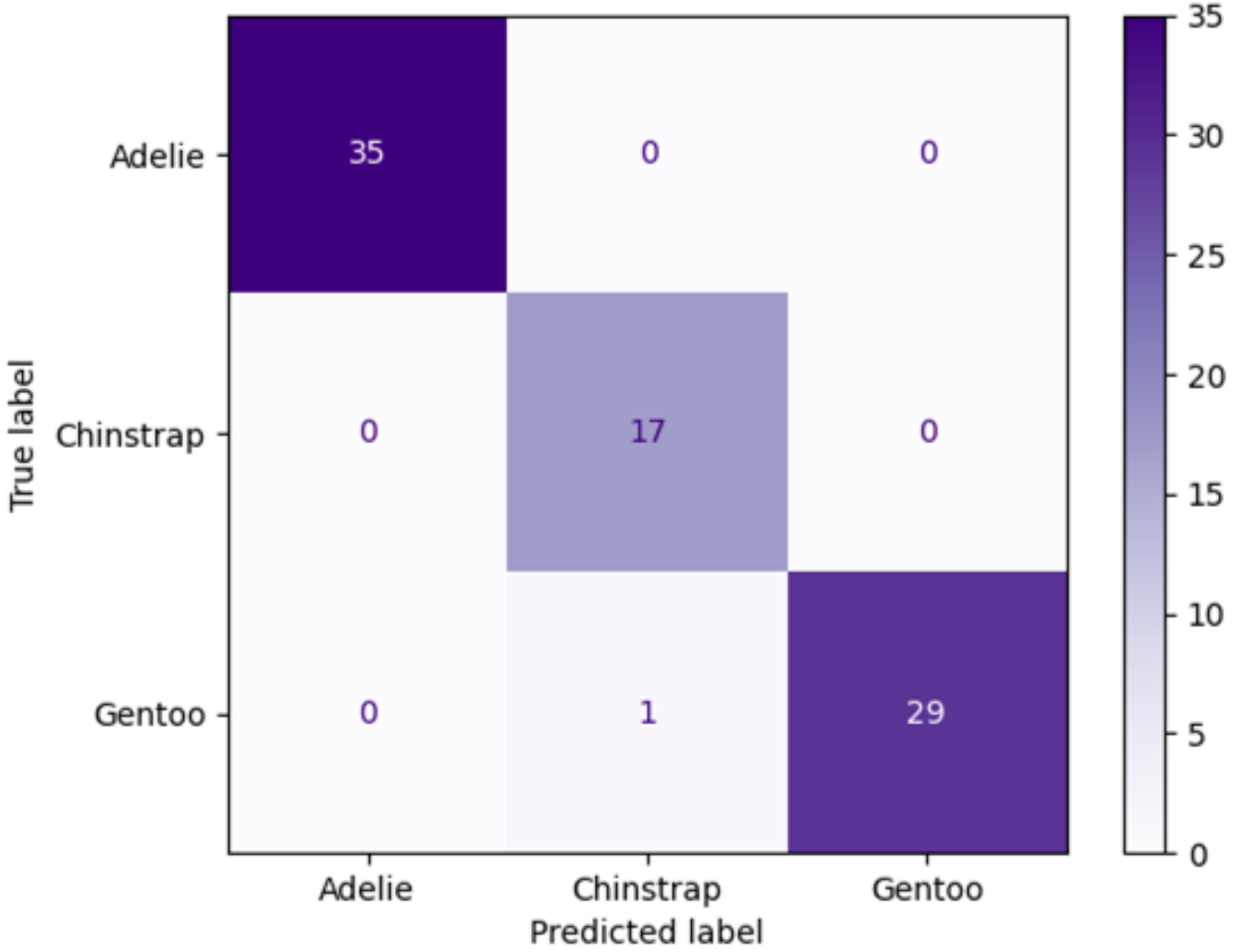

Adelie
Chinstrap
Gentoo
True label
47
0
0
1
22
0
0
0
41
Adelie
Chinstrap
Gentoo
Predicted label
40
30
20
10
0

train：test = 3 ： 1

decicision tree model 0

tree.DecisionTreeClassifier(random_state=2)

	Adelie	Chinstrap	Gentoo	All
accuracy	1	1	0.966667	0.987805
Precision	1	0.944444	1	0.987805
Recall	1	1	0.966667	0.987805
F-score	1	0.971429	0.983051	0.987805
Support	35	17	30	

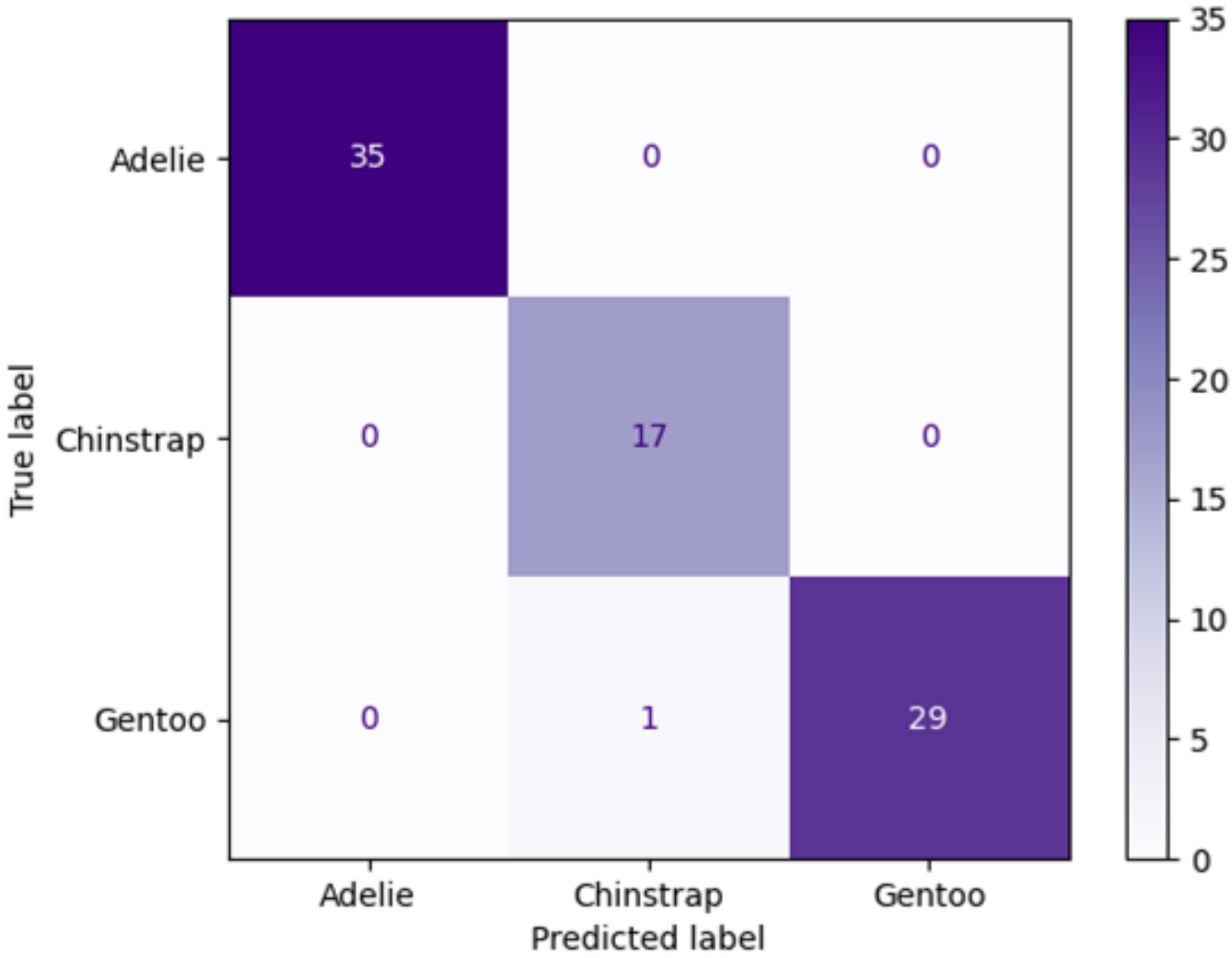

Adelie
Chinstrap
Gentoo
True label
35
0
0
0
17
0
0
1
29
Adelie
Chinstrap
Gentoo
Predicted label
35
30
25
20
15
10
5
0

train : test = 3 : 1

decicision tree model 1

tree.DecisionTreeClassifier(criterion = "entropy",random_state=2)

	Adelie	Chinstrap	Gentoo	All
accuracy	1	1	0.966667	0.987805
Precision	1	0.944444	1	0.987805
Recall	1	1	0.966667	0.987805
F-score	1	0.971429	0.983051	0.987805
Support	35	17	30	


Adelie
Chinstrap
Gentoo
True label
35
0
0
0
17
0
0
1
29
Adelie
Chinstrap
Gentoo
Predicted label
35
30
25
20
15
10
5
0

模型配適和預測

2. RANDOM FOREST

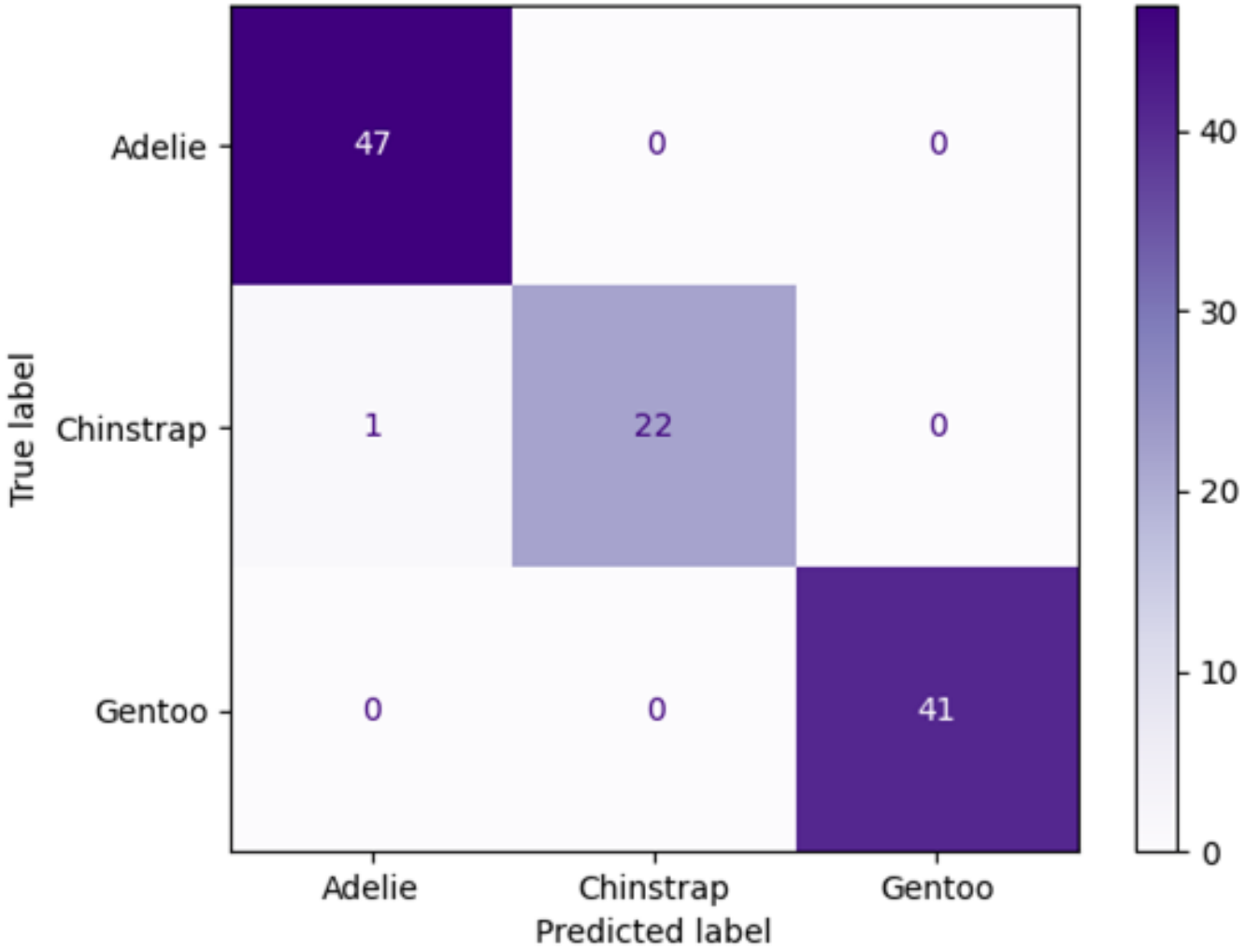
```
train : test = 2 : 1

random forest model 0 1 2 3 4 5
6 7 8

max_sample = [0.6,0.7,0.8]
max_feature = [2,3,4]

ensemble.RandomForestClassifier(
criterion="entropy",random_state
=100,max_samples,max_features)
```

	Adelie	Chinstrap	Gentoo	All
accuracy	1	0.956522	1	0.990991
Precision	0.979167	1	1	0.990991
Recall	1	0.956522	1	0.990991
F-score	0.989474	0.977778	1	0.990991
Support	47	23	41	


Adelie
Chinstrap
Gentoo
True label
47
0
0
1
22
0
0
0
41
Adelie
Chinstrap
Gentoo
Predicted label
40
30
20
10
0

	Adelie	Chinstrap	Gentoo	All
accuracy	1	1	1	1
Precision	1	1	1	1
Recall	1	1	1	1
F-score	1	1	1	1
Support	35	17	30	

train : test = 3 : 1

random forest model 0 1 3 6 7

```
max_sample = [0.6,0.7,0.8]
max_feature = [2,3,4]
```

```
ensemble.RandomForestClassifier(criterion="entropy",random_state=100,max_samples,max_features)
```

	Adelie	Chinstrap	Gentoo	All
accuracy	1	1	0.966667	0.987805
Precision	0.972222	1	1	0.987805
Recall	1	1	0.966667	0.987805
F-score	0.985915	1	0.983051	0.987805
Support	35	17	30	

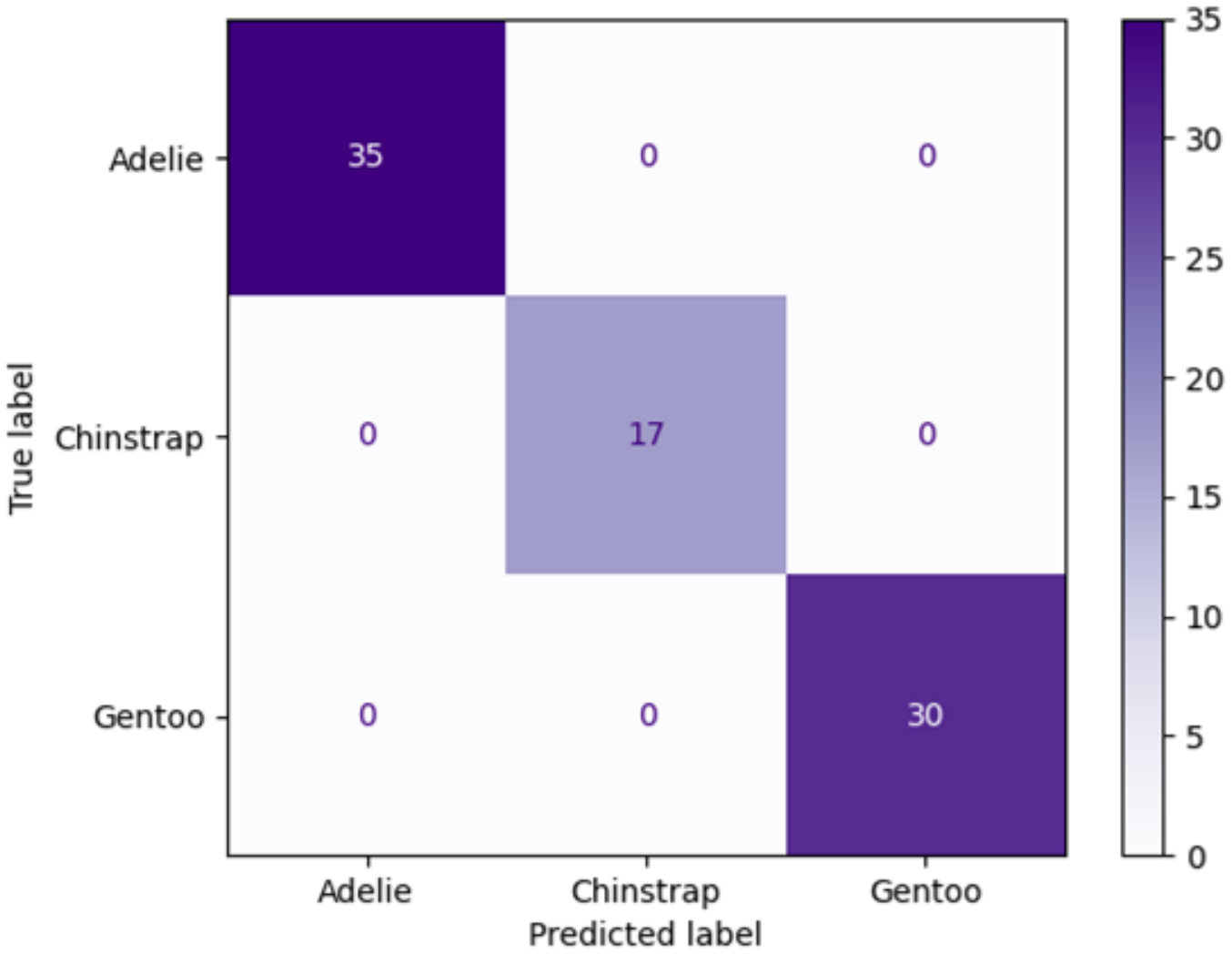
train：test = 3 ： 1

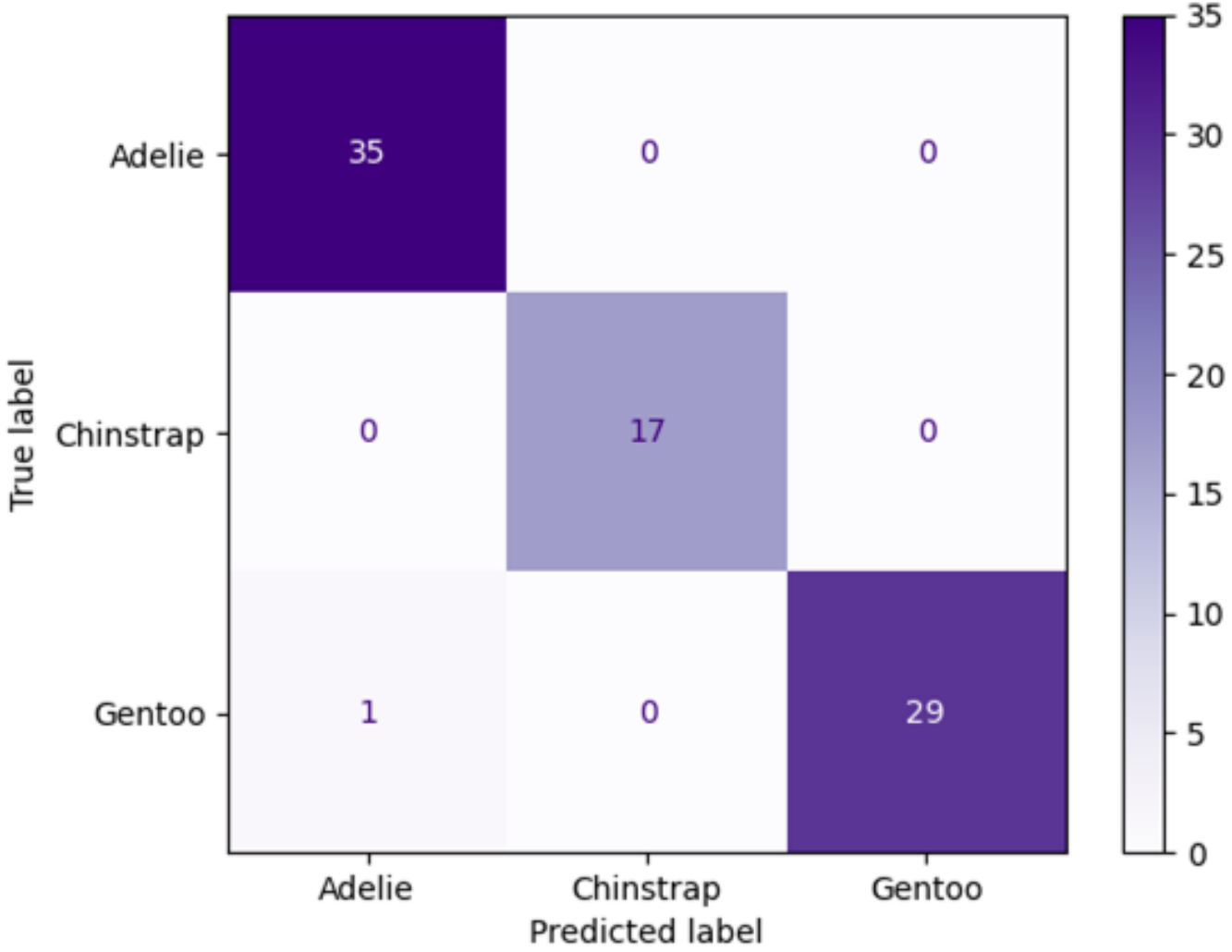
---

random forest model 2 4 5 8

max_sample = [0.6,0.7,0.8]
max_feature = [2,3,4]

ensemble.RandomForestClassifier(criterion="entropy",random_state=100,max_samples,max_features)


Adelie
Chinstrap
Gentoo
True label
35
0
0
0
17
0
0
0
30
Adelie
Chinstrap
Gentoo
Predicted label
35
30
25
20
15
10
5
0

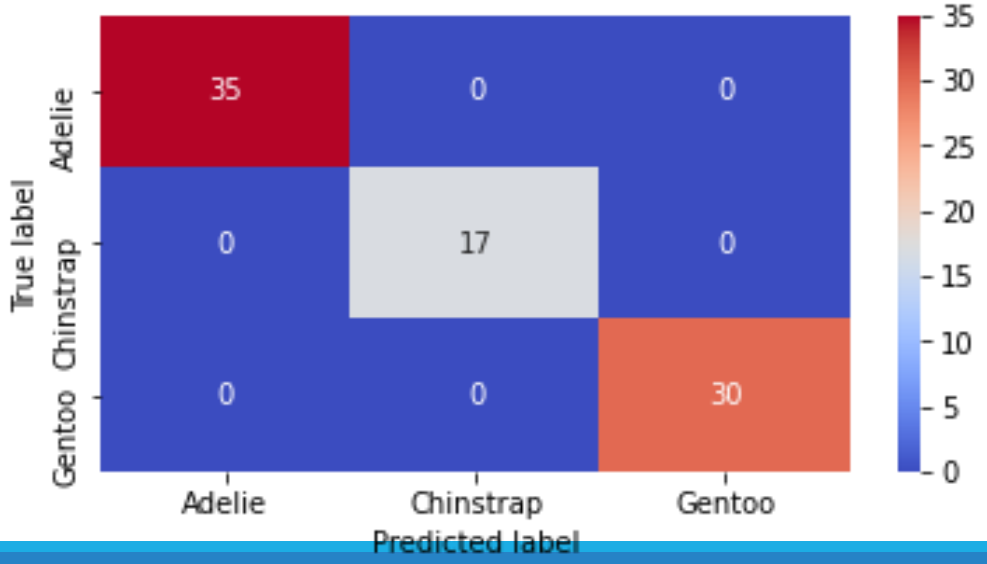

Adelie
Chinstrap
Gentoo
True label
35
0
0
0
17
0
1
0
29
Adelie
Chinstrap
Gentoo
Predicted label
35
30
25
20
15
10
5
0

模型配適和預測

3.BAGGINGCLASSIFIER

BaggingClassifier

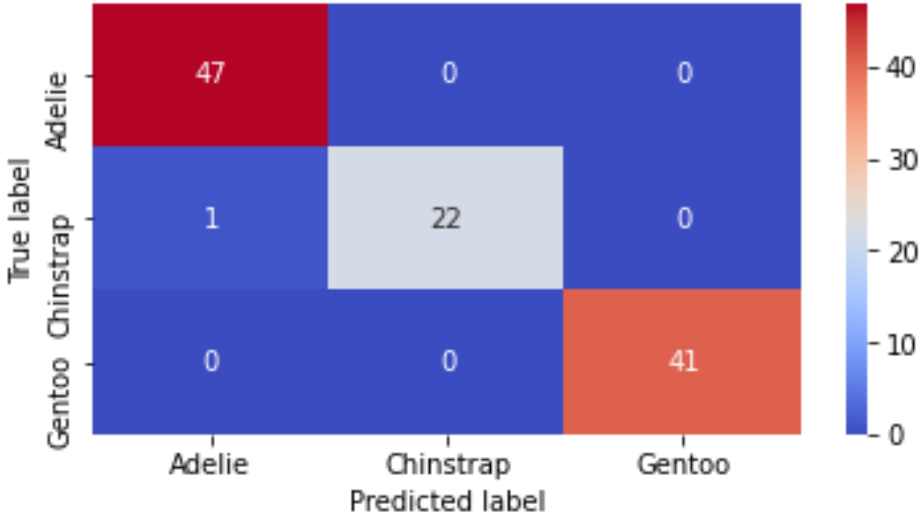
train：test = 3 ： 1
n_estimators = [8,12,16]
max_samples = [0.5,0.6,0.7,0.8]
max_features = [0.5,0.6,0.7,0.8]
BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

BaggingClassifier

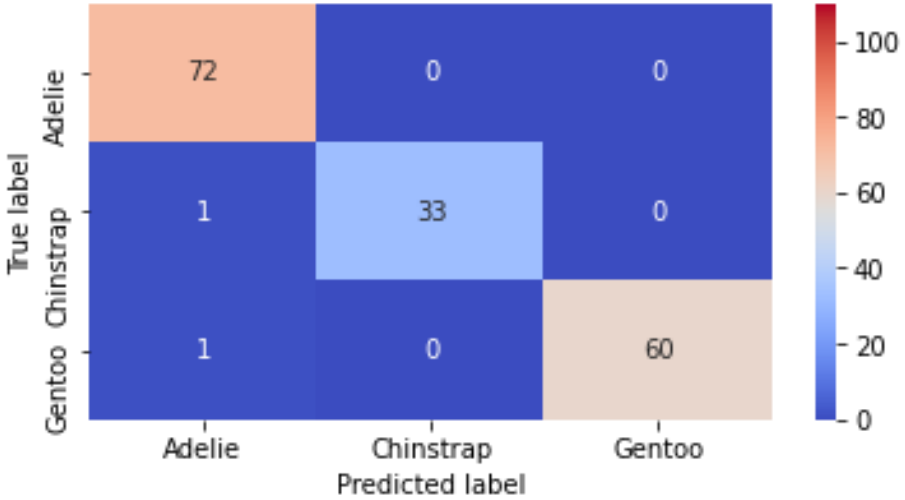
train：test = 2 ： 1
n_estimators = [8,12,16]
max_samples = [0.5,0.6,0.7,0.8]
max_features = [0.5,0.6,0.7,0.8]
BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

BaggingClassifier

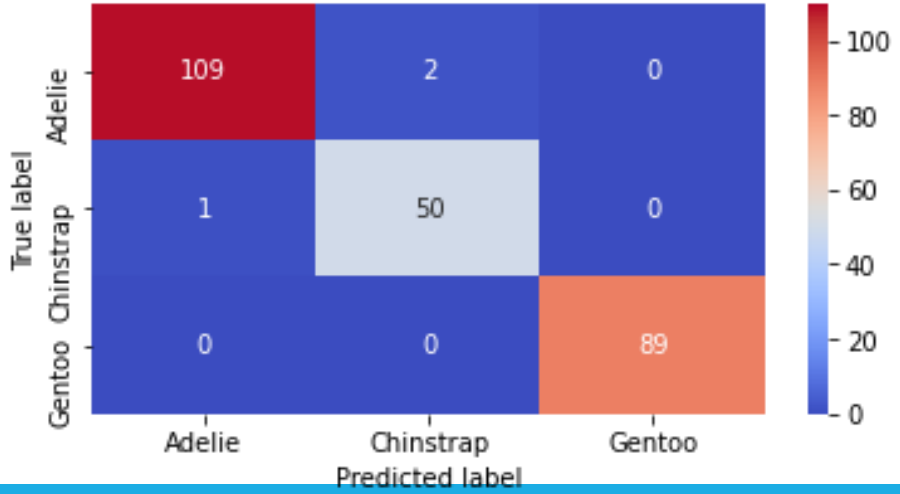
train：test = 1 ： 1
n_estimators = [8,12,16]
max_samples = [0.5,0.6,0.7,0.8]
max_features = [0.5,0.6,0.7,0.8]
BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



	precision	recall	f1-score	support
Adelie	0.97	1.00	0.99	72
Chinstrap	1.00	0.97	0.99	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	0.99	0.98	0.99	167
weighted avg	0.99	0.99	0.99	167

BaggingClassifier

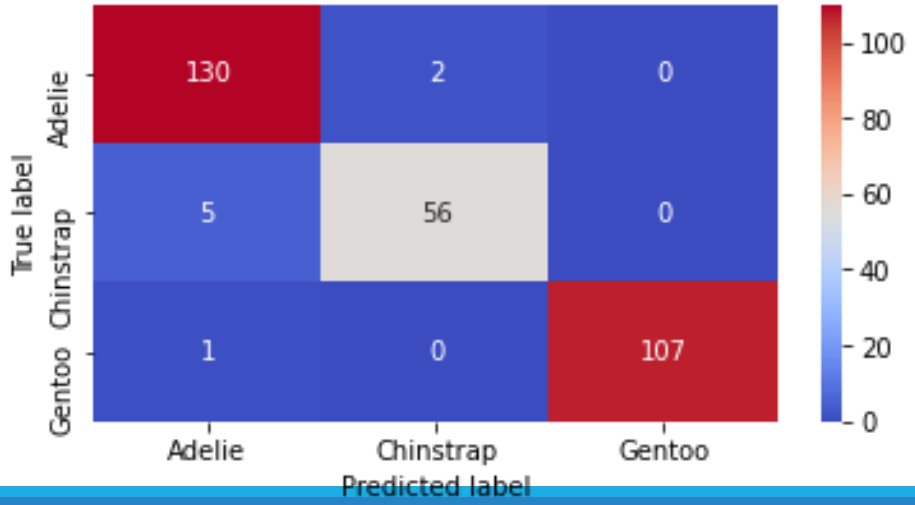
train：test = 1 ： 3
n_estimators = [8,12,16]
max_samples = [0.5,0.6,0.7,0.8]
max_features = [0.5,0.6,0.7,0.8]
BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



	precision	recall	f1-score	support
Adelie	0.99	0.98	0.99	111
Chinstrap	0.96	0.98	0.97	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.99	251
macro avg	0.98	0.99	0.99	251
weighted avg	0.99	0.99	0.99	251

BaggingClassifier

train：test = 1 ： 9
n_estimators = [8,12,16]
max_samples = [0.5,0.6,0.7,0.8]
max_features = [0.5,0.6,0.7,0.8]
BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



	precision	recall	f1-score	support
Adelie	0.96	0.98	0.97	132
Chinstrap	0.97	0.92	0.94	61
Gentoo	1.00	0.99	1.00	108
accuracy			0.97	301
macro avg	0.97	0.96	0.97	301
weighted avg	0.97	0.97	0.97	301

BaggingClassifier

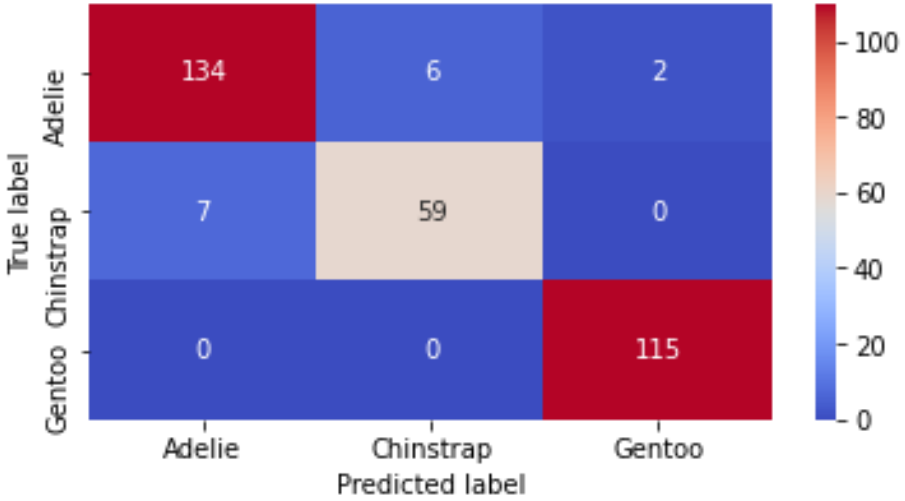
train只取其中十筆，其餘為test

n_estimators = [8,12,16]

max_samples = [0.5,0.6,0.7,0.8]

max_features = [0.5,0.6,0.7,0.8]

BaggingClassifier(n_estimators,max_samples,max_features,random_state=100)



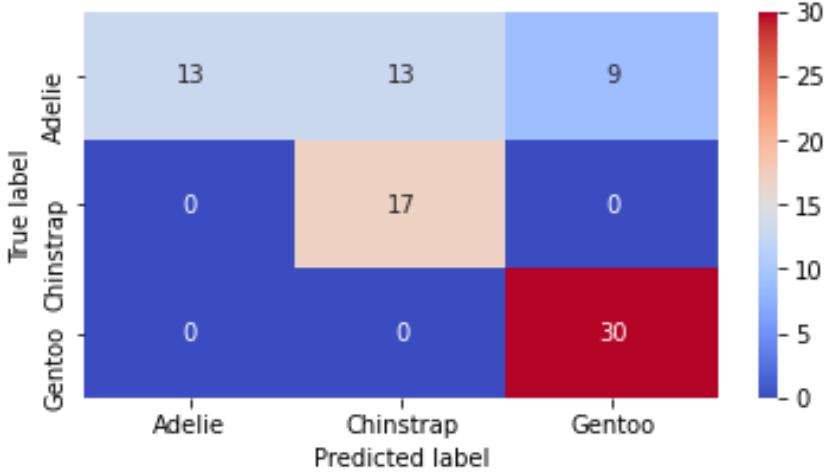
	precision	recall	f1-score	support
Adelie	0.95	0.94	0.95	142
Chinstrap	0.91	0.89	0.90	66
Gentoo	0.98	1.00	0.99	115
accuracy			0.95	323
macro avg	0.95	0.95	0.95	323
weighted avg	0.95	0.95	0.95	323

模型配適和預測

4.GAUSSIANNB

GaussianNB

train：test = 3 ： 1
GaussianNB()

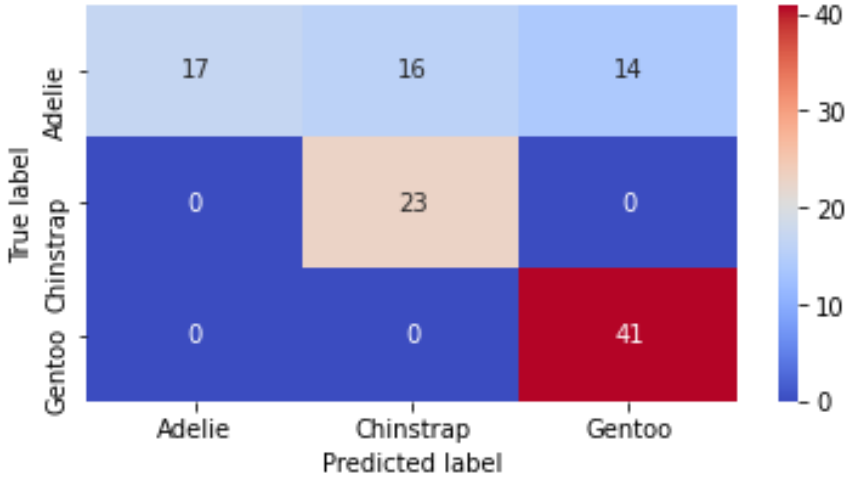


	precision	recall	f1-score	support
Adelie	1.00	0.37	0.54	35
Chinstrap	0.57	1.00	0.72	17
Gentoo	0.77	1.00	0.87	30
accuracy			0.73	82
macro avg	0.78	0.79	0.71	82
weighted avg	0.83	0.73	0.70	82

GaussianNB

train：test = 2 ： 1

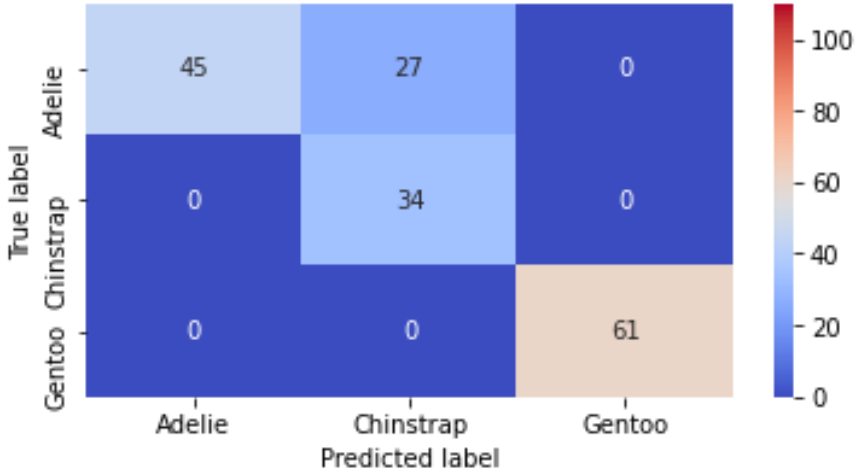
GaussianNB()



	precision	recall	f1-score	support
Adelie	1.00	0.36	0.53	47
Chinstrap	0.59	1.00	0.74	23
Gentoo	0.75	1.00	0.85	41
accuracy			0.73	111
macro avg	0.78	0.79	0.71	111
weighted avg	0.82	0.73	0.69	111

GaussianNB

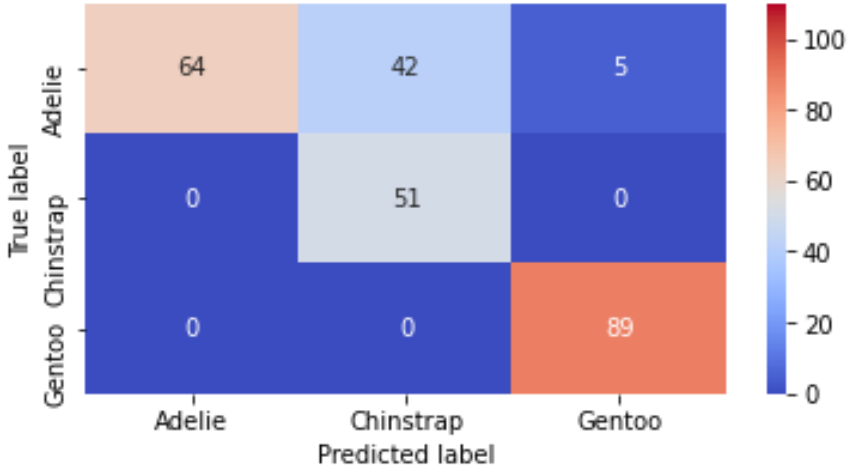
train：test = 1 ： 1
GaussianNB()



	precision	recall	f1-score	support
Adelie	1.00	0.62	0.77	72
Chinstrap	0.56	1.00	0.72	34
Gentoo	1.00	1.00	1.00	61
accuracy			0.84	167
macro avg	0.85	0.88	0.83	167
weighted avg	0.91	0.84	0.84	167

GaussianNB

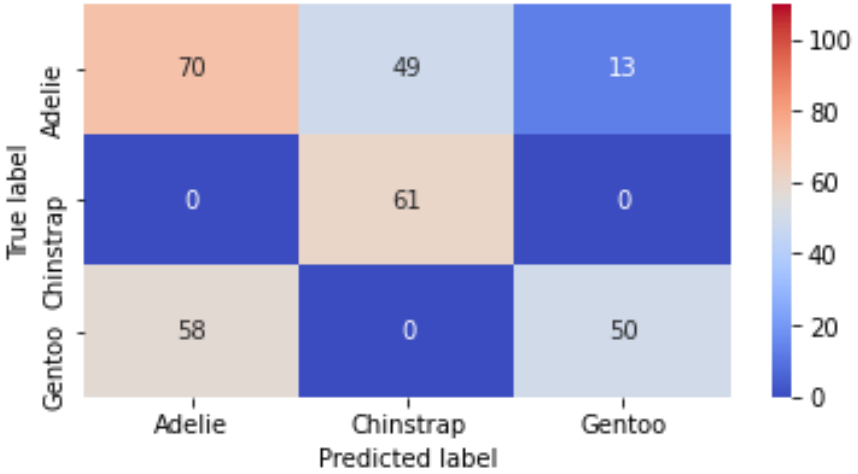
train：test = 1 ： 3
GaussianNB()



	precision	recall	f1-score	support
Adelie	1.00	0.58	0.73	111
Chinstrap	0.55	1.00	0.71	51
Gentoo	0.95	1.00	0.97	89
accuracy			0.81	251
macro avg	0.83	0.86	0.80	251
weighted avg	0.89	0.81	0.81	251

GaussianNB

train：test = 1 ： 9
GaussianNB()

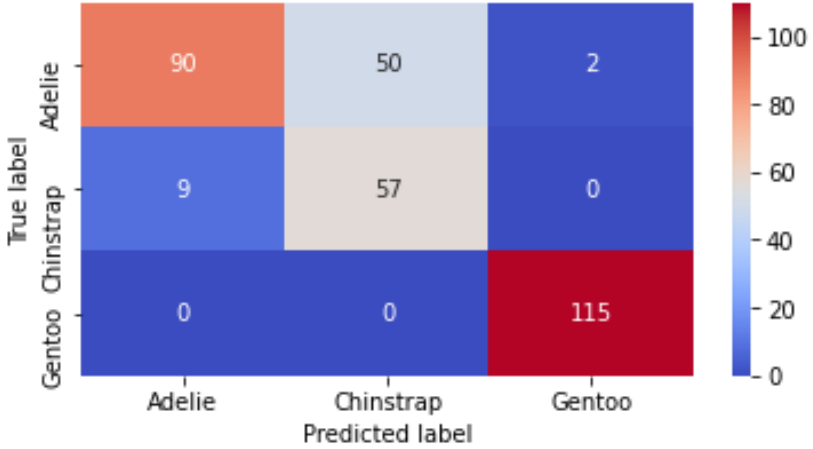


	precision	recall	f1-score	support
Adelie	0.55	0.53	0.54	132
Chinstrap	0.55	1.00	0.71	61
Gentoo	0.79	0.46	0.58	108
accuracy			0.60	301
macro avg	0.63	0.66	0.61	301
weighted avg	0.64	0.60	0.59	301

GaussianNB

train只取其中十筆，其餘為test

GaussianNB()



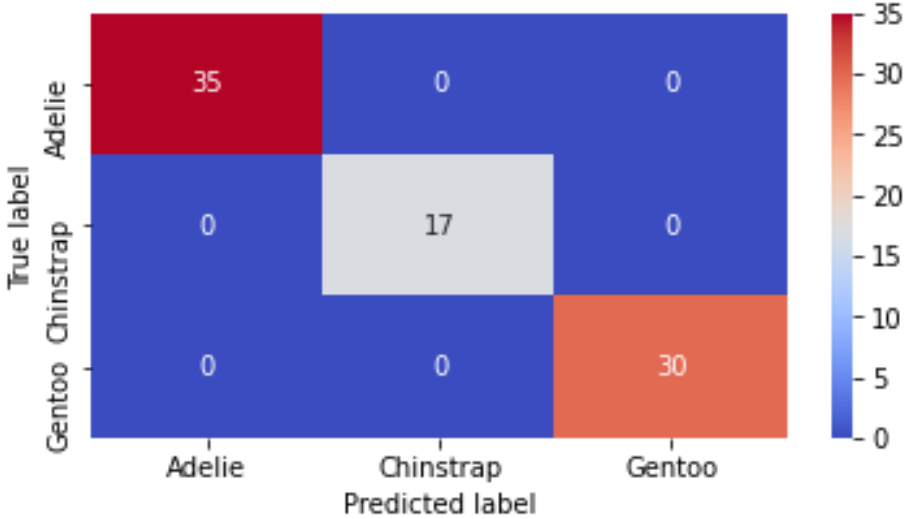
	precision	recall	f1-score	support
Adelie	0.91	0.63	0.75	142
Chinstrap	0.53	0.86	0.66	66
Gentoo	0.98	1.00	0.99	115
accuracy			0.81	323
macro avg	0.81	0.83	0.80	323
weighted avg	0.86	0.81	0.82	323

模型配適和預測

5.GRADIENTBOOSTINGCLASSIFIER

GradientBoostingClassifier

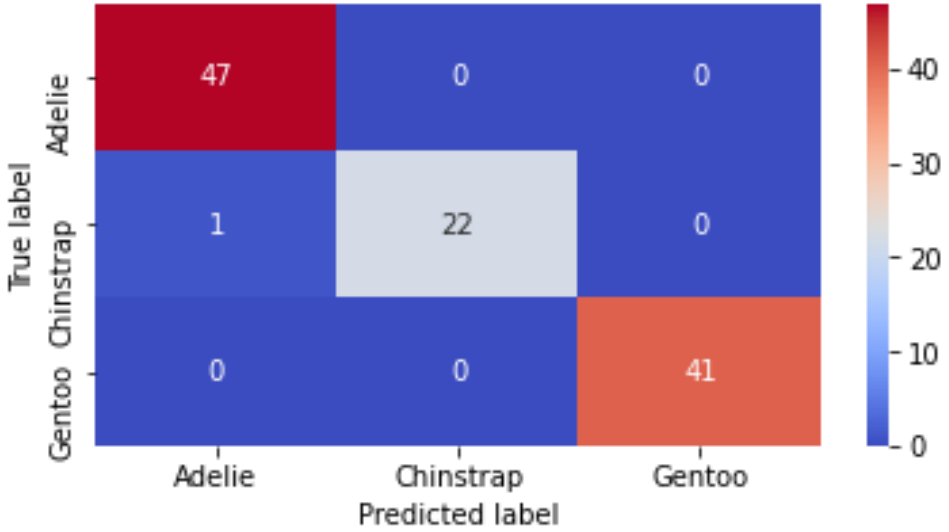
train：test = 3 ： 1
n_estimators = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

GradientBoostingClassifier

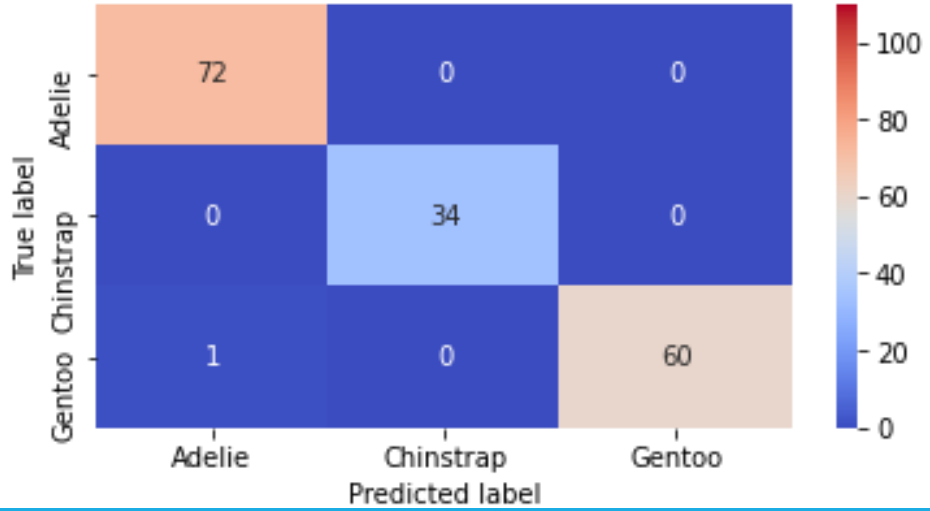
train：test = 2 ： 1
n_estimators = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

GradientBoostingClassifier

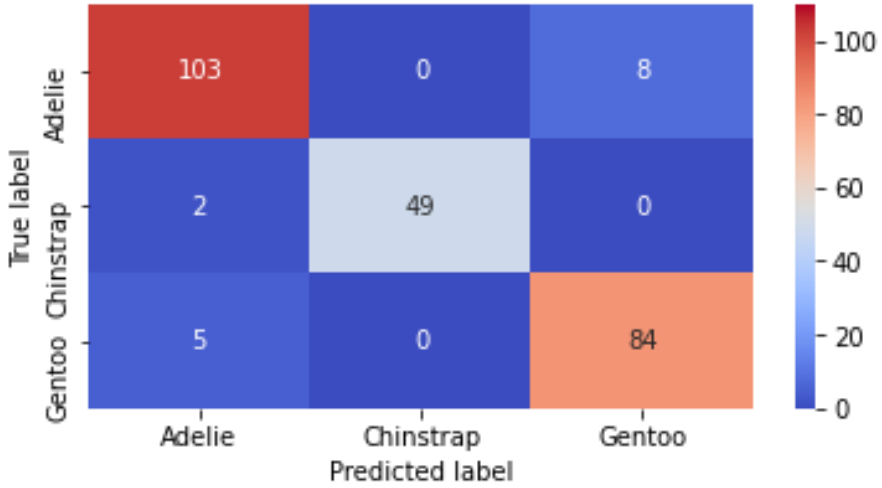
train：test = 1 ： 1
n_estimators = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.99	1.00	0.99	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	1.00	0.99	0.99	167
weighted avg	0.99	0.99	0.99	167

GradientBoostingClassifier

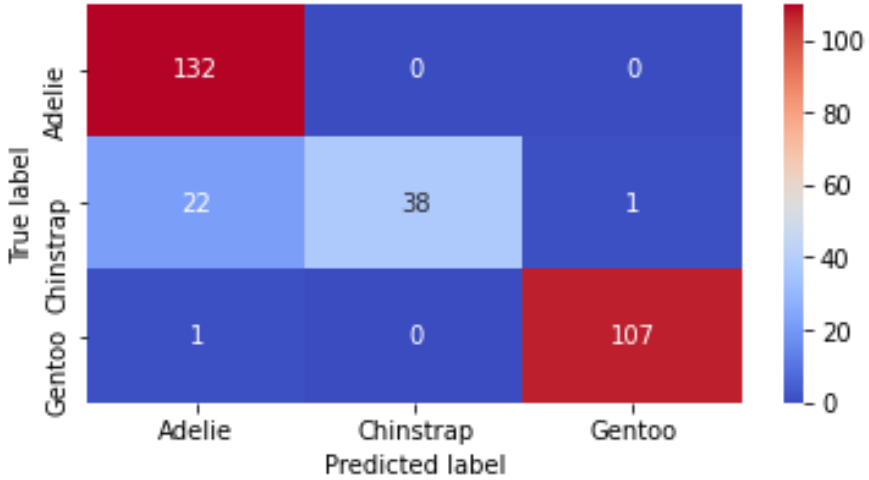
train：test = 1 ： 3
n_estimators = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.94	0.93	0.93	111
Chinstrap	1.00	0.96	0.98	51
Gentoo	0.91	0.94	0.93	89
accuracy			0.94	251
macro avg	0.95	0.94	0.95	251
weighted avg	0.94	0.94	0.94	251

GradientBoostingClassifier

train：test = 1 ： 9
n_estimators = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.85	1.00	0.92	132
Chinstrap	1.00	0.62	0.77	61
Gentoo	0.99	0.99	0.99	108
accuracy			0.92	301
macro avg	0.95	0.87	0.89	301
weighted avg	0.93	0.92	0.91	301

GradientBoostingClassifier

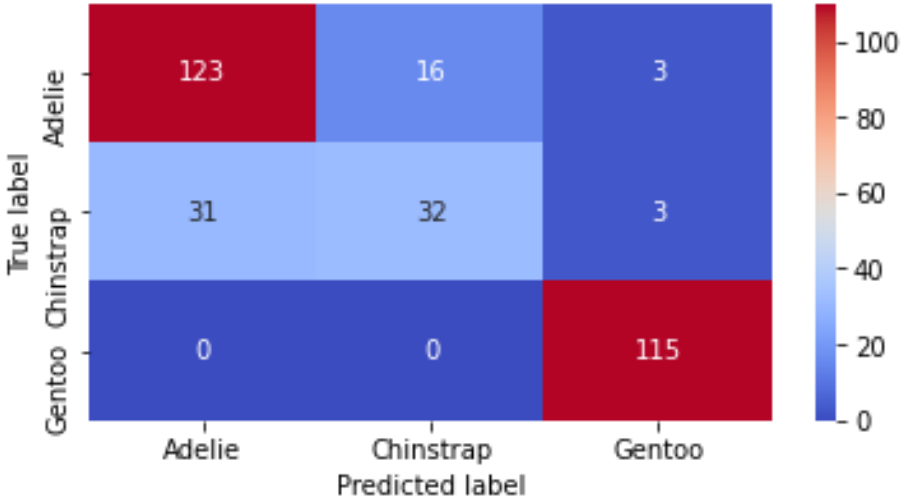
train只取其中十筆，其餘為test

n_estimators = [8,12,16]

learning_rate = [0.6,0.7,0.8,1]

max_depth = [2,3,4]

GradientBoostingClassifier(n_estimators,learning_rate, max_depth,random_state=100)



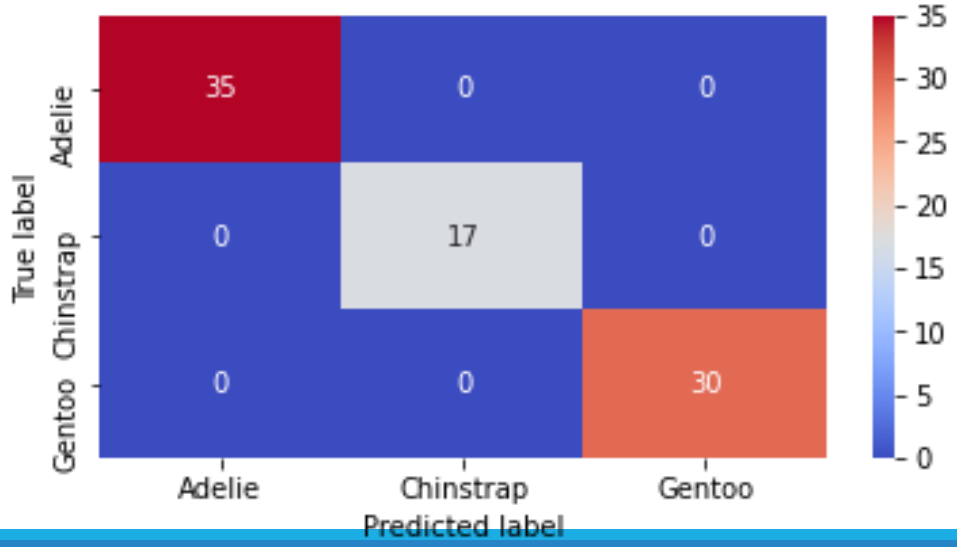
	precision	recall	f1-score	support
Adelie	0.80	0.87	0.83	142
Chinstrap	0.67	0.48	0.56	66
Gentoo	0.95	1.00	0.97	115
accuracy			0.84	323
macro avg	0.81	0.78	0.79	323
weighted avg	0.83	0.84	0.83	323

模型配適和預測

6.HISTGRADIENTBOOSTINGCLASSIFIER

HistGradientBoostingClassifier

train：test = 3 ： 1
max_iter = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
max_iter,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

HistGradientBoostingClassifier

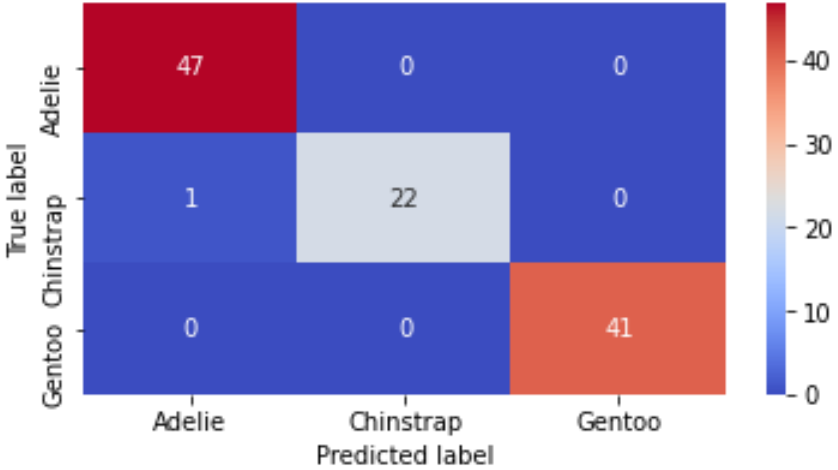
train：test = 2 ： 1

max_iter = [8,12,16]

learning_rate = [0.6,0.7,0.8,1]

max_depth = [2,3,4]

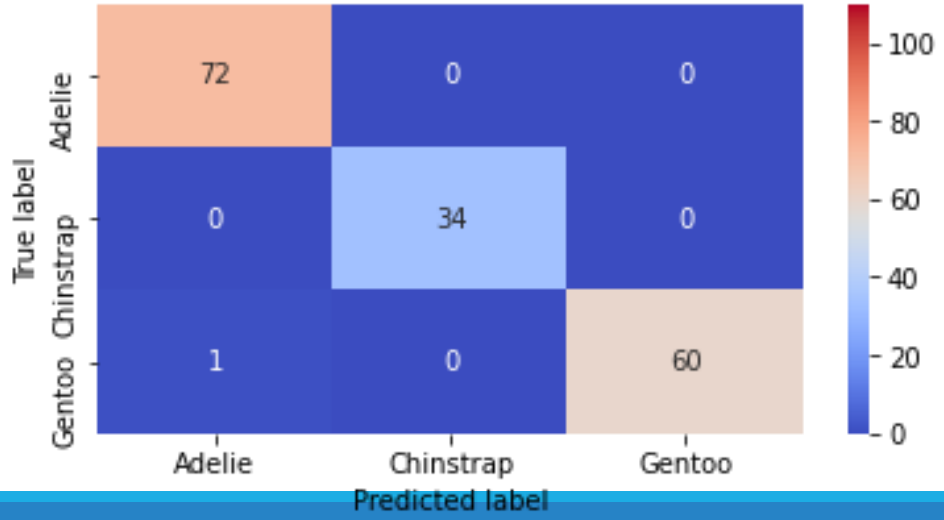
HistGradientBoostingClassifier(max_iter,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

HistGradientBoostingClassifier

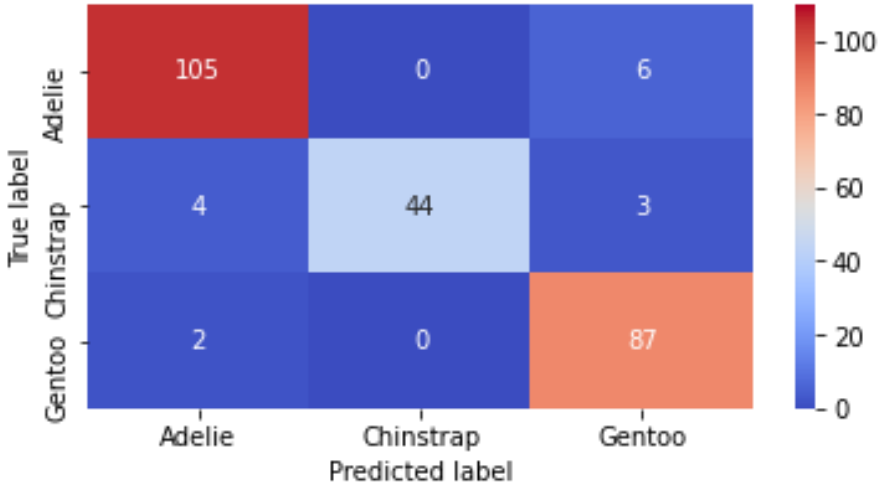
train：test = 1 ： 1
max_iter = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
HistGradientBoostingClassifier(max_iter,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.99	1.00	0.99	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	1.00	0.99	0.99	167
weighted avg	0.99	0.99	0.99	167

HistGradientBoostingClassifier

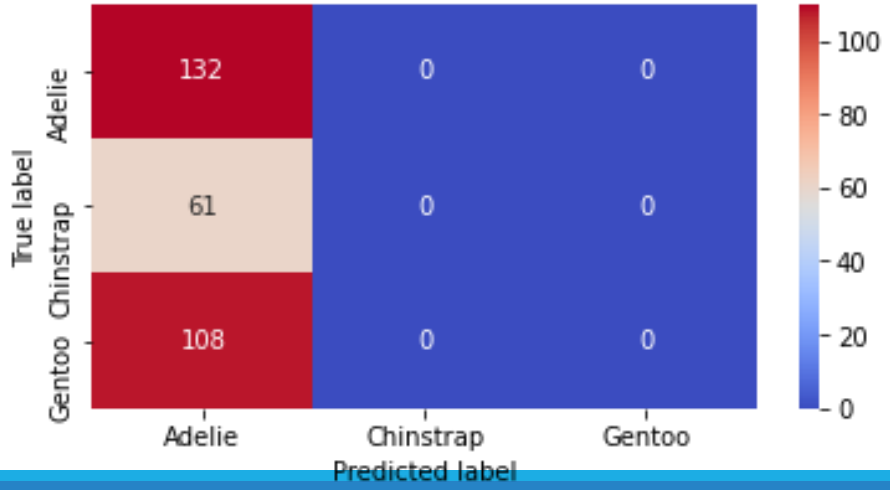
train：test = 1 ： 3
max_iter = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
HistGradientBoostingClassifier(max_iter,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.95	0.95	0.95	111
Chinstrap	1.00	0.86	0.93	51
Gentoo	0.91	0.98	0.94	89
accuracy			0.94	251
macro avg	0.95	0.93	0.94	251
weighted avg	0.94	0.94	0.94	251

HistGradientBoostingClassifier

train：test = 1 ： 9
max_iter = [8,12,16]
learning_rate = [0.6,0.7,0.8,1]
max_depth = [2,3,4]
HistGradientBoostingClassifier(max_iter,learning_rate, max_depth,random_state=100)



	precision	recall	f1-score	support
Adelie	0.44	1.00	0.61	132
Chinstrap	0.00	0.00	0.00	61
Gentoo	0.00	0.00	0.00	108
accuracy			0.44	301
macro avg	0.15	0.33	0.20	301
weighted avg	0.19	0.44	0.27	301

HistGradientBoostingClassifier

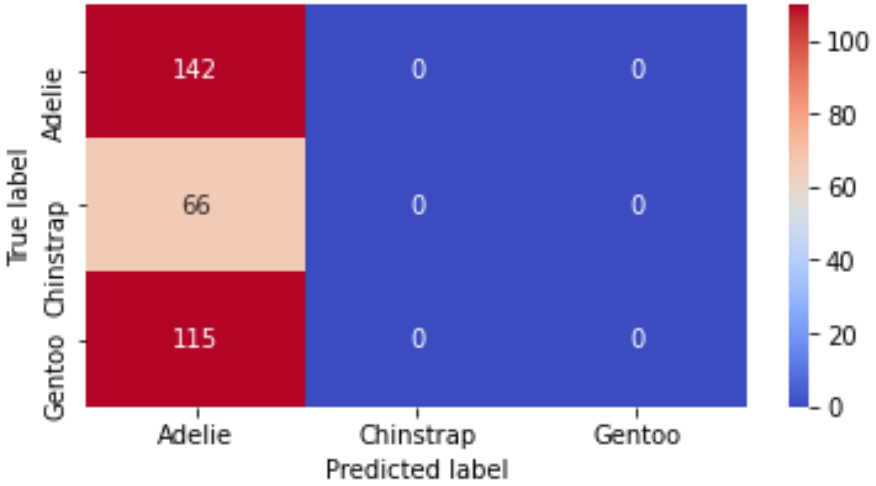
train只取其中十筆，其餘為test

max_iter = [8,12,16]

learning_rate = [0.6,0.7,0.8,1]

max_depth = [2,3,4]

HistGradientBoostingClassifier(max_iter,learning_rate, max_depth,random_state=100)

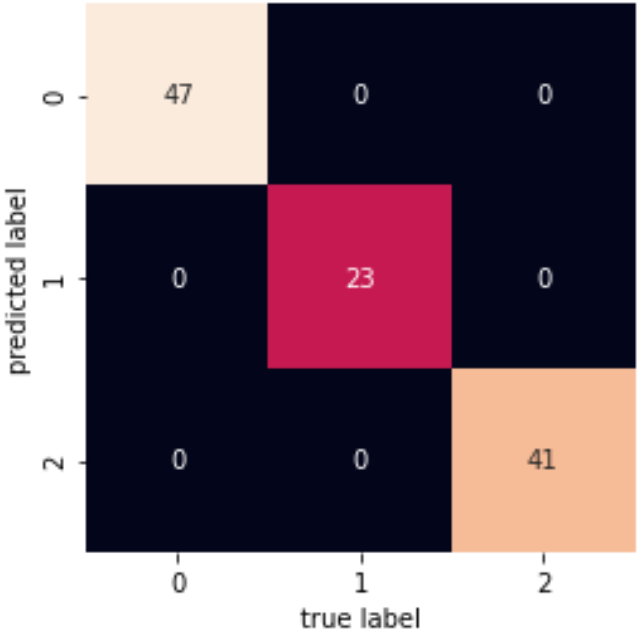


	precision	recall	f1-score	support
Adelie	0.44	1.00	0.61	142
Chinstrap	0.00	0.00	0.00	66
Gentoo	0.00	0.00	0.00	115
accuracy			0.44	323
macro avg	0.15	0.33	0.20	323
weighted avg	0.19	0.44	0.27	323

模型配適和預測

7.KNN

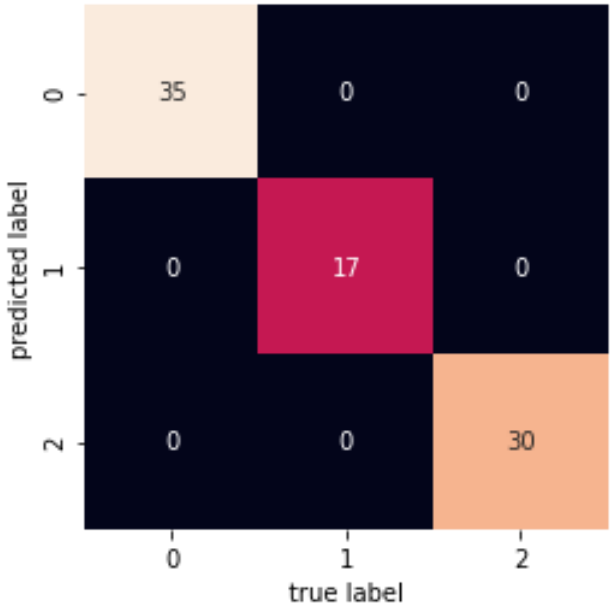
Knn train：test = 2 ： 1



	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	47
Chinstrap	1.00	1.00	1.00	23
Gentoo	1.00	1.00	1.00	41
accuracy			1.00	111
macro avg	1.00	1.00	1.00	111
weighted avg	1.00	1.00	1.00	111

最佳n_neighbors值為：5 accuracy=：1.00

Knn train：test = 3： 1



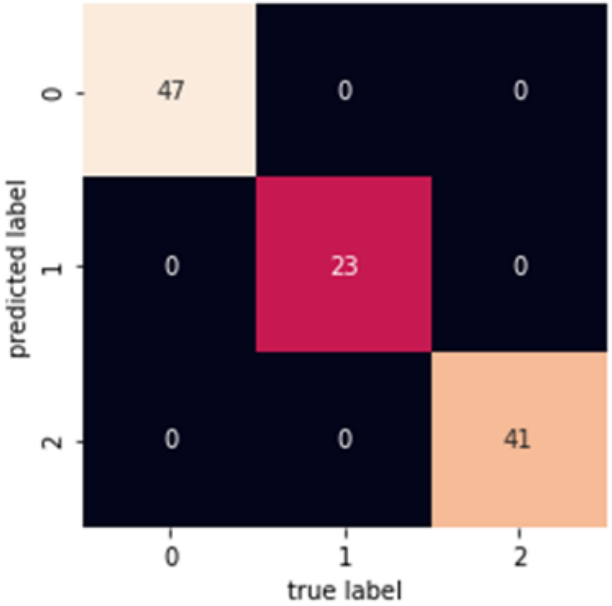
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

最佳n_neighbors值為：5 accuracy=：1.00

模型配適和預測

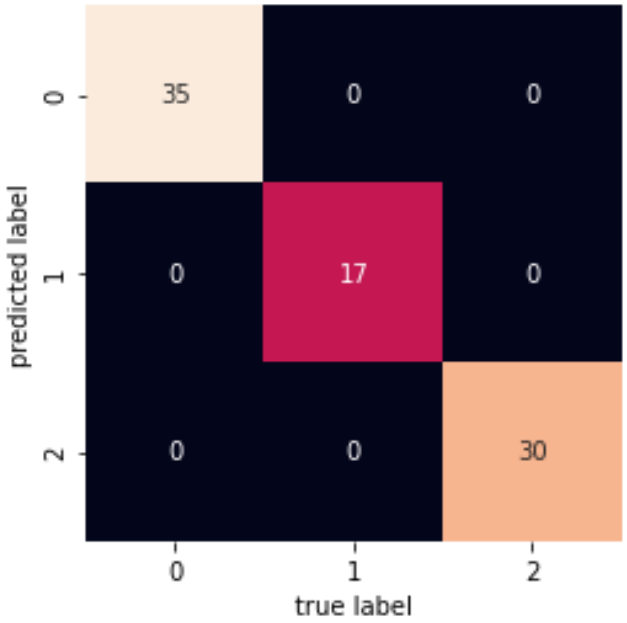
8.SVM

Svm - Linear train：test = 2 ： 1



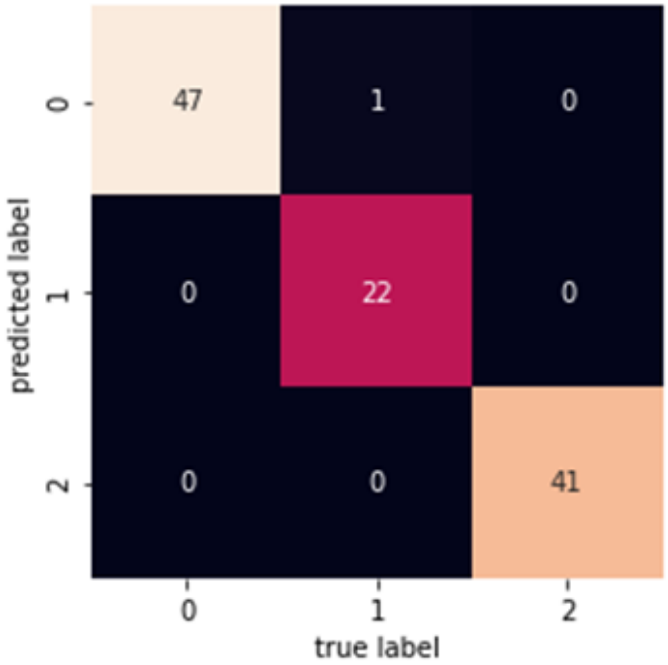
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	47
Chinstrap	1.00	1.00	1.00	23
Gentoo	1.00	1.00	1.00	41
accuracy			1.00	111
macro avg	1.00	1.00	1.00	111
weighted avg	1.00	1.00	1.00	111

Svm - Linear train：test = 3 ： 1



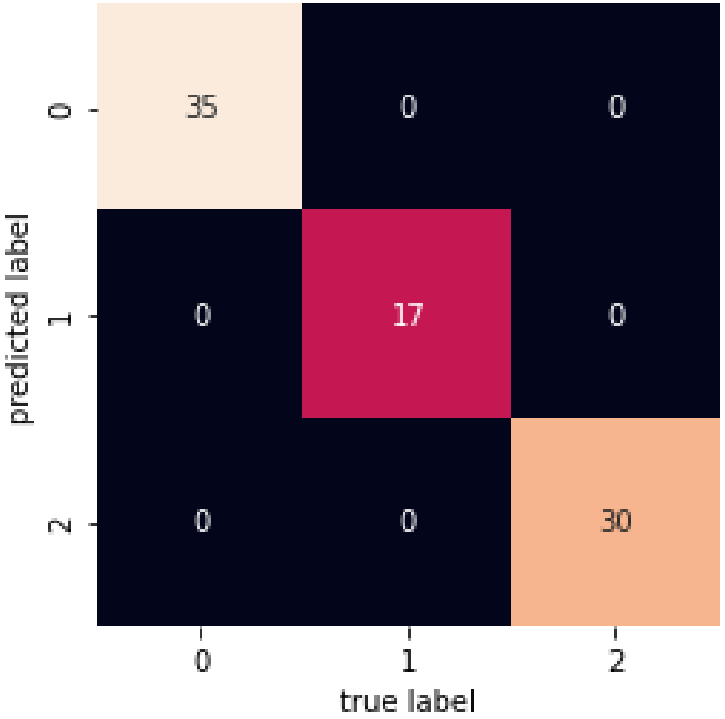
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

Svm - rbf train：test = 2 ： 1



	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

Svm - rbf train：test = 3： 1

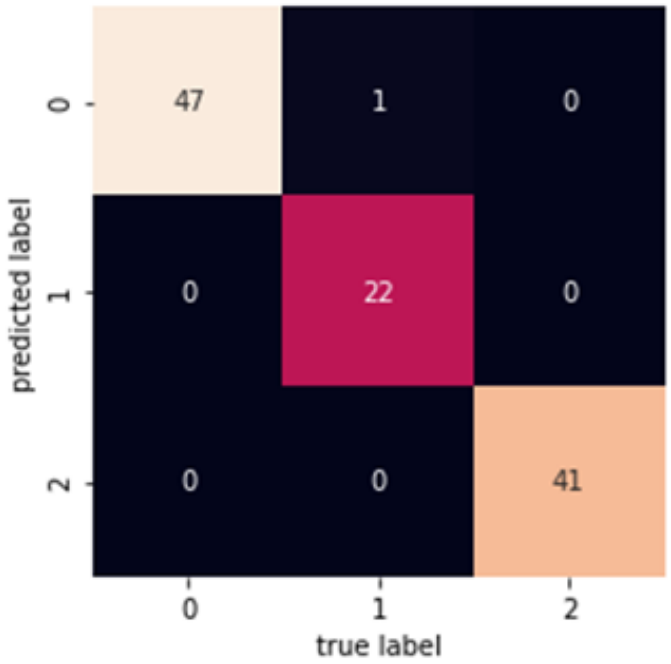


	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

模型配適和預測

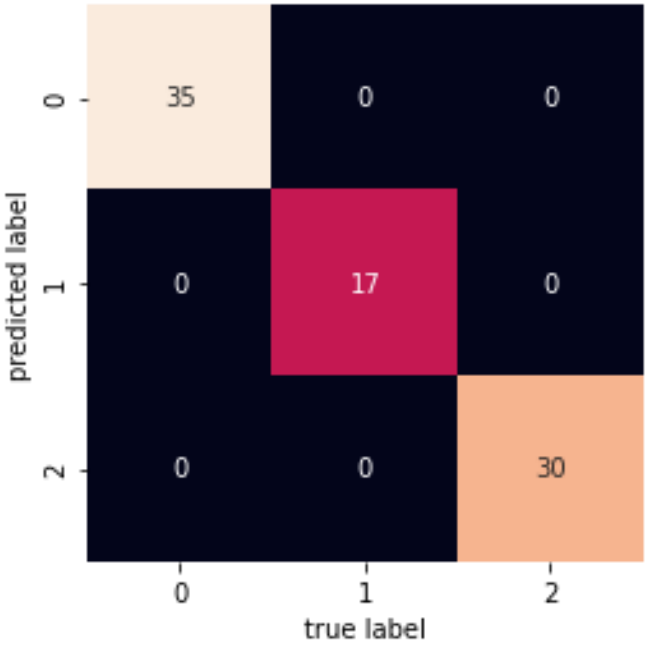
9.LOGISTIC REGRESSION

logistic regression train：test = 2： 1



	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

logistic regression train：test = 3： 1



	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

結論

1. 在畫出敘述統計量的圖表後，我們對於資料的型態有更進一步的了解，讓我們能夠在後續的資料分析上，更能夠知道資料在哪些模型與演算法上，可能將會有比較好的結果。

2. 在試過多種模型後，我們發現大多數的模型在預測訓練資料時，都能有不錯的表現，只有GaussianNB的表現較差。

3. 而在大部分模型，train和test切割成3:1，對訓練資料分類結果較佳。