組別：第二組 成員：H24051184 統計110 馬靖宇 H24089030 統計111 翁瑋廷
H24089014 統計111 許祐誠

1. Brief introduction of the problem

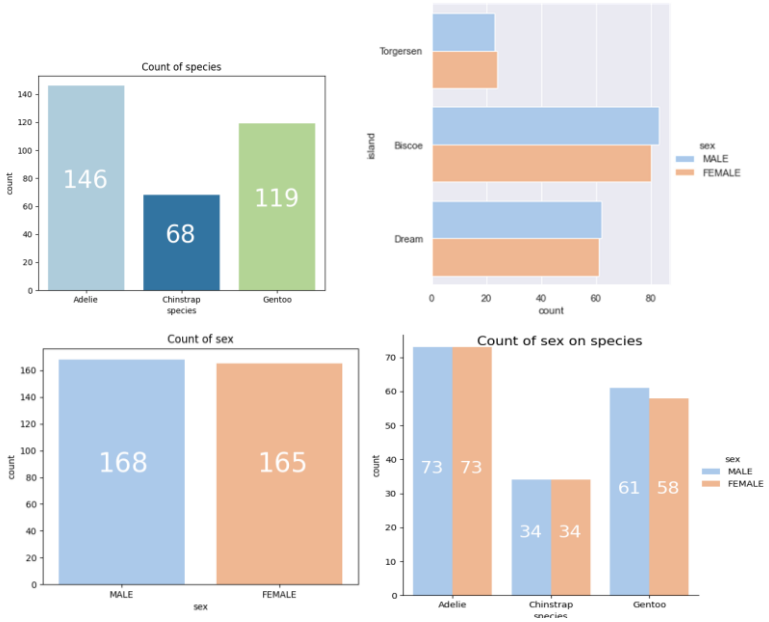
南極洲的企鵝，由於全球暖化使其走向滅絕。我們透過本次資料科學導論報告，深入的研究這些瀕臨滅絕的各種不同的企鵝， 利用「帕爾默群島（南極洲）企鵝數據」，觀察三種不同企鵝的各項數據，透過各種機器學習演算法以及不同的訓練、測試資料切割方式，找出哪些演算法在訓練資料較少時，依舊有好的分類企鵝結果。

2. Data description and preprocessing

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。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。

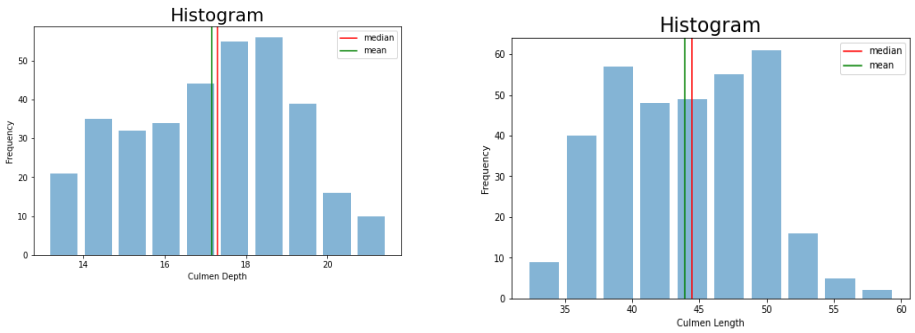
3. Insights discovered from the data

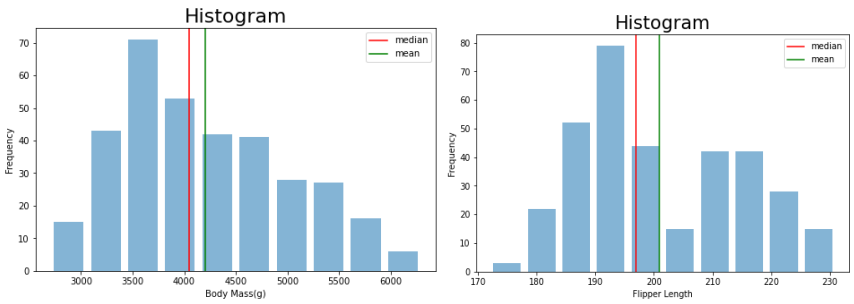
a.



企鵝種類數量以Adelie最多，Gentoo次之，Chinstrap最少；性別的數量、種類對性別的數量、島嶼對性別的數量都相當平均

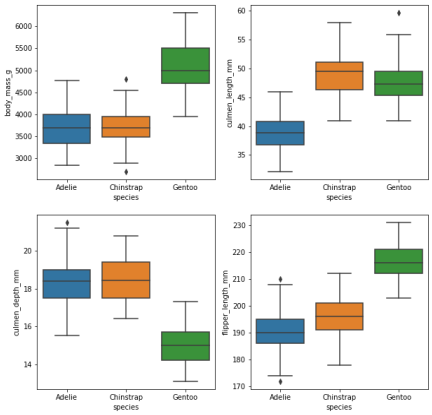
b.





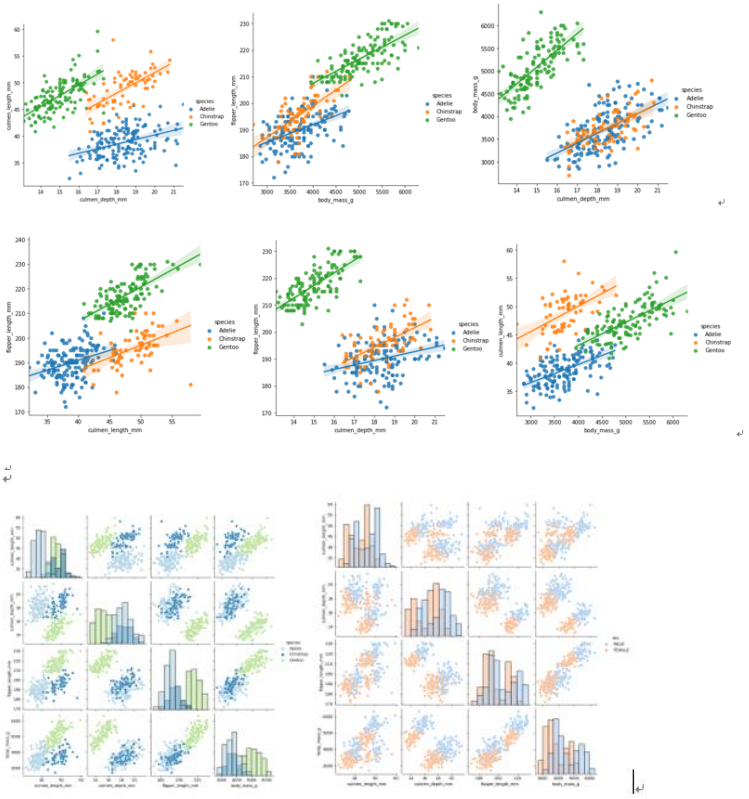
體重和flipper_length的分配較為右偏，鳥嘴長度的分配有一點雙峰的感覺

c.



Gentoo的體重較重、culmen_depth較厚且flipper_length較長；Adelie的culmen_length較短

d.



從散佈圖中可以看出在兩兩不同特徵下，可以將三種企鵝大致區分開來(在分群上能有所幫助)，性別則較無法區分

4. Methodology details

a. 模型建置

i. Decision tree：

想比較不同的criterion，information gain 和 gini index，在分類資料上的差別。 並指定 random state 固定住模型。

第一個模型超參數：random state = 2

第二個模型超參數：random state = 2 , criterion = "entropy"

ii. Random forest：

根據Decision tree 的結果，使用不同的criterion 並沒有對分類結果有差別。所以我random forest的criterion參數只用entropy。max_feature代表隨機森林內單顆決策樹使用特徵的最大數量。max_samples代表隨機森林內單顆決策樹所使用的訓練資料數。如果max_sample是介於0，1之間的浮點數，單顆決策樹所使用的訓練資料數為max_sample * 訓練資料總數。

第一個模型超參數：criterion="entropy" , random_state = 100 , max_samples = 0.7 , max_features = 3

iii. BaggingClassifier :

調整的重要參數有n_estimators最大迭代次數、max_samples抽取訓練每個基本估計量的樣本數、max_features訓練每個基本估計量的要素數量，參數值使用以下所有的組合進行比較 : 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]

iv. GaussianNB :

沒有重要參數，所以僅以GaussianNB()進行資料訓練和預測

v. GradientBoostingClassifier :

調整的重要參數有n_estimators最大迭代次數、learning_rate權重縮減係、max_depth樹的最大深度，參數值使用以下所有的組合進行比較 :

n_estimators = [8,12,16]；learning_rate = [0.6,0.7,0.8,1]；max_depth = [2,3,4]

vi. HistGradientBoostingClassifier :

調整的重要參數有max_iter算法收斂的最大迭代次數、learning_rate權重縮減係數、max_depth樹的最大深度，參數值使用以下所有的組合進行比較 :

max_iter = [8,12,16]；learning_rate = [0.6,0.7,0.8,1]；max_depth = [2,3,4]

vii. Knn

在使用KNN演算法，我們僅調整n_neighbors，而我們將利用gridsearchcv來選擇適合的n值，範圍為 [1,2,3,4,5,6]

viii. svm

使用svm的模型，調整的參數為kernel，kernel有兩種，一個為'linear'，另一個為'rbf'，以此兩種方式進行比較

ix. logistic regressiom

調整的重要參數有收斂最大跌帶次數max_iter=10、正則化係數的倒數c=1

5. Evaluation and Results(Evaluation metrics, Train/Test settings, Baselines. Experimental Results. Insights derived from the results.)

a. 分類指標

i. TP、TN、FP、FN

TP(True Positive) 為預測是1，真實是1的結果；FP(False Positive) 為預測是1，真實是0的結果；TN(True Negative) 為預測是0，真實是0的結果；FN(False Negative) 為預測是0，真實是1的結果。

ii. accuracy

代表正確分類的機率， accuracy = TP + TN / all test data

iii. precision

代表預測是1，真實是1的機率，precision = TP / TP + FP

iv. recall

代表真實是1，預測是1的機率，recall = TP / TP + FN

v. f1-score

代表precision 和 recall 的調和平均數，f1-score = 2×Precision×Recall / Precision + Recall

vi. weighted

將各類別的precision、recall、f1-score直接平均

vii. macro

將各類別的precision、recall、f1-score依照support佔訓練資料比例加權平均

b. 訓練和測試資料切割方式：

i. train : test = 3 : 1

將index的0,4,8…、1,5,9…、2,6,10…..設為訓練資料，index的3,7,11...設為測試資料。

ii. train : test = 2 : 1

將index的0,3,6…和1,4,7…設為訓練資料，index的2,5,8…設為測試資料。

iii. train : test = 1 : 1

將index的0,2,4…..設為訓練資料，index的1,3,5...設為測試資料。

iv. train : test = 1 : 3

將index的3,7,11...設為訓練資料，index的0,4,8…、1,5,9…、2,6,10…..設為測試資料。

v. train : test = 1 : 9

將index的9 19 29.....設為訓練資料，其餘index設為測試資料。

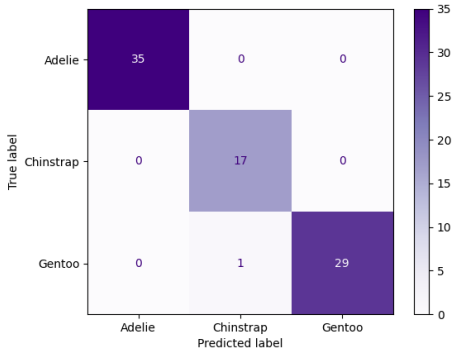
vi. 只取10筆資料訓練

將index的34 66 98 .....等10筆資料設為訓練資料，其餘index設為測試資料。

c. 模型結果

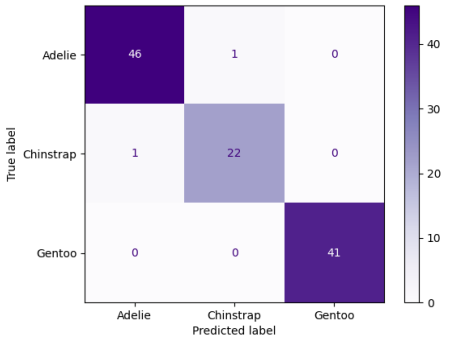
i. Decision tree

1. train : test = 3 : 1 (兩個模型結果相同)

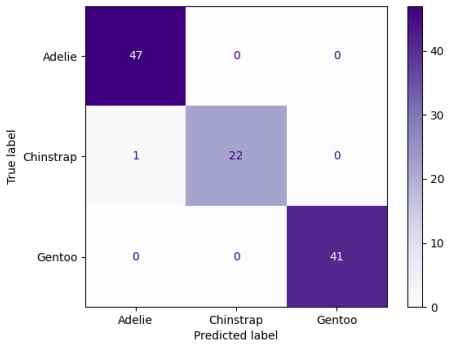


	precision	recall	f1-score	support
Adelie	1	1	1	35
Chinstrap	0.9444	1	0.9714	17
Gentoo	1	0.9667	0.9831	30
accuracy			0.9878	82
macro avg	0.9815	0.9889	0.9848	82
weighted a	0.9885	0.9878	0.9879	82

2. train : test = 2 : 1 (模型0、1)

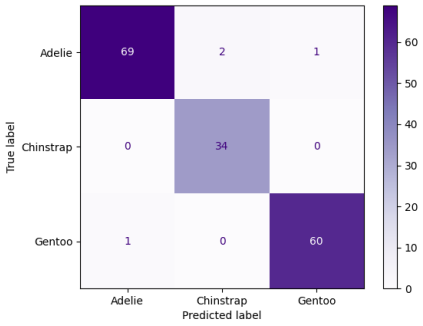


	precision	recall	f1-score	support
Adelie	0.9787	0.9787	0.9787	47
Chinstrap	0.9565	0.9565	0.9565	23
Gentoo	1	1	1	41
accuracy			0.982	111
macro avg	0.9784	0.9784	0.9784	111
weighted a	0.982	0.982	0.982	111

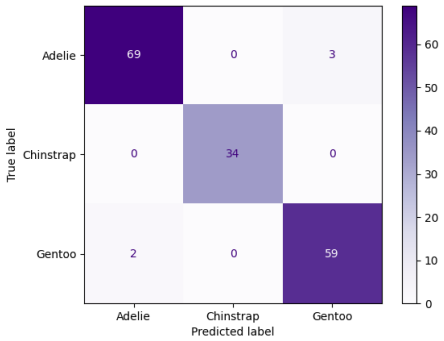


	precision	recall	f1-score	support
Adelie	0.9792	1	0.9895	47
Chinstrap	1	0.9565	0.9778	23
Gentoo	1	1	1	41
accuracy			0.991	111
macro avg	0.9931	0.9855	0.9891	111
weighted a	0.9912	0.991	0.9909	111

3. train : test = 1 : 1(模型0、1)

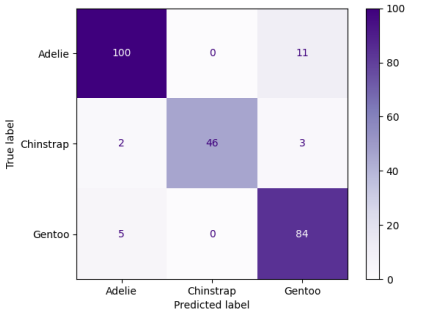


	precision	recall	f1-score	support
Adelie	0.9857	0.9583	0.9718	72
Chinstrap	0.9444	1	0.9714	34
Gentoo	0.9836	0.9836	0.9836	61
accuracy			0.976	167
macro avg	0.9713	0.9806	0.9756	167
weighted a	0.9765	0.976	0.9761	167



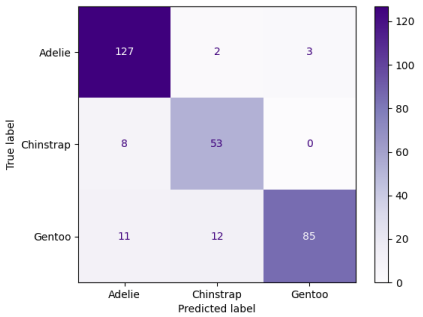
	precision	recall	f1-score	support
Adelie	0.9718	0.9583	0.965	72
Chinstrap	1	1	1	34
Gentoo	0.9516	0.9672	0.9593	61
accuracy			0.9701	167
macro avg	0.9745	0.9752	0.9748	167
weighted a	0.9702	0.9701	0.9701	167

4. train : test = 1 : 3 (兩個模型結果相同)



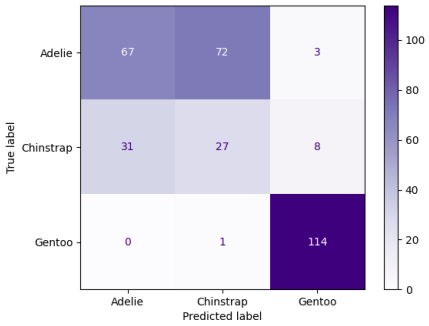
	precision	recall	f1-score	support
Adelie	0.9346	0.9009	0.9174	111
Chinstrap	1	0.902	0.9485	51
Gentoo	0.8571	0.9438	0.8984	89
accuracy			0.9163	251
macro avg	0.9306	0.9156	0.9214	251
weighted a	0.9204	0.9163	0.917	251

5. train : test = 1 : 9 (兩個模型結果相同)



	precision	recall	f1-score	support
Adelie	0.8699	0.9621	0.9137	132
Chinstrap	0.791	0.8689	0.8281	61
Gentoo	0.9659	0.787	0.8673	108
accuracy			0.8804	301
macro avg	0.8756	0.8727	0.8697	301
weighted a	0.8884	0.8804	0.8797	301

6. 只取10筆資料訓練(兩個模型結果相同)



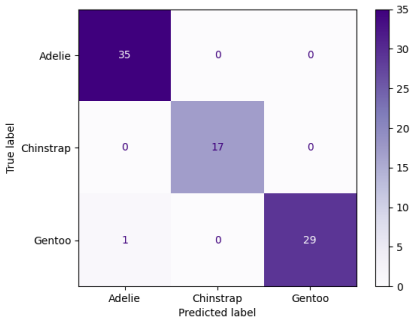
	precision	recall	f1-score	support
Adelie	0.6837	0.4718	0.5583	142
Chinstrap	0.27	0.4091	0.3253	66
Gentoo	0.912	0.9913	0.95	115
accuracy			0.644	323
macro avg	0.6219	0.6241	0.6112	323
weighted a	0.6804	0.644	0.6502	323

7. 小結：

觀察訓練、測試資料三比一、二比一和一比一的模型結果，可以發現各類別的僅有不超過五筆資料分錯。所有分類指標都還能保持在0.9以上。而訓練、測試資料一比三、一比九的模型結果逐漸變差，分類指標數值逐漸下降。只取10筆資料的模型結果相當差，準確率僅有0.644，其他分類指標也表現很差，因此決策樹模型並不適合此情況。

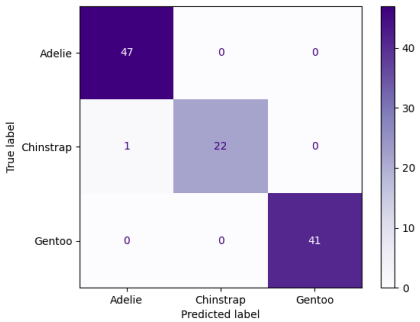
ii. Random forest

1. train : test = 3 : 1



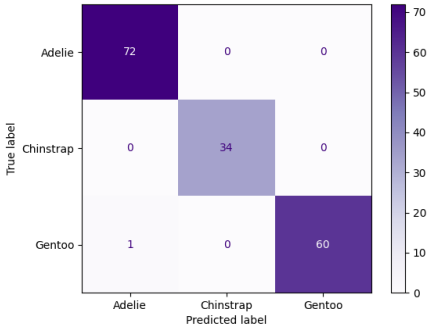
	precision	recall	f1-score	support
Adelie	0.9722	1	0.9859	35
Chinstrap	1	1	1	17
Gentoo	1	0.9667	0.9831	30
accuracy			0.9878	82
macro avg	0.9907	0.9889	0.9897	82
weighted a	0.9881	0.9878	0.9878	82

2. train : test = 2 : 1



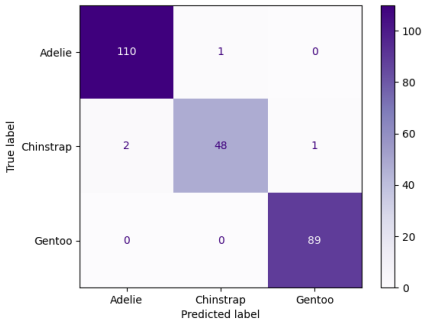
	precision	recall	f1-score	support
Adelie	0.9792	1	0.9895	47
Chinstrap	1	0.9565	0.9778	23
Gentoo	1	1	1	41
accuracy			0.991	111
macro avg	0.9931	0.9855	0.9891	111
weighted a	0.9912	0.991	0.9909	111

3. train : test = 1 : 1



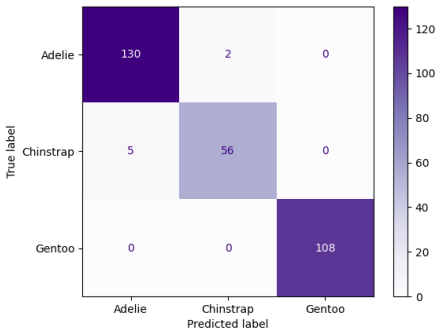
	precision	recall	f1-score	support
Adelie	0.9863	1	0.9931	72
Chinstrap	1	1	1	34
Gentoo	1	0.9836	0.9917	61
accuracy			0.994	167
macro avg	0.9954	0.9945	0.9949	167
weighted a	0.9941	0.994	0.994	167

4. train : test = 1 : 3



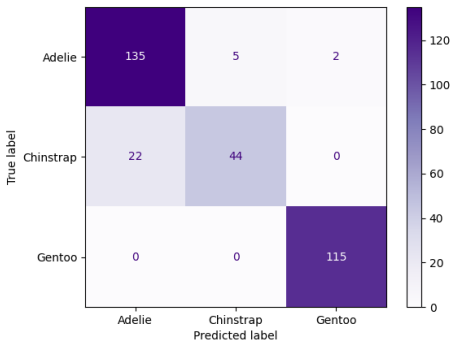
	precision	recall	f1-score	support
Adelie	0.9821	0.991	0.9865	111
Chinstrap	0.9796	0.9412	0.96	51
Gentoo	0.9889	1	0.9944	89
accuracy			0.9841	251
macro avg	0.9835	0.9774	0.9803	251
weighted a	0.984	0.9841	0.9839	251

5. train : test = 1 : 9



	precision	recall	f1-score	support
Adelie	0.963	0.9848	0.9738	132
Chinstrap	0.9655	0.918	0.9412	61
Gentoo	1	1	1	108
accuracy			0.9767	301
macro avg	0.9762	0.9676	0.9717	301
weighted a	0.9768	0.9767	0.9766	301

6. 只取10筆資料訓練



	precision	recall	f1-score	support
Adelie	0.8599	0.9507	0.903	142
Chinstrap	0.898	0.6667	0.7652	66
Gentoo	0.9829	1	0.9914	115
accuracy			0.9102	323
macro avg	0.9136	0.8725	0.8865	323
weighted a	0.9115	0.9102	0.9063	323

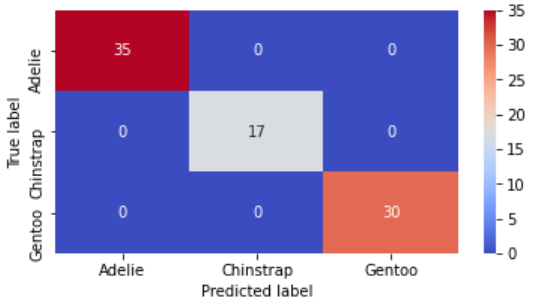
7. 小結：

觀察訓練、測試資料三比一、二比一、一比一、一比三和一比九的模型結果，可以發現各類別的僅有不超過五筆資料分錯。所有分類指標都還能保持在0.9以上。只取10筆資料的模型結果也還不錯，準確率仍有0.9102，Adelie和Gentoo的其他分

類指標也表現不錯。僅有真實是Chrinstrap的企鵝比較容易被預測成Adelie企鵝。隨機森林模型相較於決策樹模型適合此情況。

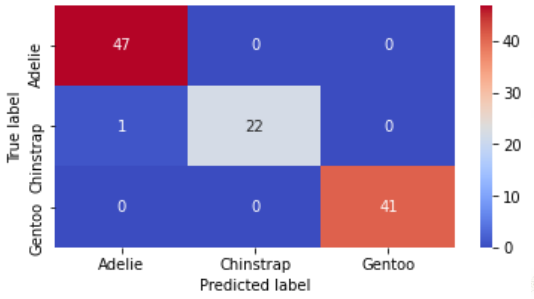
iii. BaggingClassifier

1. train : test = 3 : 1



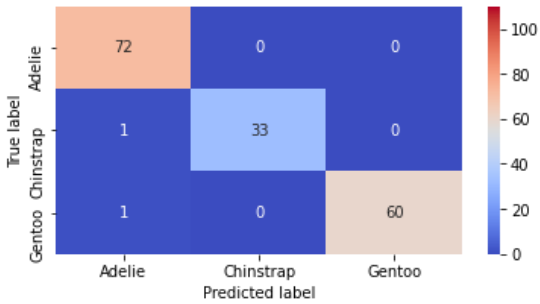
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

2. train : test = 2 : 1



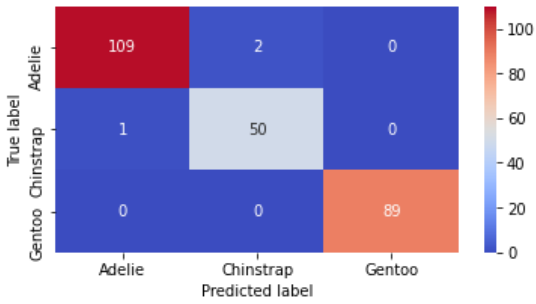
	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

3. train : test = 1 : 1



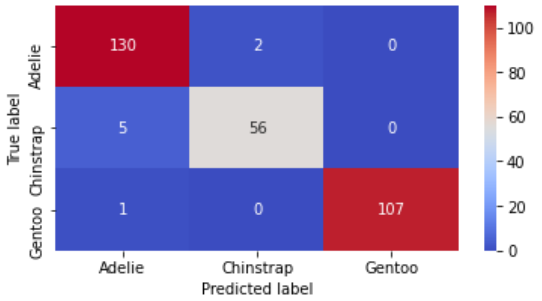
	precision	recall	f1-score	support
Adelie	0.97	1.00	0.99	72
Chinstrap	1.00	0.97	0.99	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	0.99	0.98	0.99	167
weighted avg	0.99	0.99	0.99	167

4. train : test = 1 : 3



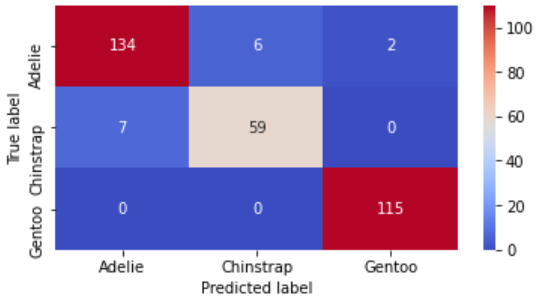
	precision	recall	f1-score	support
Adelie	0.99	0.98	0.99	111
Chinstrap	0.96	0.98	0.97	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.99	251
macro avg	0.98	0.99	0.99	251
weighted avg	0.99	0.99	0.99	251

5. train : test = 1 : 9



	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

6. 只取10筆資料訓練



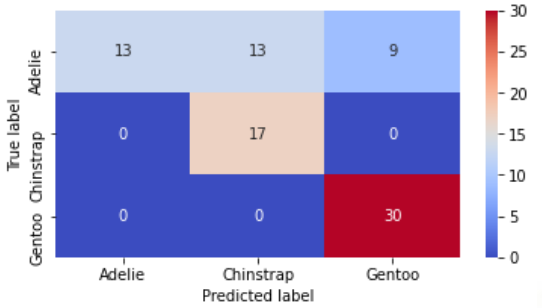
	precision	recall	f1-score	support
Adelie	0.95	0.94	0.95	142
Chinstrap	0.91	0.89	0.90	66
Gentoo	0.98	1.00	0.99	115
accuracy			0.95	323
macro avg	0.95	0.95	0.95	323
weighted avg	0.95	0.95	0.95	323

7. 小結：

在這六種比例的訓練資料下，都能有不錯的表現，三種企鵝的precision、recall和f1-score幾乎都能大於0.90，在train取其中十筆的情況下，accuracy仍可以高達0.95，只有8隻原本是屬於Adelie的企鵝和7隻原本是屬於Chinstrap的企鵝被判錯，所以BaggingClassifier非常適合用來訓練和預測企鵝的種類。

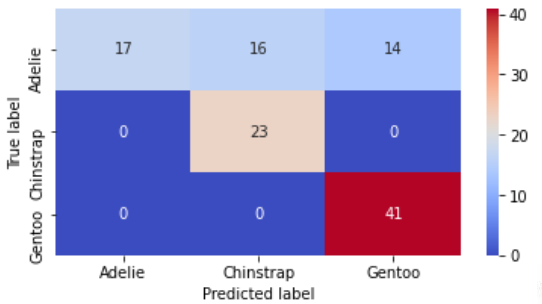
iv. GaussianNB

1. train : test = 3 : 1



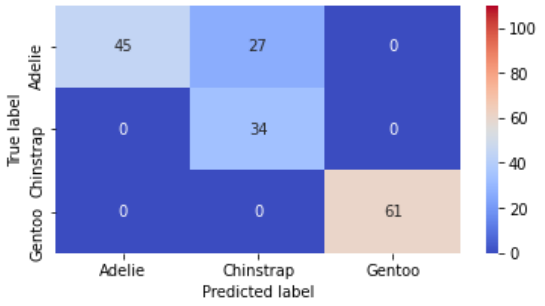
	precision	recall	f1-score	support
Adelie	1.00	0.37	0.54	35
Chinstrap	0.57	1.00	0.72	17
Gentoo	0.77	1.00	0.87	30
accuracy			0.73	82
macro avg	0.78	0.79	0.71	82
weighted avg	0.83	0.73	0.70	82

2. train : test = 2 : 1



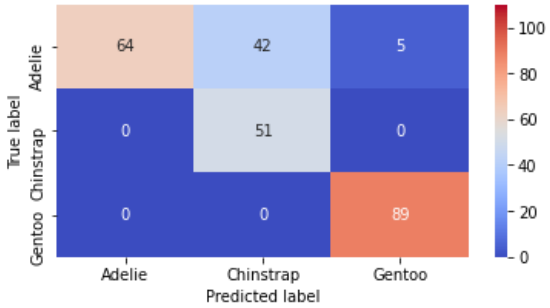
	precision	recall	f1-score	support
Adelie	1.00	0.36	0.53	47
Chinstrap	0.59	1.00	0.74	23
Gentoo	0.75	1.00	0.85	41
accuracy			0.73	111
macro avg	0.78	0.79	0.71	111
weighted avg	0.82	0.73	0.69	111

3. train : test = 1 : 1



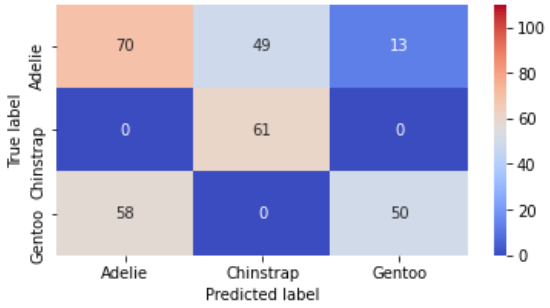
	precision	recall	f1-score	support
Adelie	1.00	0.62	0.77	72
Chinstrap	0.56	1.00	0.72	34
Gentoo	1.00	1.00	1.00	61
accuracy			0.84	167
macro avg	0.85	0.88	0.83	167
weighted avg	0.91	0.84	0.84	167

4. train : test = 1 : 3



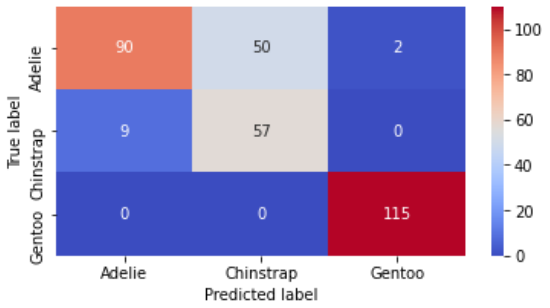
	precision	recall	f1-score	support
Adelie	1.00	0.58	0.73	111
Chinstrap	0.55	1.00	0.71	51
Gentoo	0.95	1.00	0.97	89
accuracy			0.81	251
macro avg	0.83	0.86	0.80	251
weighted avg	0.89	0.81	0.81	251

5. train : test = 1 : 9



	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

6. 只取10筆資料訓練



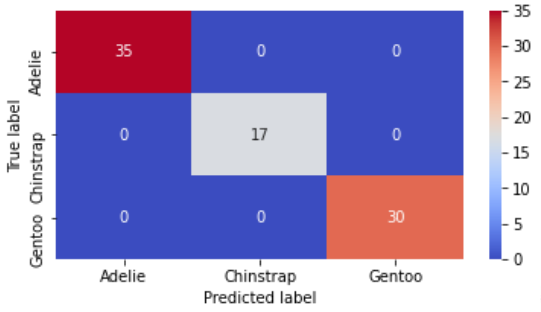
	precision	recall	f1-score	support
Adelie	0.91	0.63	0.75	142
Chinstrap	0.53	0.86	0.66	66
Gentoo	0.98	1.00	0.99	115
accuracy			0.81	323
macro avg	0.81	0.83	0.80	323
weighted avg	0.86	0.81	0.82	323

7. 小結 :

GaussianNB與其他演算法相較之下表現最差，其中又以train : test = 1 : 9的表現最差，accuracy只有0.60，與其他訓練比例相比，多了許多原本是屬於Gentoo的企鵝被判錯，其表現比train只取其中十筆的結果還差，主要是因為在train : test = 1 : 9所訓練的資料，在GaussianNB模型中較無法正確預估其他隻企鵝的種類。

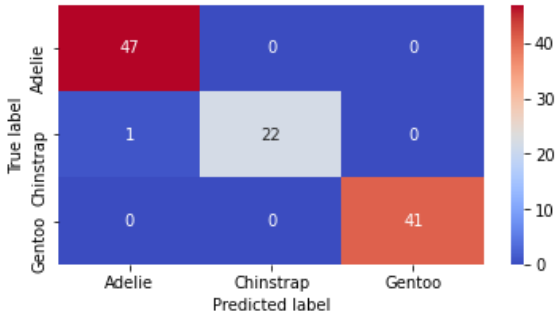
v. GradientBoostingClassifier

1. train : test = 3 : 1



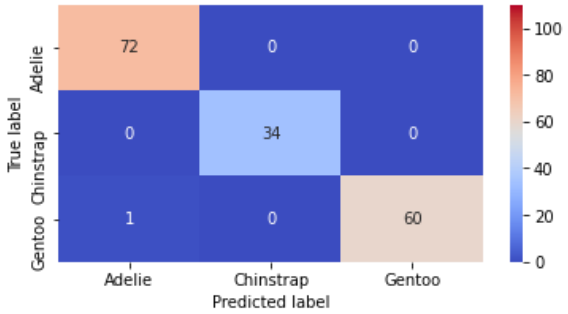
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

2. train : test = 2 : 1



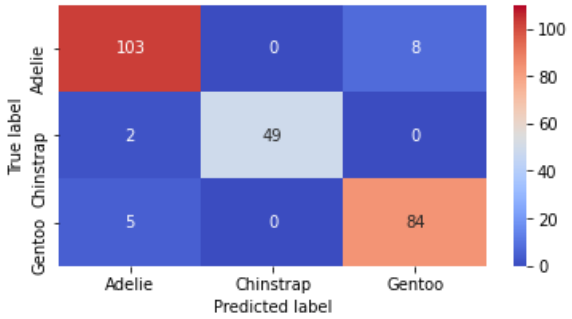
	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

3. train : test = 1 : 1



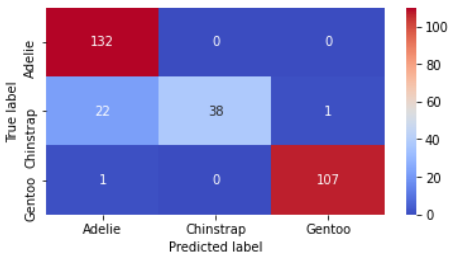
	precision	recall	f1-score	support
Adelie	0.99	1.00	0.99	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	1.00	0.99	0.99	167
weighted avg	0.99	0.99	0.99	167

4. train : test = 1 : 3



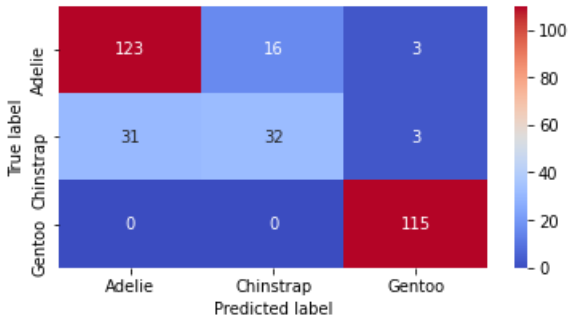
	precision	recall	f1-score	support
Adelie	0.94	0.93	0.93	111
Chinstrap	1.00	0.96	0.98	51
Gentoo	0.91	0.94	0.93	89
accuracy			0.94	251
macro avg	0.95	0.94	0.95	251
weighted avg	0.94	0.94	0.94	251

5. train : test = 1 : 9



	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

6. 只取10筆資料訓練



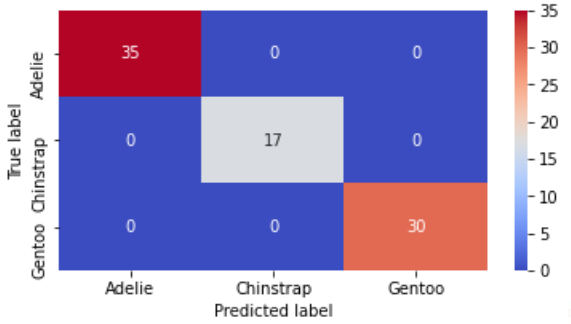
	precision	recall	f1-score	support
Adelie	0.80	0.87	0.83	142
Chinstrap	0.67	0.48	0.56	66
Gentoo	0.95	1.00	0.97	115
accuracy			0.84	323
macro avg	0.81	0.78	0.79	323
weighted avg	0.83	0.84	0.83	323

7. 小結：

GradientBoostingClassifier只有在train只取其中十筆的情況下表現較差，accuracy降至0.84，有19隻原本是屬於Adelie的企鵝和34隻原本是屬於Chinstrap的企鵝被判錯，但只有訓練十筆資料能有這樣的結果已經算很不錯了。

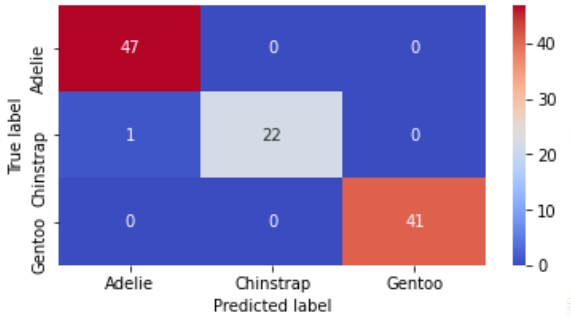
vi. HistGradientBoostingClassifier

1. train : test = 3 : 1



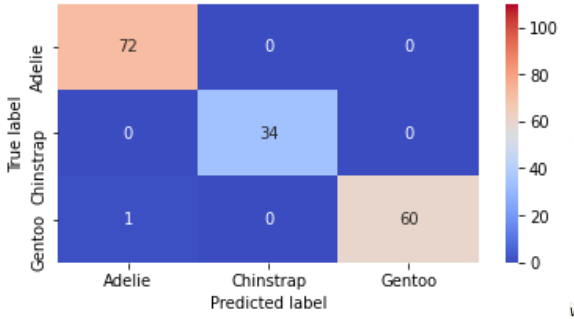
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

2. train : test = 2 : 1



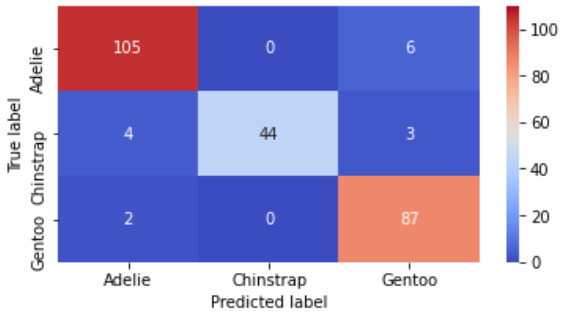
	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

3. train : test = 1 : 1



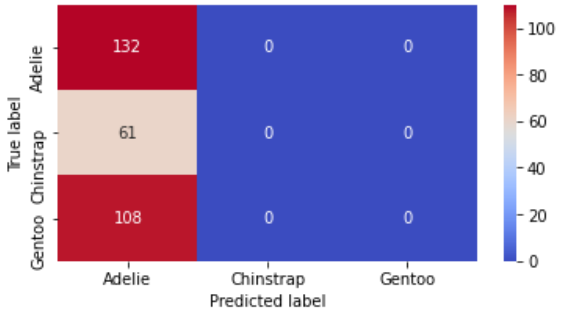
	precision	recall	f1-score	support
Adelie	0.99	1.00	0.99	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	0.98	0.99	61
accuracy			0.99	167
macro avg	1.00	0.99	0.99	167
weighted avg	0.99	0.99	0.99	167

4. train : test = 1 : 3



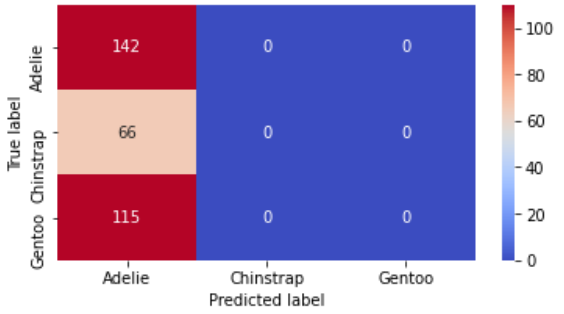
	precision	recall	f1-score	support
Adelie	0.95	0.95	0.95	111
Chinstrap	1.00	0.86	0.93	51
Gentoo	0.91	0.98	0.94	89
accuracy			0.94	251
macro avg	0.95	0.93	0.94	251
weighted avg	0.94	0.94	0.94	251

5. train : test = 1 : 9



	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

6. 只取10筆資料訓練



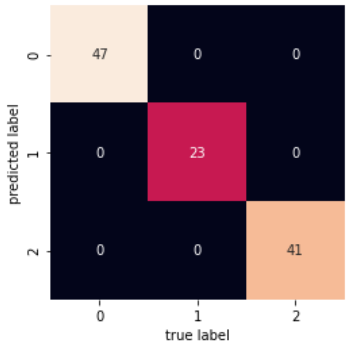
	precision	recall	f1-score	support
Adelie	0.44	1.00	0.61	142
Chinstrap	0.00	0.00	0.00	66
Gentoo	0.00	0.00	0.00	115
accuracy			0.44	323
macro avg	0.15	0.33	0.20	323
weighted avg	0.19	0.44	0.27	323

7. 小結：

HistGradientBoostingClassifier從train : test = 3 : 1到train : test = 1 : 3都能預估得很好，三種企鵝的precision、recall和f1-score幾乎都能大於0.90，但在train : test = 1 : 9和train只取其中十筆時，預測的效果極差，不論原本是屬於哪種企鵝，都會被判成是Adelie，所以HistGradientBoostingClassifier需要較多筆的訓練資料，才能在訓練和預測企鵝的種類上有不錯的表現。

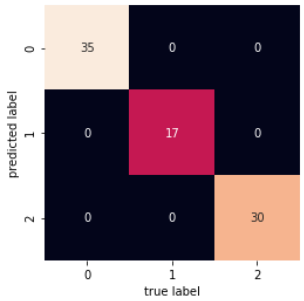
vii. knn

1.train : test = 2 : 1



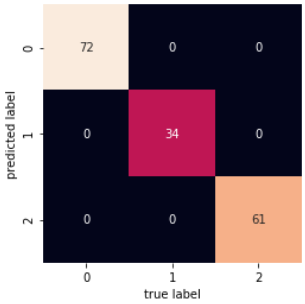
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	47
Chinstrap	1.00	1.00	1.00	23
Gentoo	1.00	1.00	1.00	41
accuracy			1.00	111
macro avg	1.00	1.00	1.00	111
weighted avg	1.00	1.00	1.00	111

2.train : test = 3:1



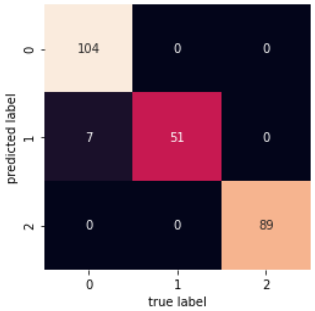
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

3.train : test = 1:1



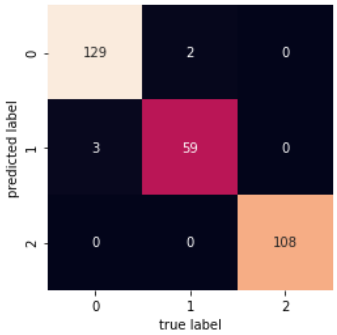
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	1.00	1.00	61
accuracy			1.00	167
macro avg	1.00	1.00	1.00	167
weighted avg	1.00	1.00	1.00	167

4.train : test = 1:3



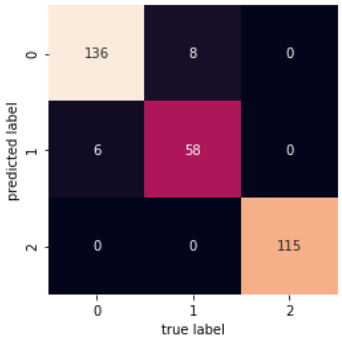
	precision	recall	f1-score	support
Adelie	1.00	0.94	0.97	111
Chinstrap	0.88	1.00	0.94	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.97	251
macro avg	0.96	0.98	0.97	251
weighted avg	0.98	0.97	0.97	251

5.train : test = 1:9



	precision	recall	f1-score	support
Adelie	0.98	0.98	0.98	132
Chinstrap	0.95	0.97	0.96	61
Gentoo	1.00	1.00	1.00	108
accuracy			0.98	301
macro avg	0.98	0.98	0.98	301
weighted avg	0.98	0.98	0.98	301

6.只取10筆資料訓練



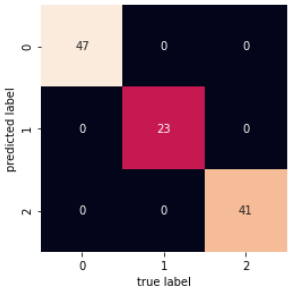
	precision	recall	f1-score	support
Adelie	0.94	0.96	0.95	142
Chinstrap	0.91	0.88	0.89	66
Gentoo	1.00	1.00	1.00	115
accuracy			0.96	323
macro avg	0.95	0.95	0.95	323
weighted avg	0.96	0.96	0.96	323

7.小結

KNN演算法再所有的資料分割的表現上基本上都非常優秀，但再將訓練資料與測試資料為1:9時，由於訓練資料筆數少，因此表現會稍稍的欠佳，會有五隻企鵝的種類被判定錯誤，而當只取10筆訓練資料時，會有14隻企鵝被判定錯誤，預測效果稍稍的較差了些。

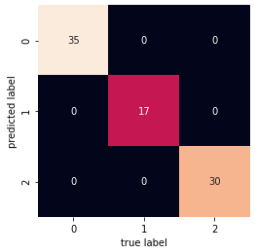
viii. svm(kernal = ‘linear’)

1.train : test = 2 : 1



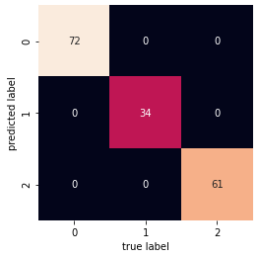
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	47
Chinstrap	1.00	1.00	1.00	23
Gentoo	1.00	1.00	1.00	41
accuracy			1.00	111
macro avg	1.00	1.00	1.00	111
weighted avg	1.00	1.00	1.00	111

2.train : test = 3:1



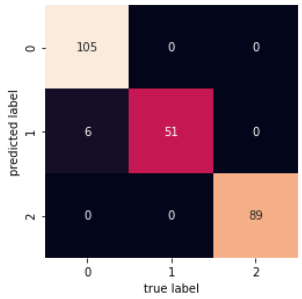
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

3.train : test = 1:1



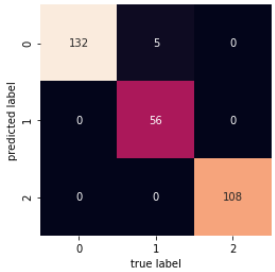
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	1.00	1.00	61
accuracy			1.00	167
macro avg	1.00	1.00	1.00	167
weighted avg	1.00	1.00	1.00	167

4.train : test = 1:3



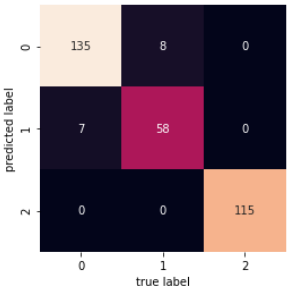
	precision	recall	f1-score	support
Adelie	1.00	0.95	0.97	111
Chinstrap	0.89	1.00	0.94	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.98	251
macro avg	0.96	0.98	0.97	251
weighted avg	0.98	0.98	0.98	251

5.train : test = 1:9



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

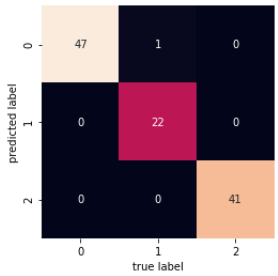
6.只取10筆資料訓練



	precision	recall	f1-score	support
Adelie	0.94	0.95	0.95	142
Chinstrap	0.89	0.88	0.89	66
Gentoo	1.00	1.00	1.00	115
accuracy			0.95	323
macro avg	0.95	0.94	0.94	323
weighted avg	0.95	0.95	0.95	323

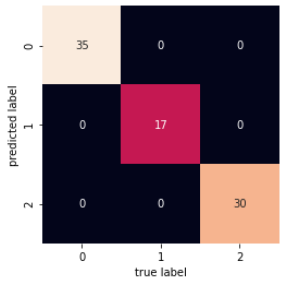
ix. svm(kernal = ‘rbf’)

1.train : test = 2 : 1



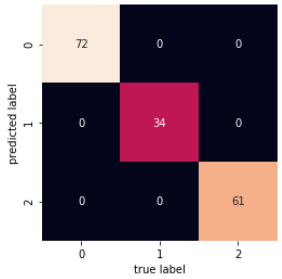
	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

2.train : test = 3:1



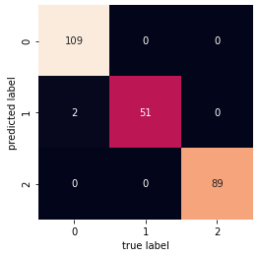
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

3.train : test = 1:1



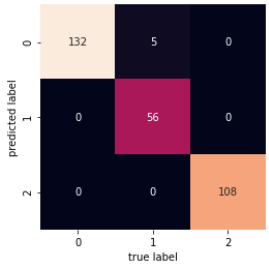
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	1.00	1.00	61
accuracy			1.00	167
macro avg	1.00	1.00	1.00	167
weighted avg	1.00	1.00	1.00	167

4.train : test = 1:3



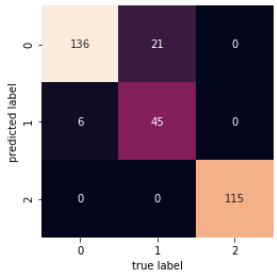
	precision	recall	f1-score	support
Adelie	1.00	0.98	0.99	111
Chinstrap	0.96	1.00	0.98	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.99	251
macro avg	0.99	0.99	0.99	251
weighted avg	0.99	0.99	0.99	251

5.train : test = 1:9



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

6.抽10筆為訓練資料



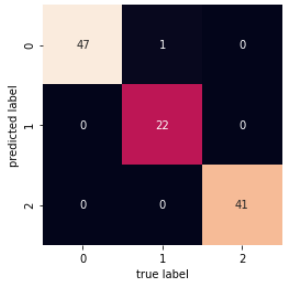
	precision	recall	f1-score	support
Adelie	0.87	0.96	0.91	142
Chinstrap	0.88	0.68	0.77	66
Gentoo	1.00	1.00	1.00	115
accuracy			0.92	323
macro avg	0.92	0.88	0.89	323
weighted avg	0.92	0.92	0.91	323

7. 小結:

SVM演算法再Kernal為'linear'與’ rbf’ 在train : test = 2:1、3:1、1:1、1:3大致上皆表現良好，不過在train : test = 1:9皆會有5隻企鵝被分錯種類，而當我們抽取10筆訓練資料時，會有比較多的企鵝被分錯，而其中linear相較於rbf會是一個較好的kernal。

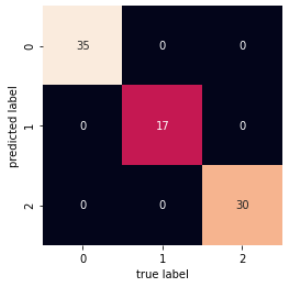
x.logistic regression

1.train : test = 2:1



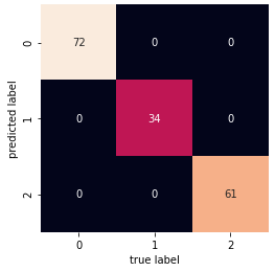
	precision	recall	f1-score	support
Adelie	0.98	1.00	0.99	47
Chinstrap	1.00	0.96	0.98	23
Gentoo	1.00	1.00	1.00	41
accuracy			0.99	111
macro avg	0.99	0.99	0.99	111
weighted avg	0.99	0.99	0.99	111

2.train : test = 3:1



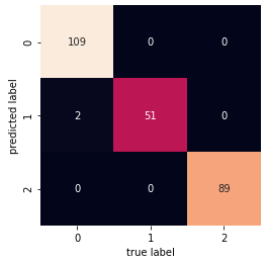
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	35
Chinstrap	1.00	1.00	1.00	17
Gentoo	1.00	1.00	1.00	30
accuracy			1.00	82
macro avg	1.00	1.00	1.00	82
weighted avg	1.00	1.00	1.00	82

3.train : test = 1:1



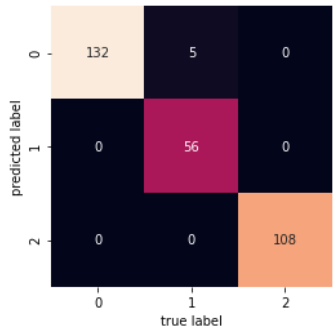
	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	72
Chinstrap	1.00	1.00	1.00	34
Gentoo	1.00	1.00	1.00	61
accuracy			1.00	167
macro avg	1.00	1.00	1.00	167
weighted avg	1.00	1.00	1.00	167

4.train : test = 1:3



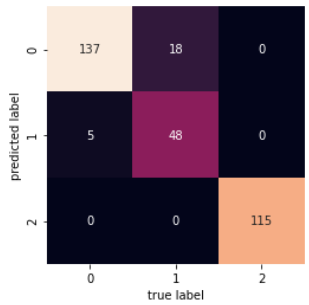
	precision	recall	f1-score	support
Adelie	1.00	0.98	0.99	111
Chinstrap	0.96	1.00	0.98	51
Gentoo	1.00	1.00	1.00	89
accuracy			0.99	251
macro avg	0.99	0.99	0.99	251
weighted avg	0.99	0.99	0.99	251

5.train : test = 1:9



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

6.抽10筆為訓練資料



	precision	recall	f1-score	support
Adelie	0.88	0.96	0.92	142
Chinstrap	0.91	0.73	0.81	66
Gentoo	1.00	1.00	1.00	115
accuracy			0.93	323
macro avg	0.93	0.90	0.91	323
weighted avg	0.93	0.93	0.93	323

7..小結

losgistic regression 模型，在train : test = 2:1、3:1、1:1、1:3大致上皆表現良好，而在train : test = 1:9會有約5隻企鵝被分類錯誤，而當只抽10筆為訓練資料時，更會有約23隻企鵝被分類錯誤，由於用較少訓練資料，因此其表現相較於其他分割方法差了許多。

(5) Conclusions and novelty

- 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。
- 在試過多種模型後，我們發現大多數的模型在訓練、測試資料3:1、2:1、1:1、1:3時，都能有不錯的表現，只有GaussianNB的表現較差。
- HistGradientBoostingClassifier在train、test切割成3:1和train只取其中十筆時，不論原本是屬於哪種企鵝，都會被判成是Adelie，所以此演算法需要較多筆的訓練資料，才能有不錯的表現。
- Random forest、BaggingClassifier和KNN在只取10筆資料訓練時，accuracy都還能高於0.9，所以這些演算法較適合用來訓練小筆的資料。

- 而在大部分模型，train和test切割成3:1，訓練資料分類結果最佳。

(6) The contribution of each team member

1. 馬靖宇：資料描述和前處理，敘述統計量做圖，測試、訓練資料切割，模型指標，決策樹、隨機森林模型建置與結果。
2. 翁瑋廷 : 統計量做圖與解釋，BaggingClassifier模型建構與結果分析、GaussianNB模型建構與結果分析 、GradientBoostingClassifier模型建構與結果分析、HistGradientBoostingClassifier模型建構與結果分析
3. 許祐誠 : k-nearest neighbors模型建構與結果分析、svm(kernal = linear 、 rbf)模型建構與結果分析、logistic regression 模型建構與結果分析。