Machine Learning
期末報告

資管四乙
408402236 施宜彣 408402547 戴婷郁408402602 陳柏光

授課老師 呂奇傑 教授

Contents

1. Variable Definition

Heart Failure Prediction Dataset

Variable (feature) definition

Variable	Definition	Description	Type
Age	個案年齡	age	Numeric
Sex	個案性別	M: male/ F: female	Category
ChestPainType	胸腔疼痛類型	TA: Typical Angina/ ATA: Atypical Angina/ NAP: Non-Anginal Pain/ ASY: Asymptomatic	Category
RestingBP	靜態血壓狀況	Unit: mm/Hg	Numeric
Cholesterol	膽固醇	Unit: mm/dl	Numeric
FastingBS	空腹血糖	1: FastingBS > 120 mg/dl 0: otherwise	Category

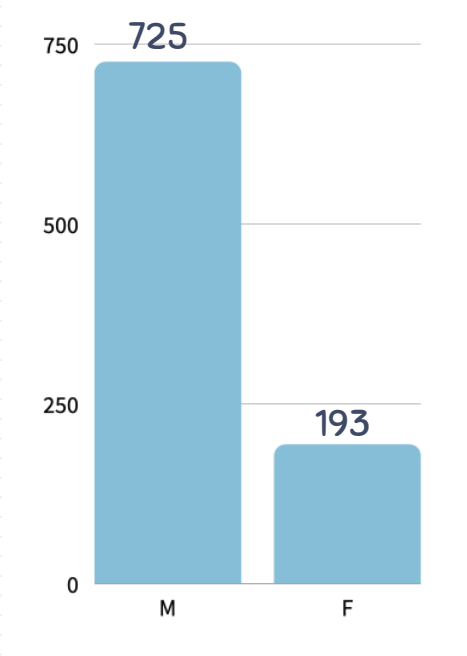
Variable (feature) definition

Variable	Definition	Description	Type
RestingECG	靜態心電圖結果	Normal: normal ST: ST abnormality LVH: 左心室肥大	Category
MaxHR	最大心率	60 – 202 (rate)	Numeric
ExerciseAngina	運動誘發心絞痛	Y: Yes / N: No	Category
Oldpeak	相對於休息來說運動引起的 ST 段抑制	ST (value measured in depression)	Numeric
ST_slope	心電圖 ST 段斜率	Up: upsloping Flat: flat Down: downsloping	Category
⭐ HeartDisease	是否有心臟病	1: Yes/ 0: No	Category

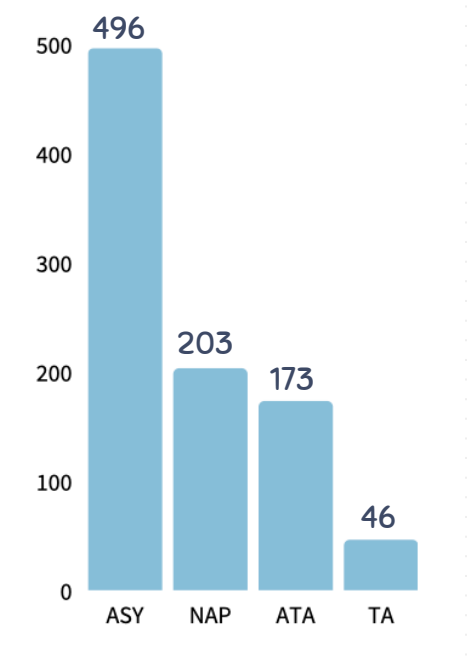
2. Descriptive Statistics

Descriptive Statistics Category

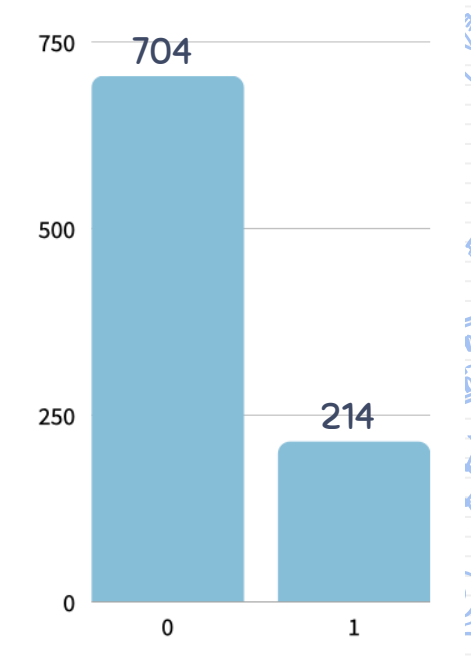
Sex



ChestPainType

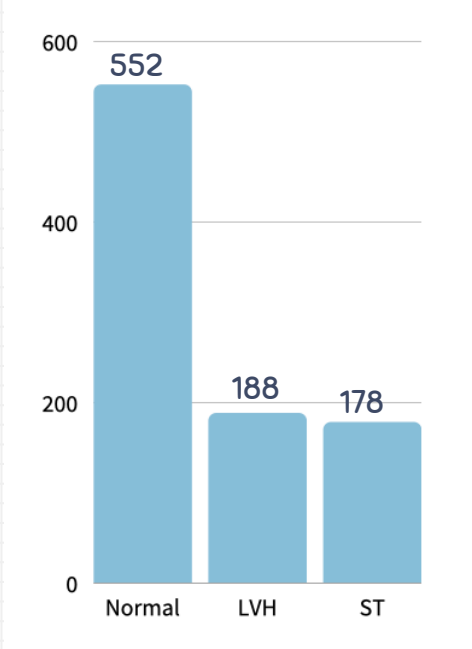


FastingBS

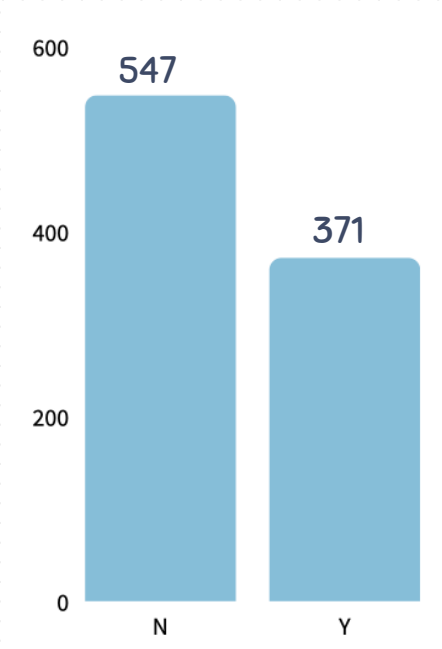


Descriptive Statistics Category

RestingECG

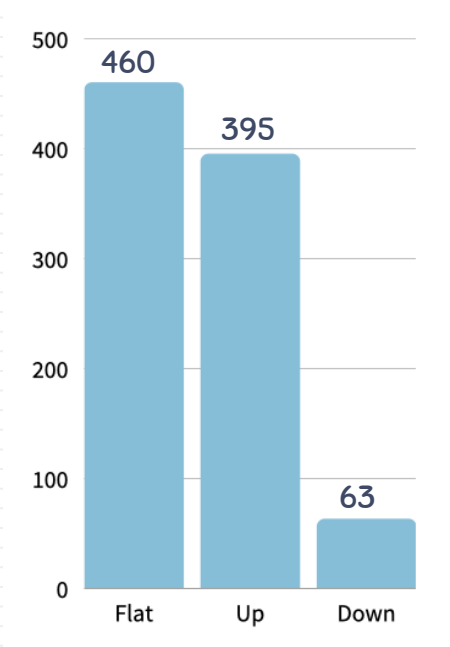


ExerciseAngina

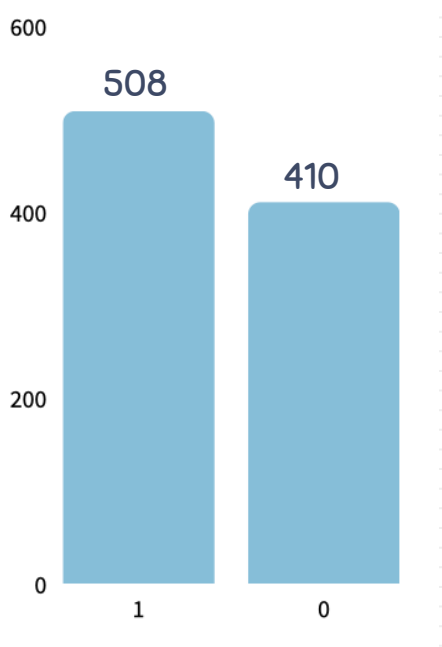


Descriptive Statistics Category

ST_Slope



HeartDisease



Descriptive Statistics Numeric

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
count	918.00	918.00	918.00	918.00	918.00
mean	53.51	132.40	198.80	136.81	0.89
std	9.43	18.51	109.38	25.46	1.07
min	28.00	0.00	0.00	60.00	-2.60
25%	47.00	120.00	173.25	120.00	0.00
50%	54.00	130.00	223.00	138.00	0.60
75%	60.00	140.00	267.00	156.00	1.50
max	77.00	200.00	603.00	202.00	6.20

3. Data Preprocessing

- Change Data Type
- Outlier Deletion
- Drop Missing Value
- Dummies
- Scaling

Data Preprocessing

Change Data Type

Before

#	Column	Non-Null Count	Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	918 non-null	int64
4	Cholesterol	918 non-null	int64
5	FastingBS	918 non-null	int64
6	RestingECG	918 non-null	object
7	MaxHR	918 non-null	int64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	918 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	int64

After

#	Column	Non-Null Count	Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	918 non-null	int64
4	Cholesterol	918 non-null	int64
5	FastingBS	918 non-null	object
6	RestingECG	918 non-null	object
7	MaxHR	918 non-null	int64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	918 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	object

Data Preprocessing

Outlier Deletion

Before

#	Column	Non-Null Count	Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	918 non-null	int64
4	Cholesterol	918 non-null	int64
5	FastingBS	918 non-null	object
6	RestingECG	918 non-null	object
7	MaxHR	918 non-null	int64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	918 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	object

After

#	Column	Non-Null Count	Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	890 non-null	float64
4	Cholesterol	735 non-null	float64
5	FastingBS	918 non-null	object
6	RestingECG	918 non-null	object
7	MaxHR	916 non-null	float64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	902 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	object

Excel 計算出離群值並刪除

Data Preprocessing

Drop Missing Value

每行資料總數剩下 702 筆

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             702 non-null    int64
 1   Sex             702 non-null    object
 2   ChestPainType   702 non-null    object
 3   RestingBP       702 non-null    float64
 4   Cholesterol     702 non-null    float64
 5   FastingBS       702 non-null    object
 6   RestingECG      702 non-null    object
 7   MaxHR           702 non-null    float64
 8   ExerciseAngina  702 non-null    object
 9   Oldpeak         702 non-null    float64
 10  ST_Slope        702 non-null    object
 11  HeartDisease    702 non-null    object
```

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
count	702.00	702.00	702.00	702.00	702.00
mean	52.72	131.56	239.71	140.55	0.83
std	9.54	15.42	50.68	24.36	0.96
min	28.00	92.00	85.00	71.00	-0.10
25%	46.00	120.00	206.00	122.00	0.00
50%	54.00	130.00	235.00	140.00	0.40
75%	59.00	140.00	272.00	160.00	1.50
max	77.00	170.00	404.00	202.00	3.60

```
new_df=df.dropna(axis=0, how='any')
```

Data Preprocessing

Dummies

Dummies	
Sex	Sex_F, Sex_M
ChestPainType	ChestPainType_ASY, ChestPainType_ATA, ChestPainType_NAP, ChestPainType_TA
FastingBS	FastingBS_0, FastingBS_1
RestingECG	RestingECG_LVH, RestingECG_Normal, RestingECG_ST
ExerciseAngina	ExerciseAngina_N, ExerciseAngina_Y
ST_Slope	ST_Slope_Down, ST_Slope_Flat, ST_Slope_Up
HeartDisease	HeartDisease_0, HeartDisease_1

Data Preprocessing

Scaling

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak	Sex_F	Sex_M
0	0.24	0.70	0.48	0.79	0.30	0.0	1.0
1	0.43	0.80	0.30	0.68	0.41	1.0	0.0
2	0.18	0.65	0.47	0.27	0.30	0.0	1.0
3	0.41	0.69	0.35	0.34	0.47	1.0	0.0
4	0.53	0.75	0.32	0.44	0.30	0.0	1.0

5 rows × 23 columns

4. Models

CART / SVM/ Logistic Regression

- Training & Testing: 80/20
- Seeds: 8 (random_state=8)
- Validation: 3-Fold (CV = 3)

CART

CART (Default)			
max_depth	max_leaf_nodes	min_samples_leaf	Average ACC
None	None	1	81.82 +/- 3.41

CART

CART（手動調參）				
max_depth = 5				
Training Average ACC		min_samples_leaf		
		2	3	4
max_leaf_nodes	15	82.71 +/- 1.82	83.07 +/- 2.2	**83.6 +/- 1.76**
	18	82.89 +/- 2.0	82.89 +/- 2.0	83.6 +/- 1.76
	21	82.0 +/- 2.02	81.82 +/- 3.06	82.53 +/- 2.81

CART

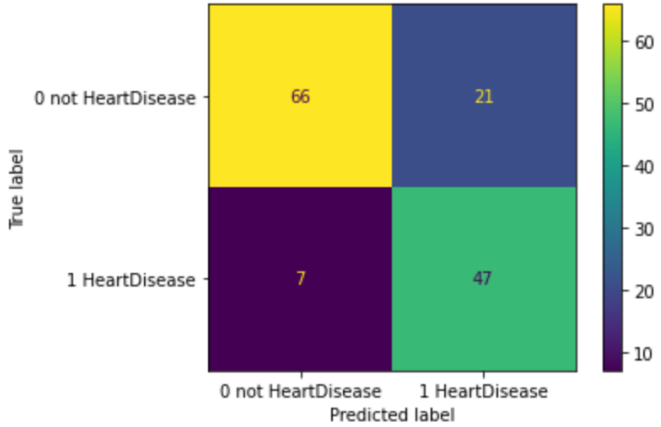
CART (Default)			
max_depth	max_leaf_nodes	min_samples_leaf	Average ACC
None	None	1	81.82 +/- 3.41

CART (手動調參)			
max_depth	**max_leaf_nodes**	**min_samples_leaf**	**Average ACC**
5	**15**	**4**	**83.6 +/- 1.76**

CART

CART〔手動調參〕			
max_depth	max_leaf_nodes	min_samples_leaf	Average ACC
5	15	4	83.6 +/- 1.76

```
Testing result:
Testing ACC: 80.14
Testing f1s: 77.05
Testing pre: 69.12
Testing sen: 87.04
Testing spe: 75.86
```



Linear SVM

Default

Linear SVM	
C=1.0	**Training Average ACC: 85.92 +/- 1.65**

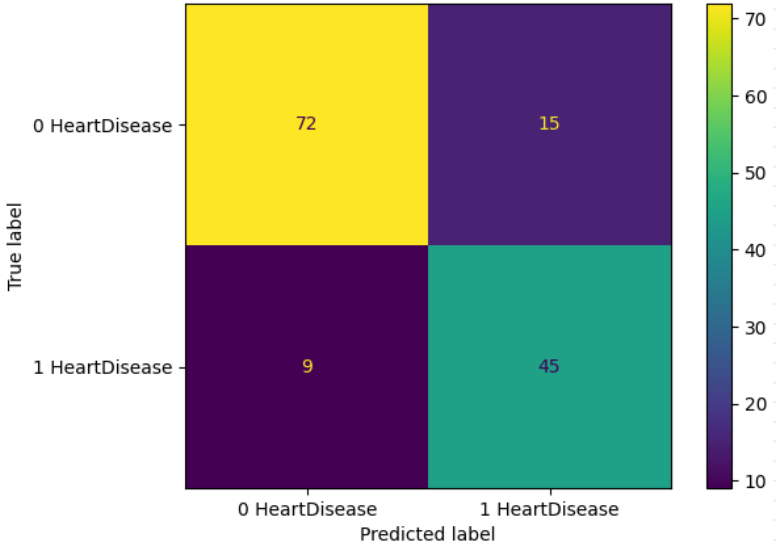
Linear SVM

Training Average ACC	
C=0.0001	86.27 +/- 2.67
C=0.001	87.17 +/- 2.31
C=0.01	87.17 +/- 1.9
C=0.1	86.1 +/-1.31
(Default)C=1.0	85.92 +/- 1.65
C=10	85.74 +/- 1.82
C=100	85.74 +/- 1.82
C=1000	85.74 +/- 1.82
C=10000	85.74 +/- 1.82

Linear SVM

```
Testing ACC:  82.98
Testing f1s:  78.95
Testing pre:  75.0
Testing sen:  83.33
Testing spe:  82.76
```

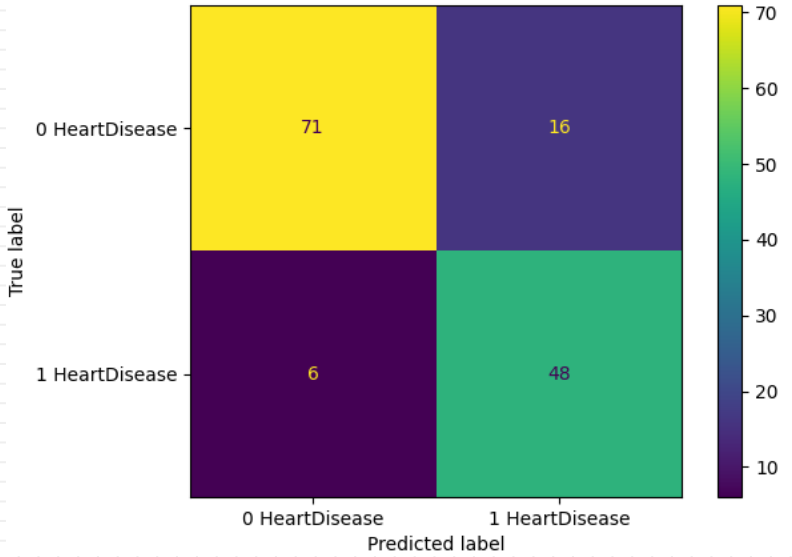
Linear SVM (手動調參)
C=0.001
Average ACC: 87.17 +/- 2.31



Linear SVM

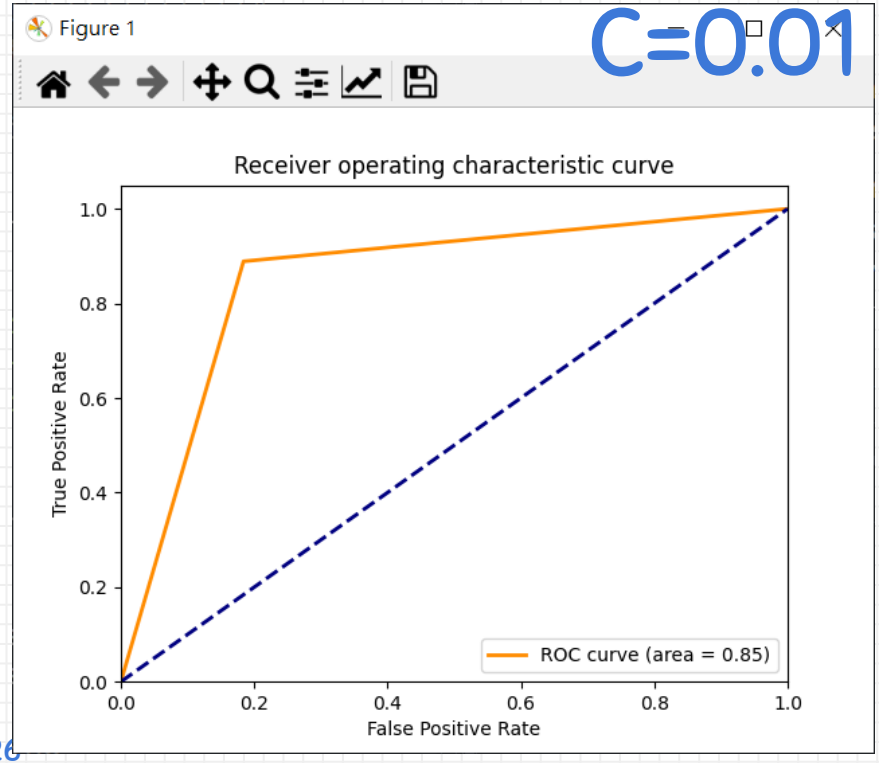
```
Testing ACC:  84.4
Testing f1s:  81.36
Testing pre:  75.0
Testing sen:  88.89
Testing spe:  81.61
```

Linear SVM (手動調參)
C=0.01
Average ACC: 87.17 +/- 1.9

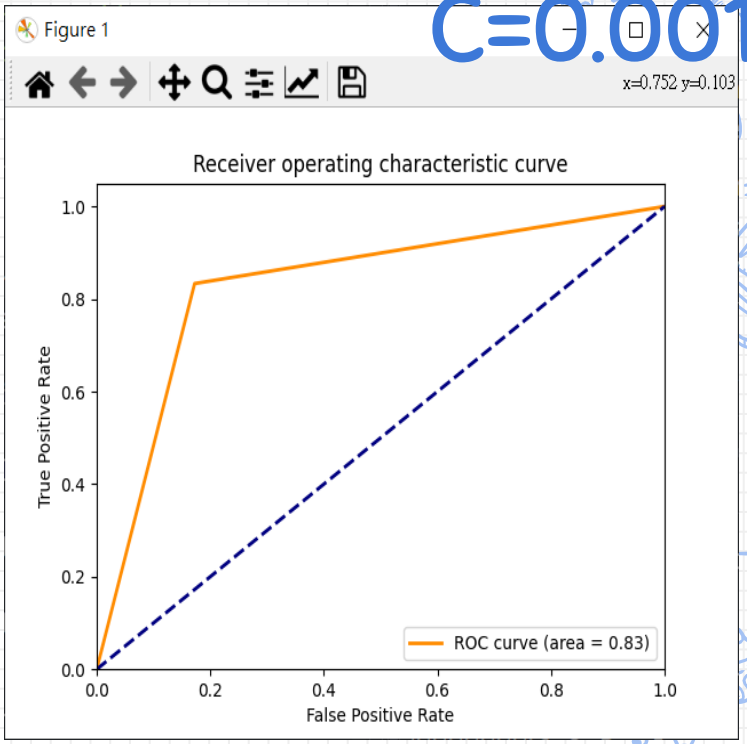


Linear SVM

C=0.01



C=0.001



觀察結果

C=0.001 的validation ACC 坐落區間大於C=0.01

C=0.001	87.17 +/- 2.31
C=0.01	87.17 +/- 1.9

但是C=0.01的Testing ACC大於C=0.001

```
C =  0.01
Testing ACC: 84.4
Testing f1s: 81.36
Testing pre: 75.0
Testing sen: 88.89
Testing spe: 81.61
```

```
C =  0.001
Testing ACC: 82.98
Testing f1s: 78.95
Testing pre: 75.0
Testing sen: 83.33
Testing spe: 82.76
```

SVM–RBF kernel

Default

RBF kernel	
C=1.0 gamma='scale'	**Training Average ACC: 85.92 +/- 0.67**

SVM-RBF kernel

Training Average ACC

Training Result	gamma=0.05	gamma=0.25	gamma=0.5	gamma=0.75
C=10	85.38 +/- 0.5	84.14 +/- 1.97	84.49 +/- 2.27	82.89 +/- 2.86
C=100	84.14 +/- 2.4	81.28 +/- 2.62	80.75 +/- 2.43	80.93 +/- 1.65
C=1000	82.35 +/- 3.06	79.68 +/- 1.31	79.68 +/- 3.06	79.32 +/- 2.77
C=10000	79.5 +/- 2.24	78.97 +/- 4.54	79.14 +/- 3.81	79.32 +/- 2.77

SVM-RBF kernel

👍 Default

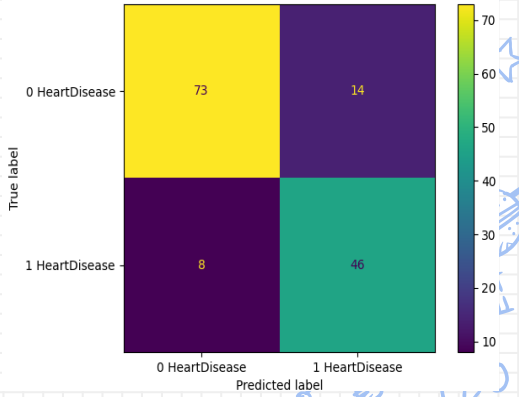
RBF kernel	
C=1.0 gamma= 'scale'	**Training Average** **ACC: 85.92 +/- 0.67**

```
Testing ACC: 85.11
Testing f1s: 81.42
Testing pre: 77.97
Testing sen: 85.19
Testing spe: 85.06
```

手動調參

RBF kernel	
C=10 gamma= 0.05	**Training Average** **ACC: 85.38 +/- 0.5**

```
Testing ACC: 84.4
Testing f1s: 80.7
Testing pre: 76.67
Testing sen: 85.19
Testing spe: 83.91
```

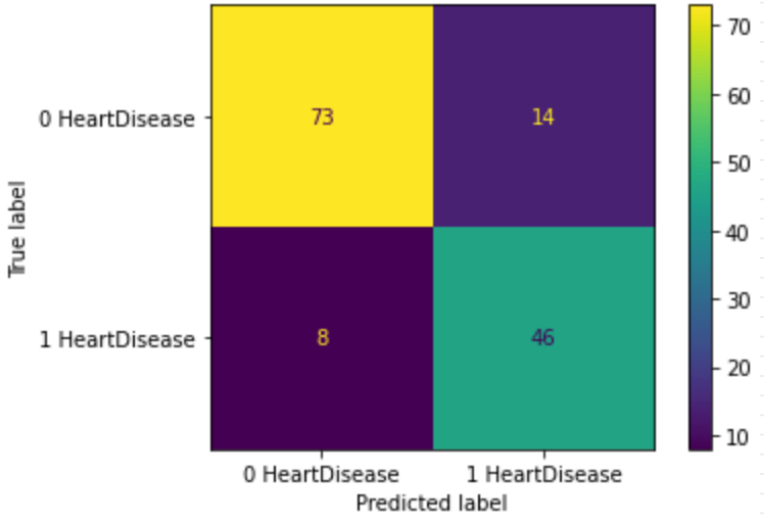


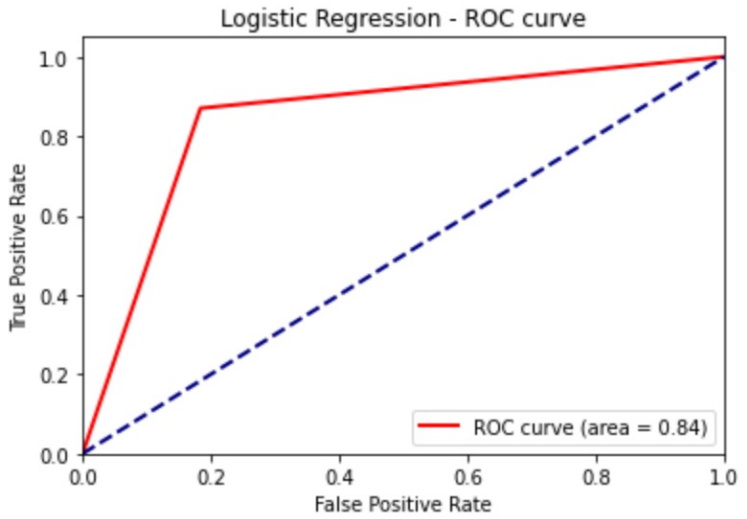
Logistic Regression

Logistic - Training	
Average ACC	**86.1 +/- 2.0**

Logistic - Testing	
Testing ACC	**83.69**
Testing F1	**80.34**
Testing pre	**74.6**
Testing sen	**87.04**
Testing spe	**81.61**

Logistic Regression





The Contrast

模型比較			
Model	**Hyperparameter**	**Training Average ACC**	**Testing ACC**
CART	max_depth = 5 max_leaf_nodes = 15 min_samples_leaf = 4	83.6 +/- 1.76	80.4
Linear SVM	C = 0.01	👍 **87.17 +/- 1.9**	84.4
SVM - RBF	C = 1.0 gamma = scale (Default)	85.92 +/- 0.67	👍 **85.11**
Logistic	(Default)	86.1 +/- 2.0	83.69

Thank You

Question Time