




# 特徵重要度排序

劉建宏

# 研究目的

- 分析各個分類器的特徵重要度排序，
- 找到同樣重要度高的共同特徵，
- 評估該共同特徵是否對原始資料具有一定之預測能力。



# 資料來源：威斯康星州乳腺癌數據集

- 實例數量：569
  - 屬性數量：30
  - 類別分佈：212例惡性、357例良性
- 

# 數據集之屬性特徵

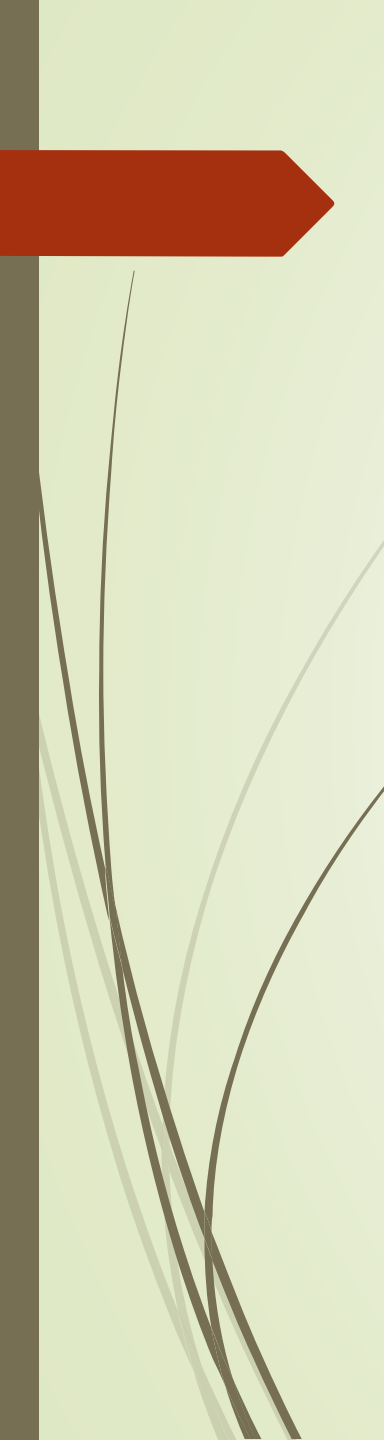
	min	max
radius (mean):	6.981	28.11
texture (mean):	9.71	39.28
perimeter (mean):	43.79	188.5
area (mean):	143.5	2501.0
smoothness (mean):	0.053	0.163
compactness (mean):	0.019	0.345
concavity (mean):	0.0	0.427
concave points (mean):	0.0	0.201
symmetry (mean):	0.106	0.304
fractal dimension (mean):	0.05	0.097
radius (standard error):	0.112	2.873
texture (standard error):	0.36	4.885
perimeter (standard error):	0.757	21.98
area (standard error):	6.802	542.2
smoothness (standard error):	0.002	0.031

	min	max
compactness (standard error):	0.002	0.135
concavity (standard error):	0.0	0.396
concave points (standard error):	0.0	0.053
symmetry (standard error):	0.008	0.079
fractal dimension (standard error):	0.001	0.03
radius (worst):	7.93	36.04
texture (worst):	12.02	49.54
perimeter (worst):	50.41	251.2
area (worst):	185.2	4254.0
smoothness (worst):	0.071	0.223
compactness (worst):	0.027	1.058
concavity (worst):	0.0	1.252
concave points (worst):	0.0	0.291
symmetry (worst):	0.156	0.664
fractal dimension (worst):	0.055	0.208

# 測試AdaBoostClassifier，進行標準化前後是否影響特徵重要度排序

標準化後的特徵重要度排序: ['worst smoothness' 'worst texture' 'mean concave points' 'area error' 'mean texture' 'mean compactness' 'fractal dimension error' 'worst perimeter' 'symmetry error' 'worst concave points' 'worst concavity' 'worst area' 'worst symmetry' 'compactness error' 'worst fractal dimension' 'mean smoothness' 'mean symmetry' 'mean area' 'radius error' 'mean concavity' 'smoothness error' 'mean fractal dimension' 'texture error' 'perimeter error' 'concavity error' 'concave points error' 'worst radius' 'mean perimeter' 'worst compactness' 'mean radius']

標準化前的特徵重要度排序: ['worst smoothness' 'worst texture' 'mean concave points' 'area error' 'mean texture' 'mean compactness' 'fractal dimension error' 'worst perimeter' 'symmetry error' 'worst concave points' 'worst concavity' 'worst area' 'worst symmetry' 'compactness error' 'worst fractal dimension' 'mean smoothness' 'mean symmetry' 'mean area' 'radius error' 'mean concavity' 'smoothness error' 'mean fractal dimension' 'texture error' 'perimeter error' 'concavity error' 'concave points error' 'worst radius' 'mean perimeter' 'worst compactness' 'mean radius']



AdaBoost標準化後特徵重要度排序前10名：

第 1 名 worst smoothness 重要度 0.11

第 2 名 worst texture 重要度 0.07

第 3 名 mean concave points 重要度 0.07

第 4 名 area error 重要度 0.07

第 5 名 mean texture 重要度 0.06

第 6 名 mean compactness 重要度 0.06

第 7 名 fractal dimension error 重要度 0.05

第 8 名 worst perimeter 重要度 0.05

第 9 名 symmetry error 重要度 0.05

第 10 名 worst concave points 重要度 0.04

AdaBoost標準化前特徵重要度排序前10名：

第 1 名 worst smoothness 重要度 0.11

第 2 名 worst texture 重要度 0.07

第 3 名 mean concave points 重要度 0.07

第 4 名 area error 重要度 0.07

第 5 名 mean texture 重要度 0.06

第 6 名 mean compactness 重要度 0.06

第 7 名 fractal dimension error 重要度 0.05

第 8 名 worst perimeter 重要度 0.05

第 9 名 symmetry error 重要度 0.05

第 10 名 worst concave points 重要度 0.04



# 標準化是否並不影響AdaBoost之原因

- AdaBoost的核心思想是根據先前分類器的錯誤來加權訓練數據，以便下一個分類器可以更好地處理錯誤分類的樣本。
- 這種加權過程可以讓AdaBoost在處理非線性、不平衡或具有噪聲的數據時表現出很強的魯棒性。
- AdaBoost本身不要求對數據進行標準化或縮放，但根據具體情況和數據特點，雖然可以選擇在訓練之前對數據進行適當的標準化或縮放操作，以獲得更好的結果。
- 然對於該breast cancer數據判斷良性或惡性腫瘤而言，數據是否標準化並不影響AdaBoost對特徵的重要度排序。

# 測試RandomForestClassifier，進行標準化前後是否影響 特徵重要度排序

標準化後的特徵重要度排序: ['worst perimeter' 'worst concave points' 'worst radius' 'worst area'  
'mean concave points' 'mean perimeter' 'mean concavity' 'mean area'  
'mean radius' 'area error' 'worst texture' 'worst concavity'  
'mean texture' 'radius error' 'worst smoothness' 'worst symmetry'  
'mean compactness' 'worst fractal dimension' 'worst compactness'  
'perimeter error' 'symmetry error' 'mean smoothness' 'texture error'  
'concavity error' 'concave points error' 'compactness error'  
'mean fractal dimension' 'fractal dimension error' 'smoothness error'  
'mean symmetry']

標準化前的特徵重要度排序: ['worst perimeter' 'worst concave points' 'worst radius' 'worst area'  
'mean concave points' 'mean perimeter' 'mean concavity' 'mean area'  
'mean radius' 'area error' 'worst texture' 'worst concavity'  
'mean texture' 'radius error' 'worst smoothness' 'worst symmetry'  
'mean compactness' 'worst fractal dimension' 'worst compactness'  
'perimeter error' 'symmetry error' 'mean smoothness' 'texture error'  
'concavity error' 'concave points error' 'compactness error'  
'mean fractal dimension' 'fractal dimension error' 'smoothness error'  
'mean symmetry']



RandomForestClassifier標準化後特徵重要度排序前10名：

第 1 名 worst perimeter 重要度  
0.17351898987710243

第 2 名 worst concave points 重要度  
0.12122546127957205

第 3 名 worst radius 重要度 0.10904512014019425

第 4 名 worst area 重要度 0.07300596244768266

第 5 名 mean concave points 重要度  
0.07186697208262177

第 6 名 mean perimeter 重要度  
0.07171383430487879

第 7 名 mean concavity 重要度  
0.05435985721864484

第 8 名 mean area 重要度 0.04625303010225377

第 9 名 mean radius 重要度 0.037430642075498656

第 10 名 area error 重要度 0.03560738826235755

RandomForestClassifier標準化前特徵重要度排序前10名：

第 1 名 worst perimeter 重要度  
0.17351898987710243

第 2 名 worst concave points 重要度  
0.12122546127957205

第 3 名 worst radius 重要度 0.10904512014019425

第 4 名 worst area 重要度 0.07300596244768266

第 5 名 mean concave points 重要度  
0.07186697208262177

第 6 名 mean perimeter 重要度  
0.07171383430487879

第 7 名 mean concavity 重要度  
0.05435985721864484

第 8 名 mean area 重要度 0.04625303010225377

第 9 名 mean radius 重要度 0.037430642075498656

第 10 名 area error 重要度 0.03560738826235755

# 標準化是否並不影響RandomForest之原因

- 由於RandomForest是基於決策樹的集成模型，決策樹本身對特徵的尺度和範圍不敏感。
- 每個決策樹都獨立地進行特徵分割，而不會受到特徵值的絕對大小的影響。
- RandomForest本身不要求對數據進行標準化或縮放，但根據具體情況和數據特點，雖然可以選擇在訓練之前對數據進行適當的標準化或縮放操作，以獲得更好的結果。
- 然對於該breast cancer數據判斷良性或惡性腫瘤而言，數據是否標準化並不影響RandomForest對特徵的重要度排序。

# 比較Adaboost與RandomForest的特徵重要度排序結果

AdaBoost演算法的特徵重要度排序: ['worst smoothness' 'worst texture' 'mean concave points' 'area error'

'mean texture' 'mean compactness' 'fractal dimension error'

'worst perimeter' 'symmetry error' 'worst concave points'

'worst concavity' 'worst area' 'worst symmetry' 'compactness error'

'worst fractal dimension' 'mean smoothness' 'mean symmetry' 'mean area'

'radius error' 'mean concavity' 'smoothness error'

'mean fractal dimension' 'texture error' 'perimeter error'

'concavity error' 'concave points error' 'worst radius' 'mean perimeter'

'worst compactness' 'mean radius']

RandomForest演算法的特徵重要度排序: ['worst perimeter' 'worst concave points' 'worst radius' 'worst area'

'mean concave points' 'mean perimeter' 'mean concavity' 'mean area'

'mean radius' 'area error' 'worst texture' 'worst concavity'

'mean texture' 'radius error' 'worst smoothness' 'worst symmetry'

'mean compactness' 'worst fractal dimension' 'worst compactness'

'perimeter error' 'symmetry error' 'mean smoothness' 'texture error'

'concavity error' 'concave points error' 'compactness error'

'mean fractal dimension' 'fractal dimension error' 'smoothness error'

'mean symmetry']

# Adaboost與RandomForest的特徵重要度排序結果不同之理由

- 從上述結果可以得知兩者對於特徵重要度的排序具有顯著差異。理由可能如下：
- AdaBoost傾向於在每個迭代中更關注那些容易被錯誤分類的樣本，因此對於這些特徵的重要度可能更高。
- 而Random Forest通過隨機選擇特徵子集進行訓練，可能會導致特徵的重要度分散在多個決策樹中。
- 由於AdaBoost和Random Forest可能對數據集的特徵敏感程度不同，因此可能產生不同的特徵重要度排序。

# 在AdaBoost與RandomForest的特徵重要度的排序前10名中，取得共同的特徵

AdaBoost的特徵重要度的排序前10名:

- 第 1 名 worst smoothness
- 第 2 名 worst texture
- 第 3 名 mean concave points
- 第 4 名 area error
- 第 5 名 mean texture
- 第 6 名 mean compactness
- 第 7 名 fractal dimension error
- 第 8 名 worst perimeter
- 第 9 名 symmetry error
- 第 10 名 worst concave points

RandomForest的特徵重要度的排序前10名:

- 第 1 名 worst perimeter
- 第 2 名 worst concave points
- 第 3 名 worst radius
- 第 4 名 worst area
- 第 5 名 mean concave points
- 第 6 名 mean perimeter
- 第 7 名 mean concavity
- 第 8 名 mean area
- 第 9 名 mean radius
- 第 10 名 area error

共同的特徵為：

mean concave points, area error, worst perimeter, worst concave points



# 在AdaBoost與RandomForest的特徵重要度的排序前10名中，取得共同的特徵之理由

- 在AdaBoost與RandomForest的特徵重要度的排序前10名中，取得共同的特徵為：'mean concave points', 'area error', 'worst perimeter', 'worst concave points'。
- 雖然AdaBoost和Random Forest在特徵重要度排序時使用了不同的算法原理和策略，但仍然有一些共同的特徵被兩種算法都認為是重要的。
- 這些共同的重要特徵可能在數據集中具有較高的信息量和重要性，並且在不同算法中的表現相對穩定，並且可能具有更廣泛的預測能力。



## 測試Logistic regression特徵重要度排序之事前需知

- Solver = liblinear：這是一個比較常用的求解器，適用於二分類問題。
- 計算特徵重要度：通過計算邏輯回歸模型係數的絕對值來估計特徵的重要度；
- 係數的絕對值表示特徵對目標變量的影響程度，故對係數的絕對值進行排序，以獲得特徵重要度的近似排名。

# 測試Logistic regression,進行標準化前後是否影響特徵重要度排序

標準化後的特徵重要度排序: ['radius error' 'worst texture' 'compactness error' 'worst radius' 'worst area' 'mean concavity' 'area error' 'mean concave points' 'worst concavity' 'fractal dimension error' 'worst concave points' 'worst perimeter' 'worst symmetry' 'symmetry error' 'perimeter error' 'worst smoothness' 'worst fractal dimension' 'mean texture' 'mean compactness' 'concave points error' 'mean area' 'mean perimeter' 'mean radius' 'smoothness error' 'mean symmetry' 'mean fractal dimension' 'concavity error' 'texture error' 'worst compactness' 'mean smoothness']

標準化前的特徵重要度排序: ['mean radius' 'worst concavity' 'worst compactness' 'worst radius' 'texture error' 'mean concavity' 'worst concave points' 'worst symmetry' 'worst texture' 'mean compactness' 'mean concave points' 'worst smoothness' 'worst perimeter' 'mean texture' 'mean smoothness' 'mean symmetry' 'perimeter error' 'worst fractal dimension' 'mean perimeter' 'area error' 'concavity error' 'radius error' 'concave points error' 'mean fractal dimension' 'worst area' 'mean area' 'compactness error' 'smoothness error' 'fractal dimension error' 'symmetry error']

Logistic regression標準化後特徵重要度排序前10名：

第 1 名 radius error 重要度 1.3128092363552692

第 2 名 worst texture 重要度 1.2093912738530388

第 3 名 compactness error 重要度  
1.0917895189643927

第 4 名 worst radius 重要度 0.9541058179942675

第 5 名 worst area 重要度 0.9411564471744772

第 6 名 mean concavity 重要度  
0.9326030633555452

第 7 名 area error 重要度 0.92878123031847

第 8 名 mean concave points 重要度  
0.9208500758706084

第 9 名 worst concavity 重要度 0.9017702893656258

第 10 名 fractal dimension error 重要度  
0.8077089759718497

Logistic regression標準化前特徵重要度排序前10名：

第 1 名 mean radius 重要度 1.8370763389991198

第 2 名 worst concavity 重要度 1.4903492186482363

第 3 名 worst compactness 重要度  
1.022618976623101

第 4 名 worst radius 重要度 0.9494911392220109

第 5 名 texture error 重要度 0.8430861863004296

第 6 名 mean concavity 重要度  
0.5994587649796344

第 7 名 worst concave points 重要度  
0.5740699895453317

第 8 名 worst symmetry 重要度 0.3874161620782656

第 9 名 worst texture 重要度 0.3776120878425882

第 10 名 mean compactness 重要度  
0.36175768875355413

# 標準化影響Logistic Regression之原因

- 由於Logistic Regression在進行優化時，求解器（solver）可能對數據的尺度敏感。
- 較大尺度的特徵可能會對結果產生更大的影響，這可能導致在特徵重要度排序中給予較大尺度的特徵更高的重要度，從而影響了模型係數的估計結果。這可能會導致特徵重要度排序的不確定性。
- 故未進行數據標準化可能會導致邏輯回歸模型中特徵重要度排序的不同。
- 標準化可以消除特徵尺度的差異，減小噪聲特徵的影響，並提供更穩定的優化過程。

# 使用Logistic Regression，在進行標準化前後數據特徵重要度的排序前10名中，取得共同的特徵

Logistic Regression標準化後的特徵重要度的排序前10名:

- 第 1 名 radius error
- 第 2 名 worst texture
- 第 3 名 compactness error
- 第 4 名 worst radius
- 第 5 名 worst area
- 第 6 名 mean concavity
- 第 7 名 area error
- 第 8 名 mean concave points
- 第 9 名 worst concavity
- 第 10 名 fractal dimension error

Logistic Regression標準化前的特徵重要度的排序前10名:

- 第 1 名 mean radius
- 第 2 名 worst concavity
- 第 3 名 worst compactness
- 第 4 名 worst radius
- 第 5 名 texture error
- 第 6 名 mean concavity
- 第 7 名 worst concave points
- 第 8 名 worst symmetry
- 第 9 名 worst texture
- 第 10 名 mean compactness

共同的特徵為：

worst texture, worst radius, mean concavity, worst concavity





## 使用Logistic Regression，在進行標準化前後數據特徵重要度的排序前10名中，取得共同的特徵之原因

- 使用Logistic Regression，在進行標準化前後數據特徵重要度的排序前10名中，取得共同的特徵為:'worst texture', 'worst radius', 'mean concavity', 'worst concavity'。
- 這些共同的特徵在標準化前後都被模型認為對目標變量有較高的影響力，
- 並在不同的尺度下都對目標變量的預測能力具有具有較高的重要性和穩定性。



# 在AdaBoost與Logistic Regression(標準化後)的特徵重要度的排序前10名中，取得共同的特徵

AdaBoost的特徵重要度的排序前10名:

- 第 1 名 worst smoothness
- 第 2 名 worst texture
- 第 3 名 mean concave points
- 第 4 名 area error
- 第 5 名 mean texture
- 第 6 名 mean compactness
- 第 7 名 fractal dimension error
- 第 8 名 worst perimeter
- 第 9 名 symmetry error
- 第 10 名 worst concave points

Logistic Regression標準化後的特徵重要度的排序前10名:

- 第 1 名 radius error
- 第 2 名 worst texture
- 第 3 名 compactness error
- 第 4 名 worst radius
- 第 5 名 worst area
- 第 6 名 mean concavity
- 第 7 名 area error
- 第 8 名 mean concave points
- 第 9 名 worst concavity
- 第 10 名 fractal dimension error

共同的特徵為：

worst texture, mean  
concave points, area error,  
fractal dimension error

## 在RandomForest與Logistic Regression(標準化後)的特徵重要度的排序前10名中，取得共同的特徵

RandomForest的特徵重要度的排序前10名:

- 第 1 名 worst perimeter
- 第 2 名 worst concave points
- 第 3 名 worst radius
- 第 4 名 worst area
- 第 5 名 mean concave points
- 第 6 名 mean perimeter
- 第 7 名 mean concavity
- 第 8 名 mean area
- 第 9 名 mean radius
- 第 10 名 area error

Logistic Regression標準化後的特徵重要度的排序前10名:

- 第 1 名 radius error
- 第 2 名 worst texture
- 第 3 名 compactness error
- 第 4 名 worst radius
- 第 5 名 worst area
- 第 6 名 mean concavity
- 第 7 名 area error
- 第 8 名 mean concave points
- 第 9 名 worst concavity
- 第 10 名 fractal dimension error

共同的特徵為：

worst radius, worst area,  
mean concave points,  
mean concavity, area error

## 在AdaBoost、RandomForest與Logistic Regression 的特徵重要度的排序前10名中，取得共同的特徵

AdaBoost的特徵重要度的排序前10名:

- 第 1 名 worst smoothness
- 第 2 名 worst texture
- 第 3 名 mean concave points
- 第 4 名 area error
- 第 5 名 mean texture
- 第 6 名 mean compactness
- 第 7 名 fractal dimension error
- 第 8 名 worst perimeter
- 第 9 名 symmetry error
- 第 10 名 worst concave points

RandomForest的特徵重要度的排序前10名:

- 第 1 名 worst perimeter
- 第 2 名 worst concave points
- 第 3 名 worst radius
- 第 4 名 worst area
- 第 5 名 mean concave points
- 第 6 名 mean perimeter
- 第 7 名 mean concavity
- 第 8 名 mean area
- 第 9 名 mean radius
- 第 10 名 area error

Logistic Regression標準化後  
的特徵重要度的排序前10名:

- 第 1 名 radius error
- 第 2 名 worst texture
- 第 3 名 compactness error
- 第 4 名 worst radius
- 第 5 名 worst area
- 第 6 名 mean concavity
- 第 7 名 area error
- 第 8 名 mean concave points
- 第 9 名 worst concavity
- 第 10 名 fractal dimension error

三者共同  
的特徵為：

mean  
concave  
points,  
area error



## 在AdaBoost、Random Forest和Logistic Regression的特徵重要度的排序前10名中，取得共同的特徵的理由

- 在AdaBoost、Random Forest和Logistic Regression的特徵重要度的排序前10名中，取得共同的特徵為：'mean concave points', 'area error'。
- 這些共同的特徵在不同的模型中都被認為對目標變量的預測具有高信息量。
- 它們可能包含了對目標變量有重要解釋的關鍵因素和特徵。

# 在已標準化後的數據，只取'mean concave points', 'area error'特徵進行SVM訓練

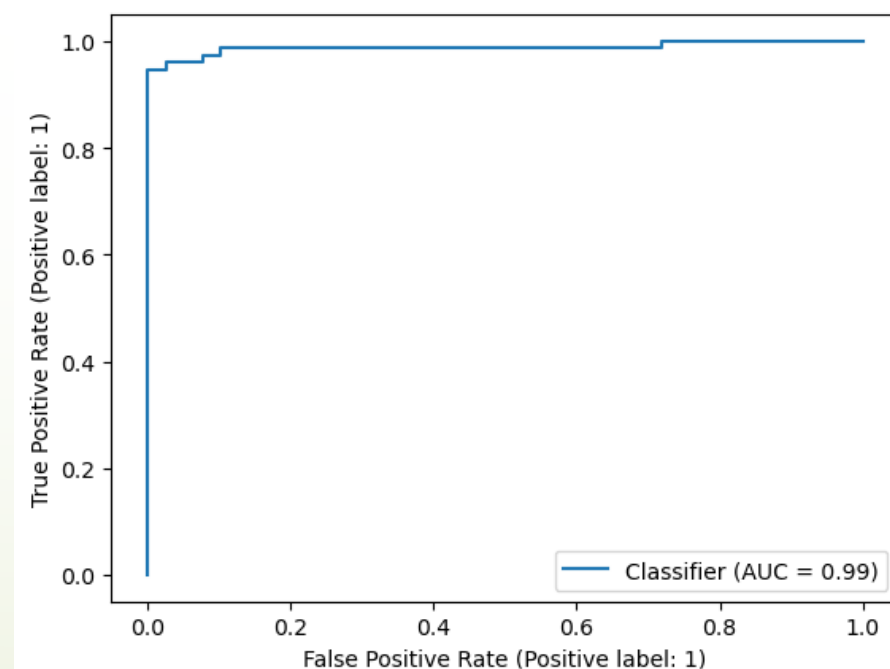
使用RBF核函數，它是默認的內核函數，是支持向量機中最常用的非線性內核函數，它可以有效地處理各種類型的數據。

training score: 0.9098901098901099  
test score: 0.9473684210526315

Confusion matrix:  
[34 5]  
[ 1 74]

K-fold Scores:  
Fold 1: 0.8947368421052632  
Fold 2: 0.9210526315789473  
Fold 3: 0.9210526315789473  
Fold 4: 0.9298245614035088  
Fold 5: 0.9292035398230089  
Mean Score: 0.9191740412979351

	precision	recall	f1-score	support
0	0.971429	0.871795	0.918919	39.0
1	0.936709	0.986667	0.961039	75.0
accuracy	0.947368			114.0
macro avg	0.954069	0.929231	0.939979	114.0
weighted avg	0.948587	0.947368	0.946629	114.0





# 在已標準化後的數據中，將全部30種屬性的特徵進行SVM訓練

training score: 0.9846153846153847

test score: 0.9736842105263158

Confusion matrix:

[39 0]

[ 3 72]

	precision	recall	f1-score	support
0	0.928571	1.000000	0.962963	39.0
1	1.000000	0.960000	0.979592	75.0
accuracy	0.973684			114.0
macro avg	0.964286	0.980000	0.971277	114.0
weighted avg	0.975564	0.973684	0.973903	114.0

K-fold Scores:

Fold 1: 0.9736842105263158

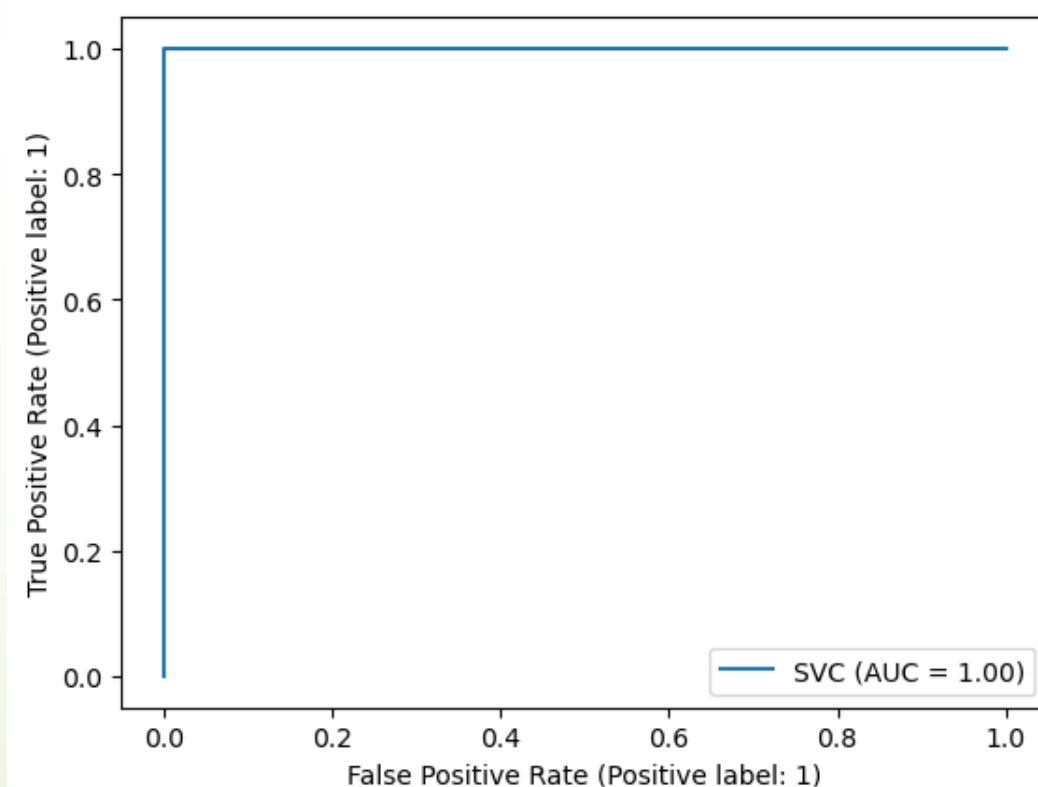
Fold 2: 0.956140350877193

Fold 3: 1.0

Fold 4: 0.9649122807017544

Fold 5: 0.9734513274336283

Mean Score: 0.9736376339077782





# 探討使用共同的特徵與全部的特徵在SVM 訓練下之差異

- 在已標準化的數據透過SVM訓練下，只使用'mean concave points'和'area error'特徵集的平均分數為0.92，而使用全部特徵的特徵集的平均分數為0.97；只使用'mean concave points'和'area error'特徵集的F1-score為0.95，而使用全部特徵的特徵集的F1-score為0.97。
- 這意味著'mean concave points'和'area error'這兩個特徵對於乳腺癌分類預測具有一定的預測能力，
- 它們可以在僅使用這兩個特徵的情況下得到相對較高的預測準確度與較高的F1-score。

# 在已標準化後的數據並進行SVM訓練之下，找到兩個特徵組合集的重要度排序

2個特徵組合的重要度排序前10名：

第 1 名 ['mean concave points' 'radius error'] 重要度 0.9736842105263158

第 2 名 ['mean concave points' 'perimeter error'] 重要度 0.9649122807017544

第 3 名 ['radius error' 'worst concave points'] 重要度 0.9649122807017544

第 4 名 ['mean compactness' 'worst area'] 重要度 0.9649122807017544

第 5 名 ['mean compactness' 'worst radius'] 重要度 0.9649122807017544

第 6 名 ['worst area' 'worst compactness'] 重要度 0.9649122807017543

第 7 名 ['mean compactness' 'worst perimeter'] 重要度 0.9561403508771931

第 8 名 ['mean concave points' 'worst texture'] 重要度 0.956140350877193

第 9 名 ['mean concave points' 'texture error'] 重要度 0.956140350877193

第 10 名 ['mean concave points' 'area error'] 重要度 0.956140350877193

## 探討共同特徵集在進行兩個特徵組合集的重要度排序SVM訓練之下，只有第10名之原因

- 雖然‘mean concave points’, ‘area error’，在已標準化後的數據並進行SVM訓練之下，該兩個特徵組合集的重要度排序為第10名，但依然有0.956分，相比較下，其他排名較前的兩個特徵組合集亦可能具有對分類預測有重要貢獻的信息。
- 可以發現在前10名的組合中，mean concave points的特徵出現最多次，共出現5次，這意味著‘mean concave points’這個特徵對於乳腺癌分類預測具有一定的預測能力，
- 它可能對腫瘤是否為良性具有重要解釋的關鍵因素。

# 在未標準化的數據，只取'mean concave points', 'area error'特徵進行decision tree訓練

由於決策樹算法基於特徵的閾值來進行分割，因此它們對  
特徵的尺度不敏感

training score: 0.9274725274725275  
test score: 0.9035087719298246

Confusion Matrix:

[32 7]  
[ 4 71]

	precision	recall	f1-score	support
0	0.888889	0.820513	0.853333	39.0
1	0.910256	0.946667	0.928105	75.0
accuracy			0.903509	114.0
macro avg	0.899573	0.883590	0.890719	114.0
weighted avg	0.902946	0.903509	0.902525	114.0

K-fold Scores:

Fold 1: 0.9035087719298246

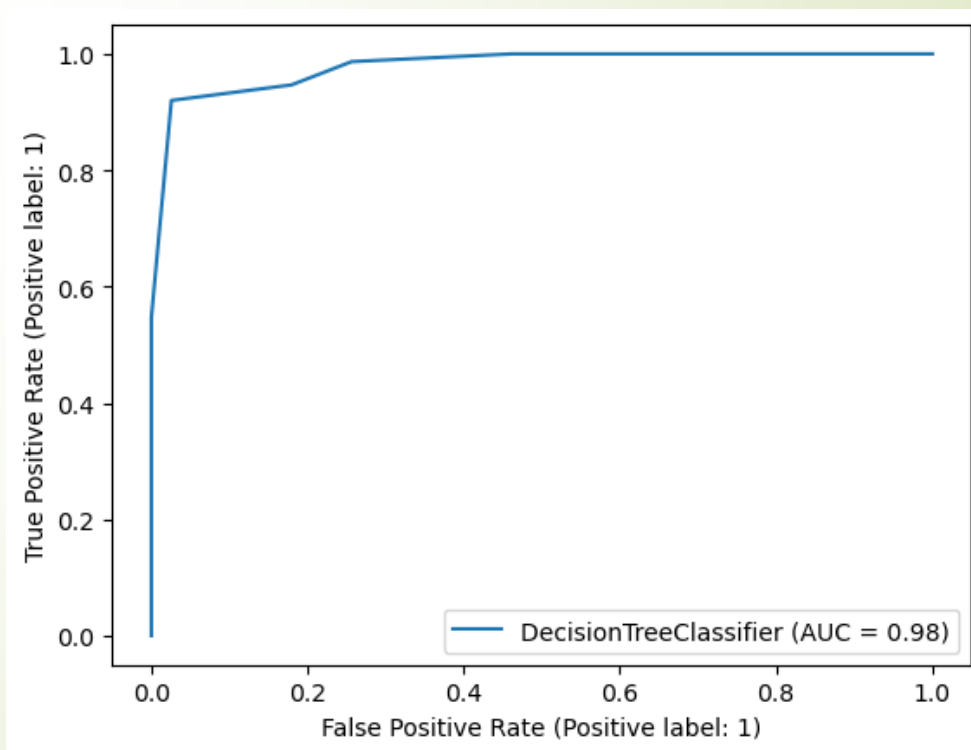
Fold 2: 0.9122807017543859

Fold 3: 0.956140350877193

Fold 4: 0.9385964912280702

Fold 5: 0.9557522123893806

Mean Score: 0.9332557056357709







# 在已標準化的數據，只取'mean concave points', 'area error'特徵進行Logistic regression訓練

由於由於Logistic Regression在進行優化時，求解器 ( solver ) 可能對數據的尺度敏感，標準化可以消除特徵尺度的差異，減小噪聲特徵的影響，並提供更穩定的優化過程。

training score: 0.9142857142857143  
test score: 0.9473684210526315

Confusion Matrix:

[34 5]  
[ 1 74]

	precision	recall	f1-score	support
0	0.971429	0.871795	0.918919	39.0
1	0.936709	0.986667	0.961039	75.0
accuracy			0.947368	114.0
macro avg	0.954069	0.929231	0.939979	114.0
weighted avg	0.948587	0.947368	0.946629	114.0

K-fold Scores:

Fold 1: 0.9824561403508771

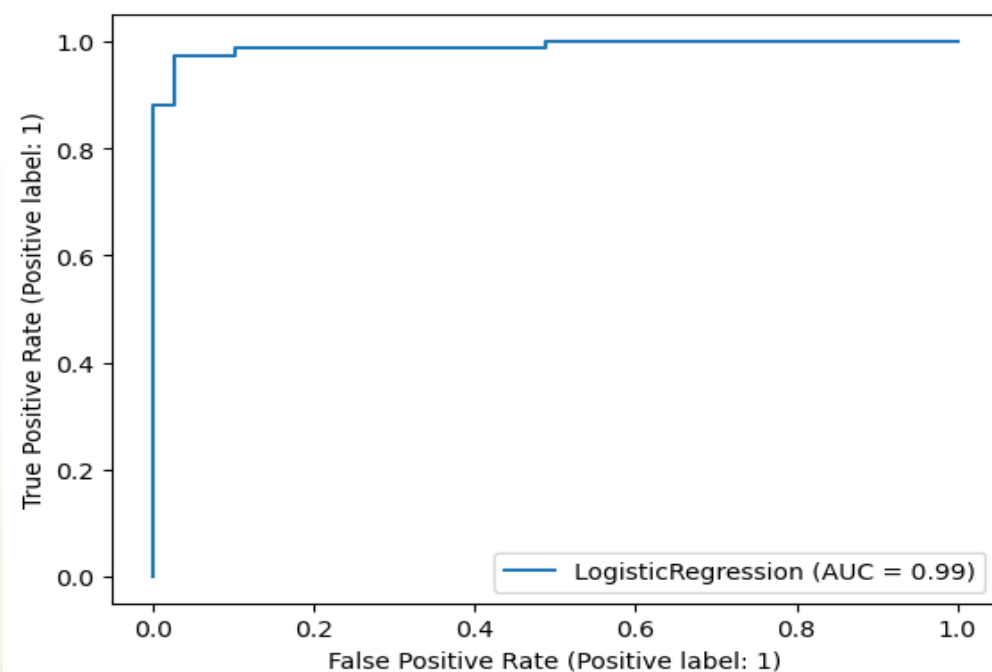
Fold 2: 0.9736842105263158

Fold 3: 0.9736842105263158

Fold 4: 0.9736842105263158

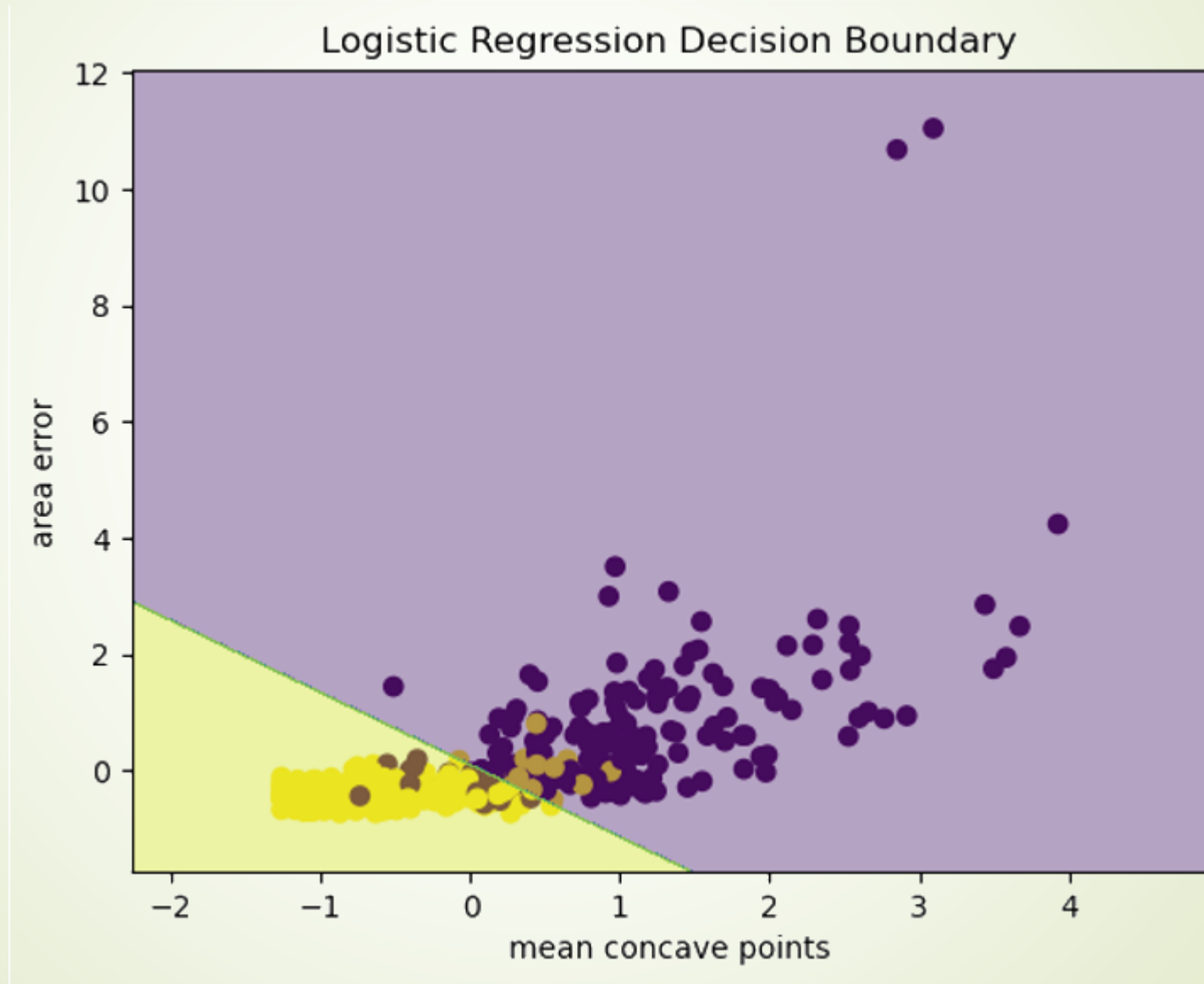
Fold 5: 0.9911504424778761

Mean Score: 0.9789318428815402





# 視覺化Logistic regression決策邊界線



# 在未標準化的數據，只取'mean concave points', 'area error'特徵進行Adaboost訓練

AdaBoost的主要關注點是樣本的權重調整，而不是特徵的縮放或標準化，因此，即使數據沒有進行縮放或標準化，AdaBoost仍然可以有效地工作。

training score: 0.9604395604395605

test score: 0.9385964912280702

Confusion Matrix:

[36 3]  
[ 4 71]

	precision	recall	f1-score	support
0	0.900000	0.923077	0.911392	39.0
1	0.959459	0.946667	0.953020	75.0
accuracy			0.938596	114.0
macro avg	0.929730	0.934872	0.932206	114.0
weighted avg	0.939118	0.938596	0.938779	114.0

K-fold Scores:

Fold 1: 0.9473684210526315

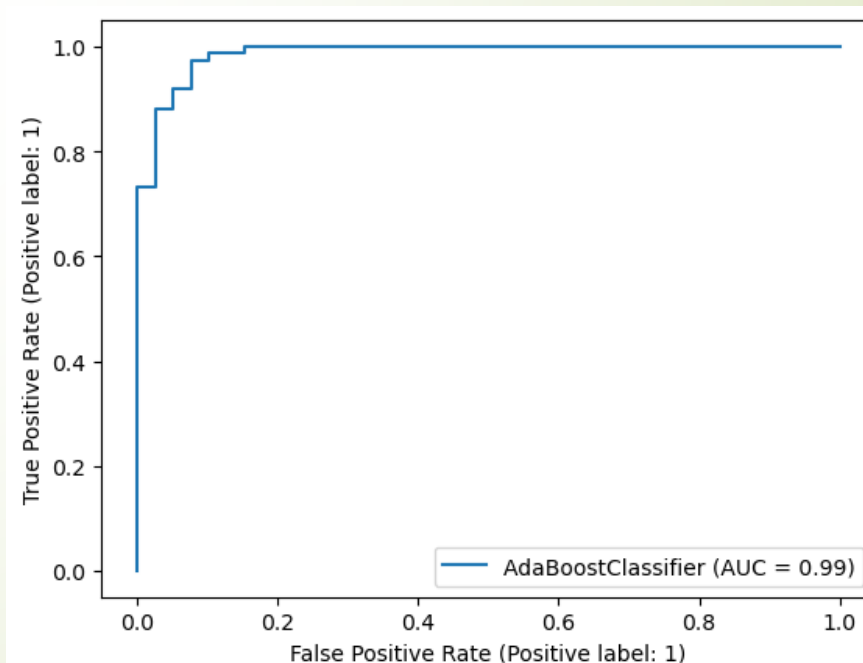
Fold 2: 0.9649122807017544

Fold 3: 0.9912280701754386

Fold 4: 0.9912280701754386

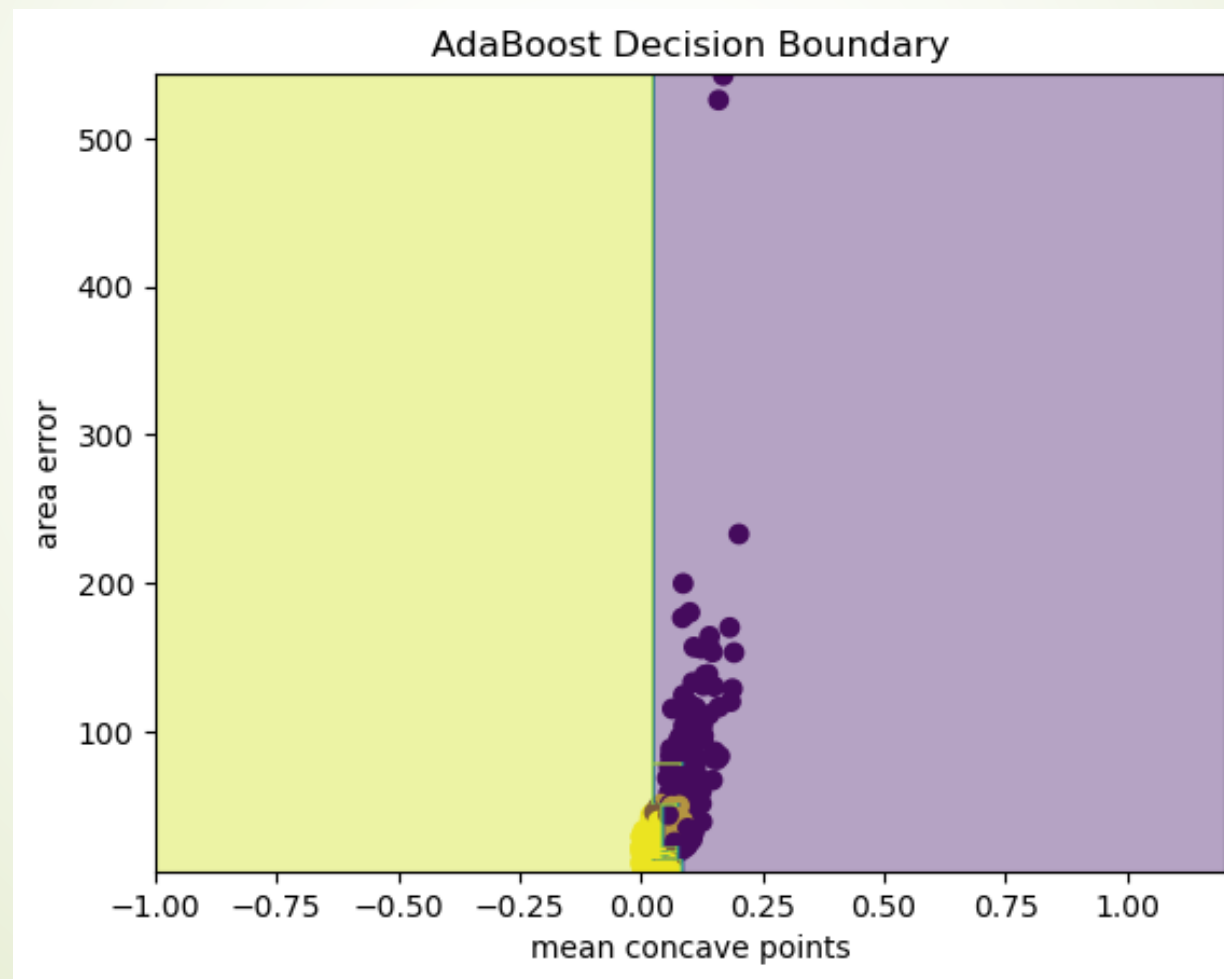
Fold 5: 0.9823008849557522

Mean Score: 0.9754075454122031

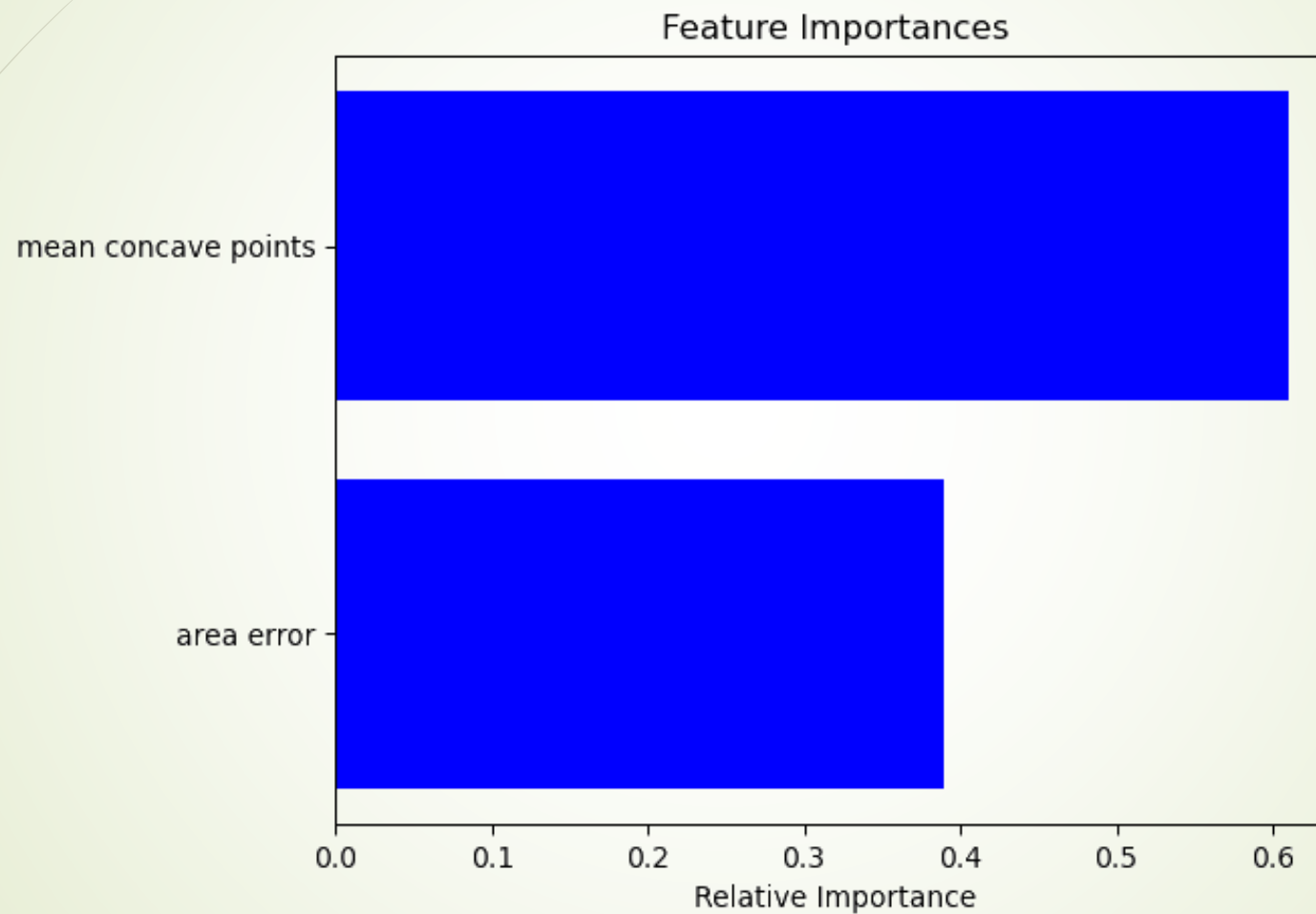


# 視覺化AdaBoost決策邊界線

AdaBoost算法本身並不直接生成決策邊界線；  
然而，可以通過繪製模型在特徵空間中的分類結果來近似  
可視化決策邊界。



## 在AdaBoost訓練下的兩個特徵的相對重要性程度



在已標準化後的數據，只取'mean concave points',  
'area error'特徵進行SVM的線性的核函數訓練

training score: 0.9120879120879121

test score: 0.9473684210526315

Confusion Matrix:

[35 4]

[ 2 73]

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.945946	0.897436	0.921053	39.0
---	----------	----------	----------	------

1	0.948052	0.973333	0.960526	75.0
---	----------	----------	----------	------

accuracy			0.947368	114.0
----------	--	--	----------	-------

macro avg	0.946999	0.935385	0.940789	114.0
-----------	----------	----------	----------	-------

weighted avg	0.947331	0.947368	0.947022	114.0
--------------	----------	----------	----------	-------

K-fold Scores:

Fold 1: 0.956140350877193

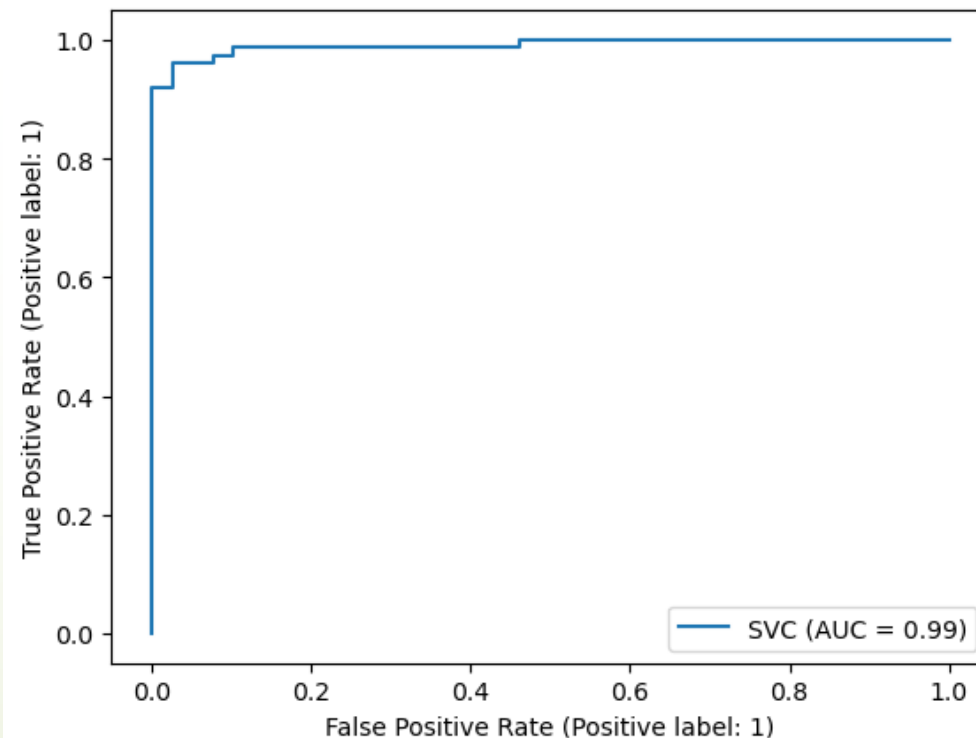
Fold 2: 0.9824561403508771

Fold 3: 0.9649122807017544

Fold 4: 0.9649122807017544

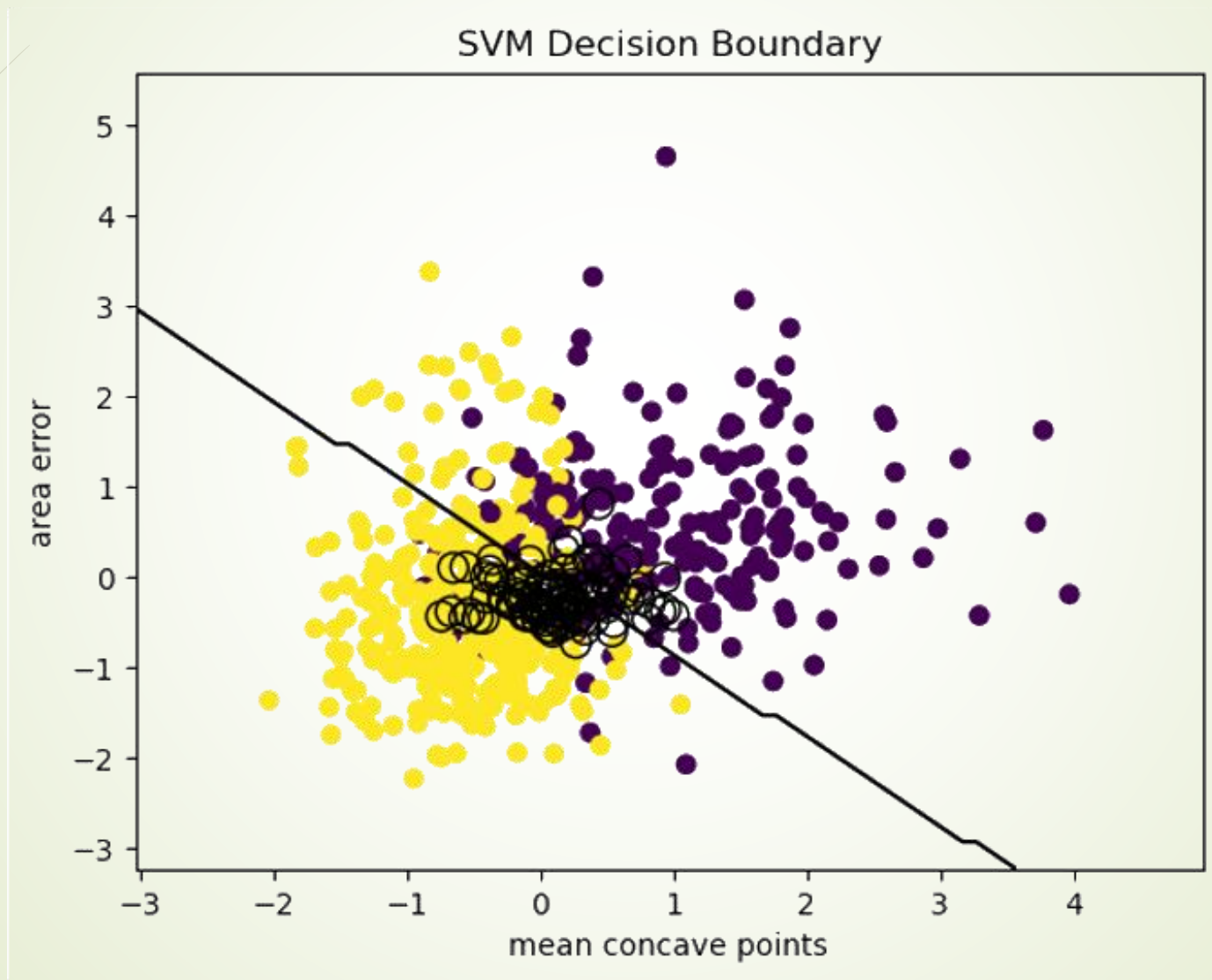
Fold 5: 0.9823008849557522

Mean Score: 0.9701443875174661





# 視覺化SVM決策邊界線



# 解釋視覺化SVM決策邊界線

- 多個黑色圈圈出現在決策邊界上的涵義是這些點被標識為支持向量 ( Support Vectors )，這有助於將它們與其他數據點區分開來，並強調它們在模型中的重要性。
- 因此，當在SVM的決策邊界上看到多個黑色圈圈時，它們表示了訓練數據中離決策邊界最近的支持向量點，這些點對於模型的決策邊界起到關鍵作用，因為它們確定了決策邊界的位置和形狀。

# 結論

- 綜合以上，本專題透過AdaBoost、RandomForest與LogisticRegression分類器的特徵重要度排序前10名，獲得同樣排序前10名的共同特徵，該共同特徵組合為'mean concave points'以及'area error'，並且使用該組合透過各個分類器，預測該組合是否皆具有一定的預測能力。
- 結果透過K-fold驗證法：SVM的RBF核函數分數為0.956；SVM的線性核函數分數為0.970；Decision tree在深度為3的分數為0.933；LogisticRegression在solver為liblinear的分數為0.978；AdaBoost的分數為0.973。
- 相對原breast cancer資料具有的30個特徵，只透過2個共同特徵的組合，就應已具有對於breast cancer是否為良性或惡性的一定預測能力。
- 故透過上述方式，應可大幅減少對於多個特徵的前處理以及取得，並可聚焦在該共同特徵的信息量與關鍵解釋因素上。

# mean concave points的特徵涵義

- 在乳腺癌數據集中，"mean concave points"（平均凹點數）是一個特徵，用於描述細胞核輪廓中的凹陷點的數量。
- 它表示細胞核邊界的不規則性和複雜性。"mean concave points" 的含義是細胞核輪廓中的凹陷點數量的平均值。
- 凹陷點是指細胞核邊界中凹下去的區域，通常與細胞核內部的不規則結構和細胞核之間的粘連有關。
- 較高的"mean concave points" 值表示細胞核輪廓中具有更多的凹陷點，意味著細胞核形狀更加不規則和複雜。這種不規則性和複雜性可能與腫瘤的惡性程度相關，因為惡性腫瘤的細胞核通常具有更多的不規則特徵和複雜的邊界。

# area error的特徵涵義

- 在乳腺癌數據集中，"area error"（面積誤差）是一個特徵，用於描述細胞核的邊界區域的變異程度。
- 它是通過測量每個細胞核邊界與最佳邊界的差異來計算得出的。"area error" 的含義是細胞核邊界區域的變異性。乳腺癌中的細胞核通常具有較大的不規則形狀和邊界模糊性。
- 因此，"area error" 可以提供關於細胞核形狀和邊界的信息。
- "area error" 的值越大，表示細胞核邊界的變異性越大，細胞核的形狀可能更加不規則。這可能與腫瘤的惡性程度有關，因為惡性腫瘤的細胞核通常具有不規則形狀和邊界模糊。