肝　　癌

統計學習與資料探勘 期末報告

2022.12.19

B082040005 高念慈
M102040035 林良朋

CONTENT

Part 01

資料介紹

2022.12.19 統計學習與資料探勘

資料集

Study.ID	Patient.ID	Sample.ID	Ablation.e	AFP.At.Pr	AFP.Norn	AFP.Norn	Diagnosis.	American.	Neoplasm	Neoplasm	American.	American.	Bilirubin.T	Bilirubin.T	Bilirubin.T	Cancer.Ty	Cancer.Ty	Child.pugh	Neoplasm	Neoplasm	Neoplasm	Neoplasm
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	10793	0	8	NA	MX	NX	Stage II	7th	T2	1.5	0	0.2	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	74	0	6	58	MX	NX	NA	6th	T2	0.7	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NA	17	NA	7	51	MX	NX	Stage I	6th	T1	0.4	0.2	1.3	Hepatobili	Hepatocel	C	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	304	0	6	55	MX	NX	Stage I	6th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	6	NA	7	54	MX	NX	Stage I	6th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	2	0	6	64	MX	N0	Stage I	6th	T1	0.6	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	1	0	7	68	MX	NX	Stage I	6th	T1	0.4	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	27600	0	7	64	MX	NX	Stage II	6th	T2	0.4	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	7	0	7	82	MX	NX	Stage II	6th	T2	0.7	0.2	1.3	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	7598	0	8	49	M0	N0	Stage IIIA	6th	T3	0.8	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	YES	3	0	7	58	MX	NX	Stage I	7th	T1	0.5	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	2	0	7	64	MX	N0	Stage I	7th	T1	0.4	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	YES	5640	0	7	45	MX	NX	Stage II	7th	T2	0.3	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	11	0	7	68	MX	N0	Stage I	7th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	11700	0	7	59	MX	N0	Stage I	7th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	114	0	7	68	MX	NX	Stage I	7th	T1	0.3	0.3	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	6	0	7	81	MX	N0	Stage I	7th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	234000	0	7	85	MX	NX	NA	7th	T1	0.3	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	12	0	7	70	MX	N0	Stage I	7th	T1	0.6	0.2	1.3	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	YES	114	0	7	70	MX	NX	Stage II	7th	T2	0.6	0.2	1.3	Hepatobili	Hepatocel	A	NA	NA	NA	NA
lihc_tcga	TCGA-2Y	TCGA-2Y	NO	21	0	7	66	MX	NX	Stage I	7th	T1	0.6	0.2	1.3	Hepatobili	Hepatocel	NA	NA	NA	NA	NA

379 rows × 103 columns

資料來源：[cBioPortal for Cancer Genomics]

主要變數25個

4. Ablation.embolization.tx.adjuvant

消融(剝蝕).栓塞.tx.佐(輔助)劑，(類別，YES/NO)

是破壞腫瘤的不同方法

6. AFP.Normal.Range.Lower.Bound

AFT 正常範圍下界，(int)

AFT : 用作肝細胞癌 (HCC) 篩查、診斷和治療隨訪的腫瘤標誌物

10. Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.Code

(類別，NX，N0，N1)

N category 描述癌症是否已經到達附近的淋巴結

14. Bilirubin.Total

總膽紅素，正常總膽紅素大約在1.2mg/dL以下，(num.)

總膽紅素 : 直接型膽紅素 + 間接型膽紅素

16. Bilirubin.Total.Norm.Range.Upper

膽紅素.總計.正常。範圍.上限，(num.)

膽紅素數值的重要性，不下於肝發炎指數

19. Child.pugh.classification.grade

肝硬化嚴重程度的 Child-Pugh 分級，(類別，A，B，C)

數字越小狀況越好，A : 5-6分 B : 7-9分 C : 10-15分

30. Specimen.collection.method.name

標本.採集.方法.名稱，做了怎麼樣的手術

肺葉切除術、肺段切除術(單) 等，(類別，6種)

32. Disease.Free.Status

無病狀態，(類別，有無復發)

38. Family.History.of.Cancer

癌症家族史，(類別，YES/NO)

40. Fraction.Genome.Altered

分數.基因組.改變，(num.)

可用於查找特定屬性，例如生存期或腫瘤分期的相關性

41. Neoplasm.Histologic.Grade

腫瘤.組織學.分級，(類別，分 1-4 級)

分化良好(由細胞接近正常)至未分化

43. Adjacent.hepatic.tissue.inflammation.extent.type

鄰近肝組織炎症範圍類型，(類別，嚴重，中，無)

主要變數25個

45. History.hepato.carcinoma.risk.factor

歷史.肝癌.風險.因素，(類別，19 類)

飲酒|乙型肝炎|丙型肝炎、非酒精性脂肪肝 等

56. Mutation.Count

稱為突變計數，可作為評判腫瘤突變的潛在指標，(int.)

60. Overall.Survival..Months.

從診斷開始（或治療開始）到結束觀察的時間

(以月為單位)，(num.)

68. Adjuvant.Postoperative.Pharmaceutical.Therapy.Administered.Indicator

術後佐藥物治療指標，(類別，YES/NO)

69. Platelet.count.preresection

在診斷腫瘤後，切除前，血小板的數值，(int)

70. Laboratory.prcoedure.platelet.result.lower.limit.of.normal.value

實驗程序中，血小板正常值的下界，(int)

71. Laboratory.prcoedure.platelet.result.upper.limit.of.normal.value

實驗程序中，血小板正常值的上界，(int)

74. Laboratory.procedure.prothrombin.time.result.value

凝血活酶時間測試，(num.)

可檢查您的血漿是否存在凝血因子異常

76. Laboratory.procedure.international.normalization.ratio.result.lower.limit.of.normal.value

凝血活酶時間測試中，(num.)

以INR（國際標準化比率）正常值下界

98. Tissue.Source.Site

組織來源地，每組織對應一種實驗項目，(類別，36種)

87. Laboratory.procedure.albumin.result.specified.value

白蛋白測定結果值，(num.)

100. Person.Neoplasm.Status

個人腫瘤狀況 ，(類別)

TUMOR FREE : 腫瘤清除/ WITH TUMOR : 有腫瘤

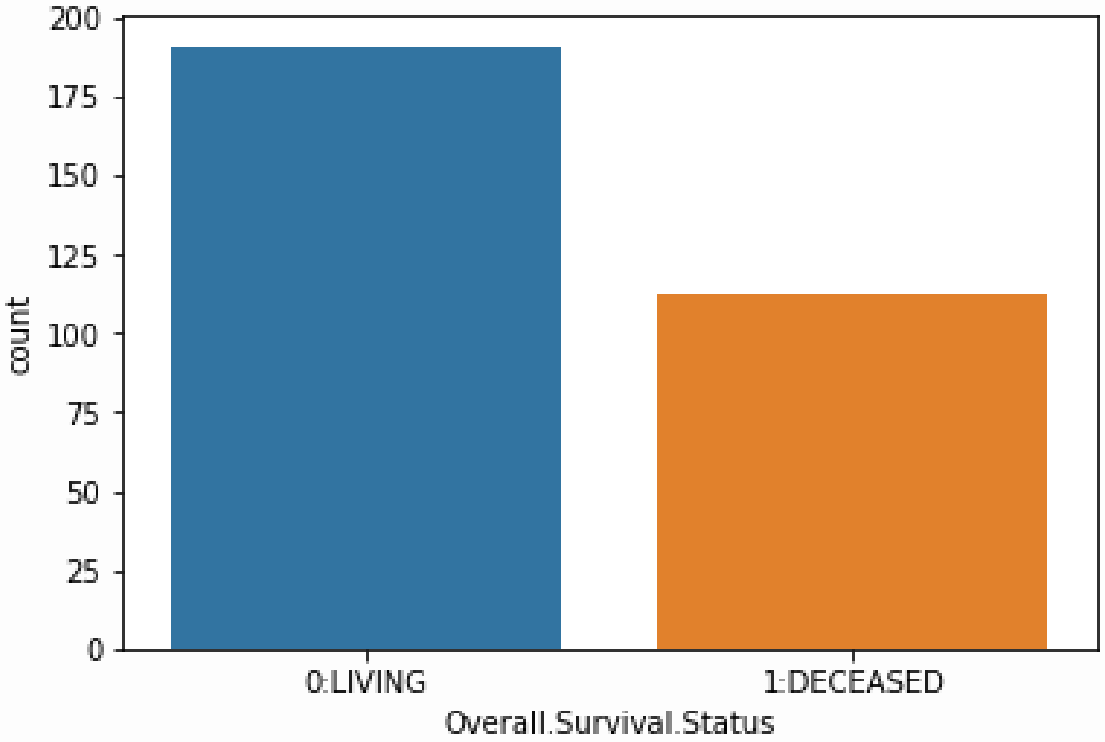
Part 02

EDA/前處理

2022.12.19 統計學習與資料探勘

目標類別不平衡

- 不平衡可能會使預測偏向某一結果



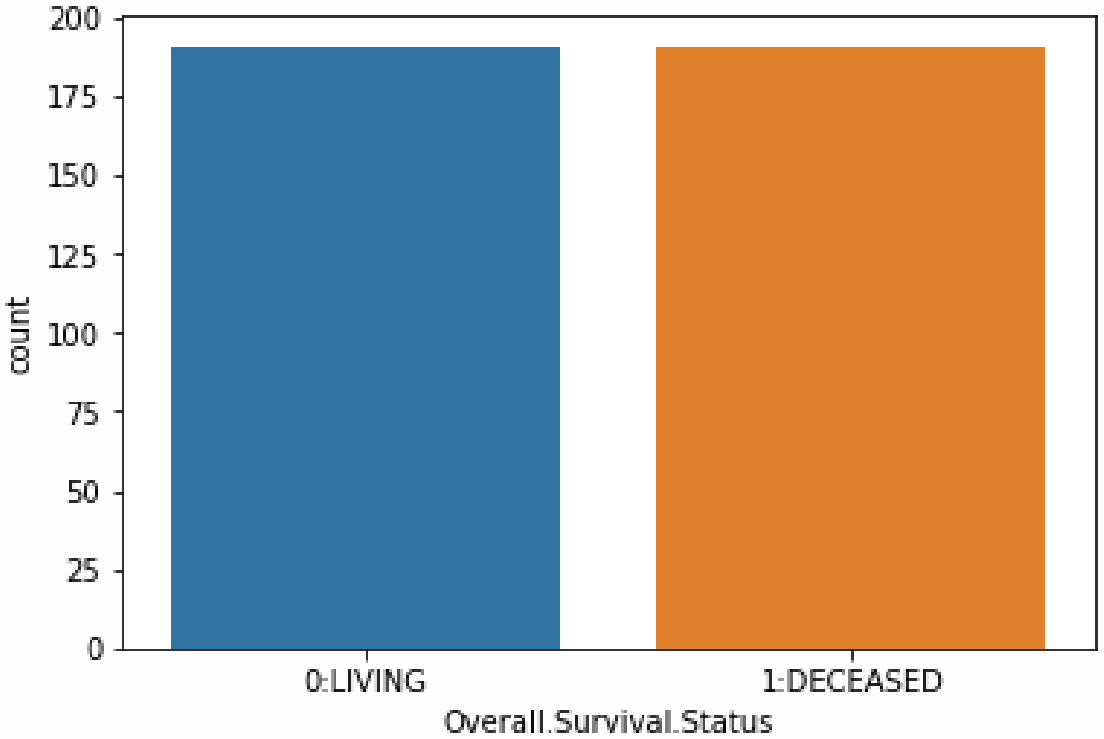
191 V.S. 112

存活狀況

Overall.Survival.Status
0:LIVING
0:LIVING
1:DECEASED
0:LIVING
0:LIVING
...
1:DECEASED
1:DECEASED
1:DECEASED
1:DECEASED
1:DECEASED

Over sampling

- 考慮到樣本不多，目標也不多



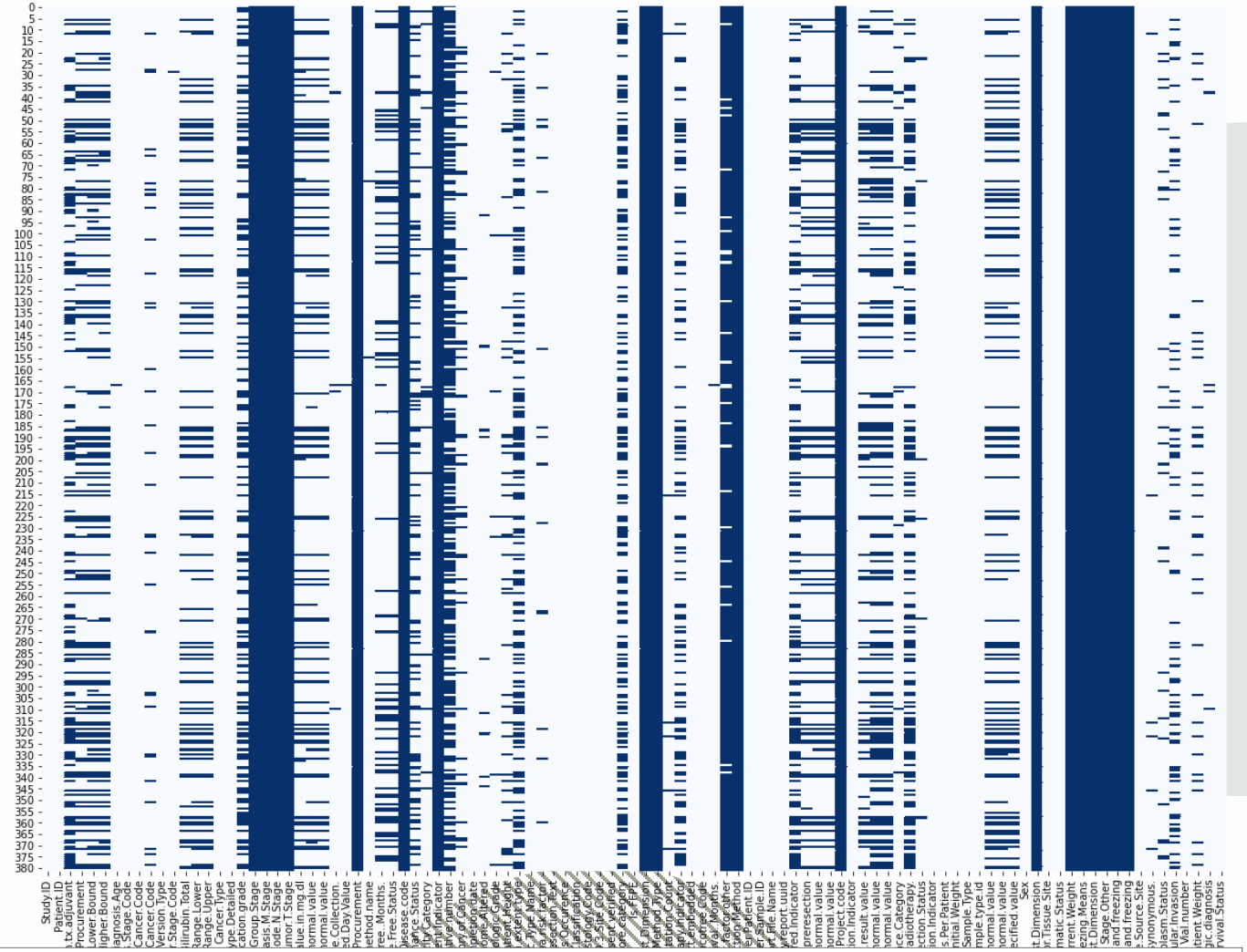
191 V.S. 191

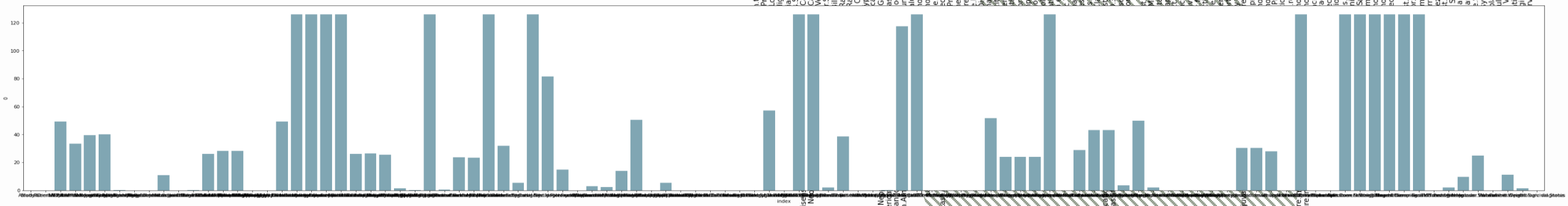
存活狀況

Overall.Survival.Status
0:LIVING
0:LIVING
1:DECEASED
0:LIVING
0:LIVING
...
1:DECEASED
1:DECEASED
1:DECEASED
1:DECEASED
1:DECEASED

缺失值分布/比例

- **大部分模型都不能處理缺失值**





第一輪刪變數

01

重複的、變異性小的
(9 成 : 382*0.9 = 343)

02

全 NA 的

03

資訊不足又有多 NA
怕亂補出事(1個)

第一輪刪了 53 個變數

KDE Plot

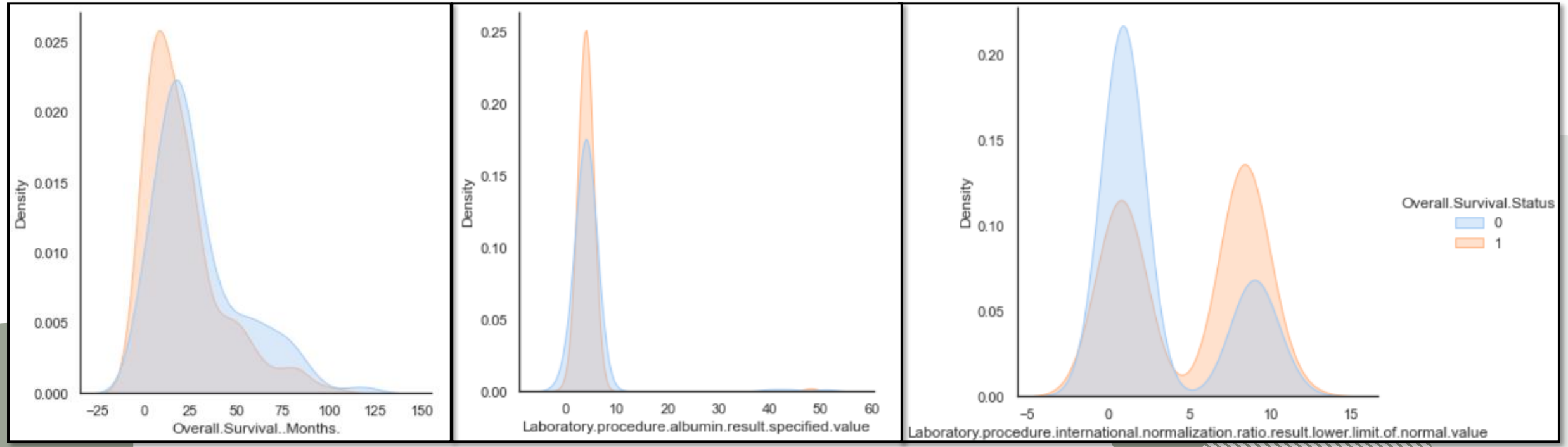
- 比較連續變數

27 個

60. Overall.Survival..Months

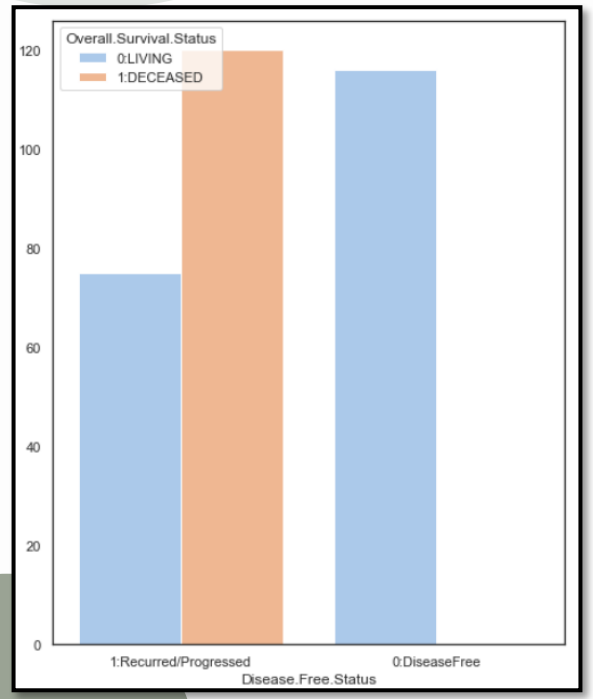
87. Laboratory.procedure.albumin.result.specified.value

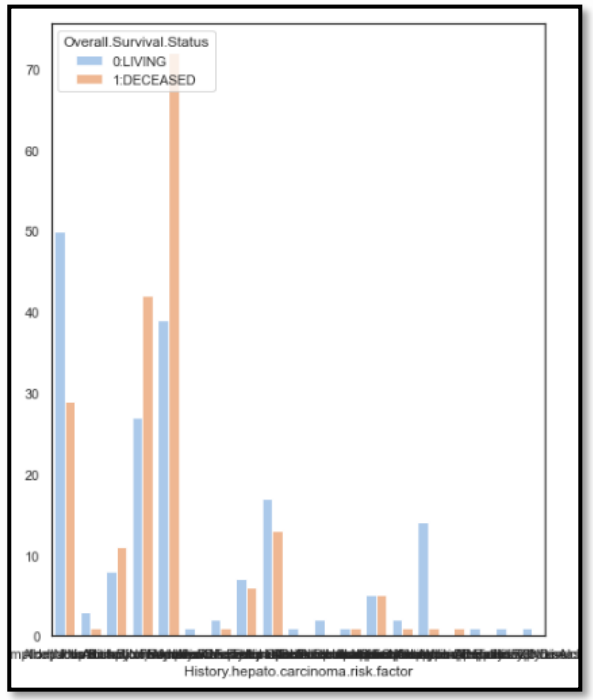
76. Laboratory.procedure.international.normalization.ratio.result.lower.limit.of.normal.value

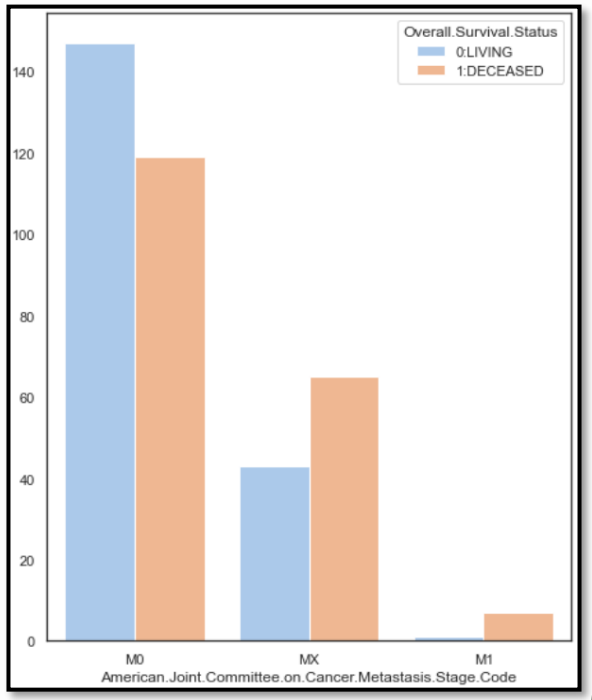


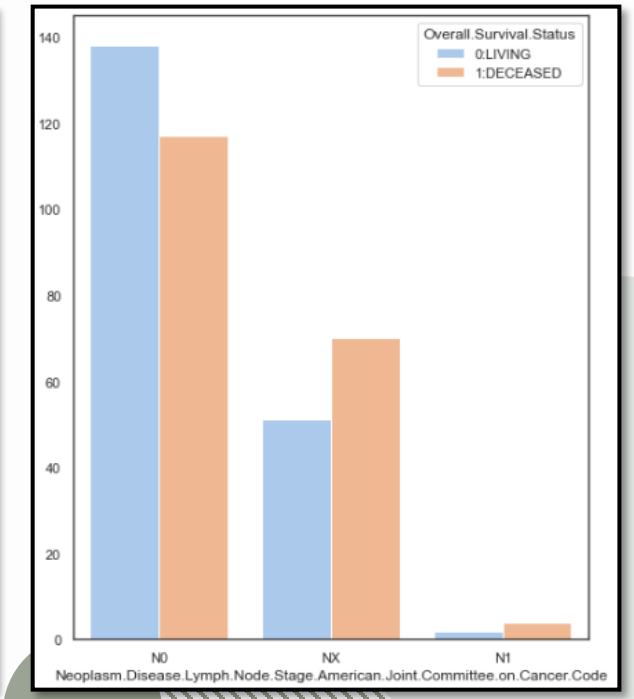
長條圖 - 觀察類別變數

32. Disease.Free.Status
45. History.hepato.carcinoma.risk.factor
9. American.Joint.Committee.on.Cancer.Metastasis.Stage.Code
10. Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.Code



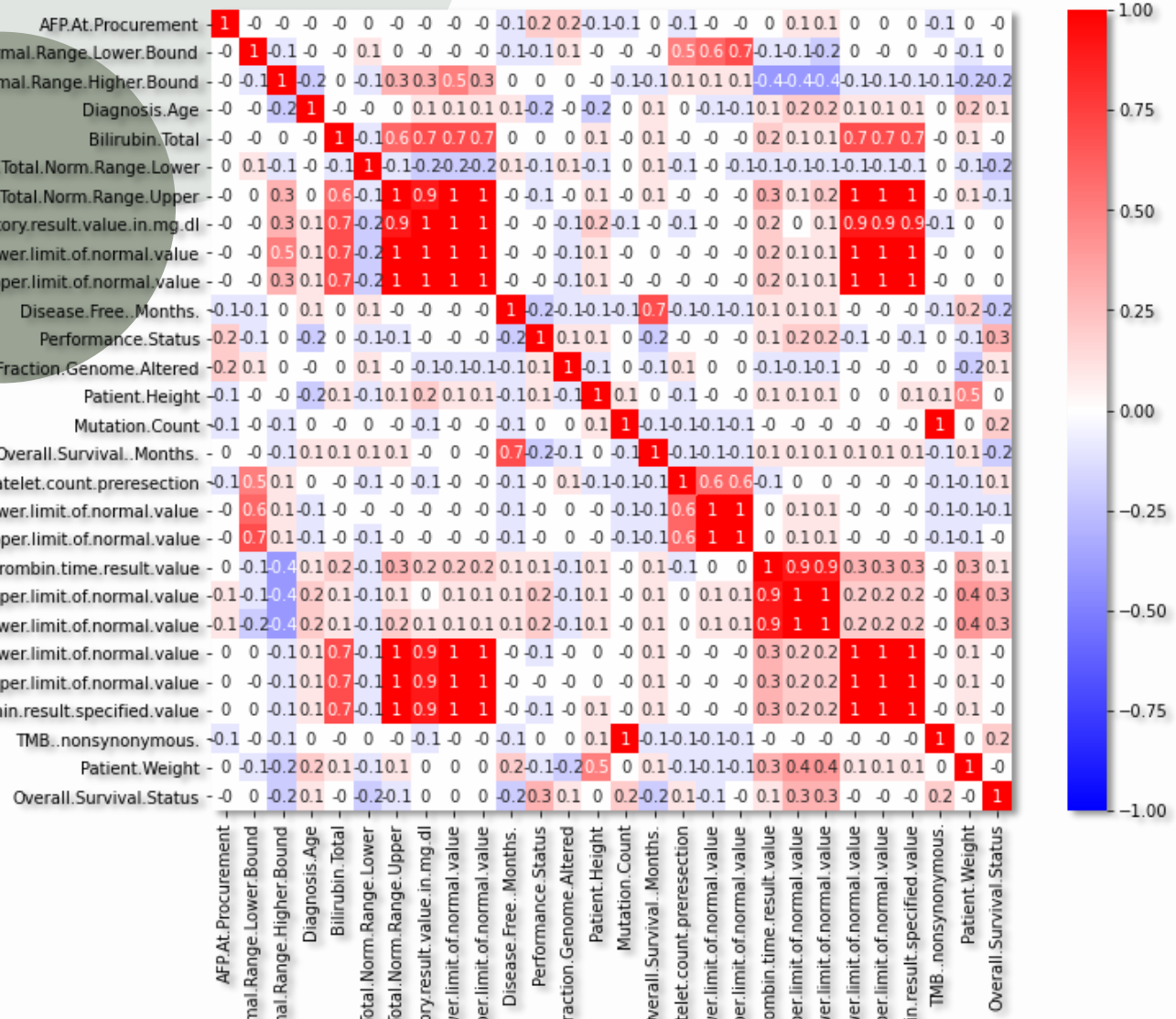






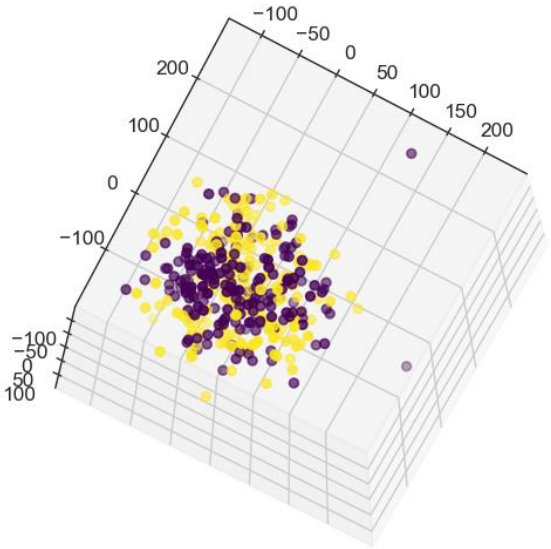
23 個

看線性 / 非線性關係

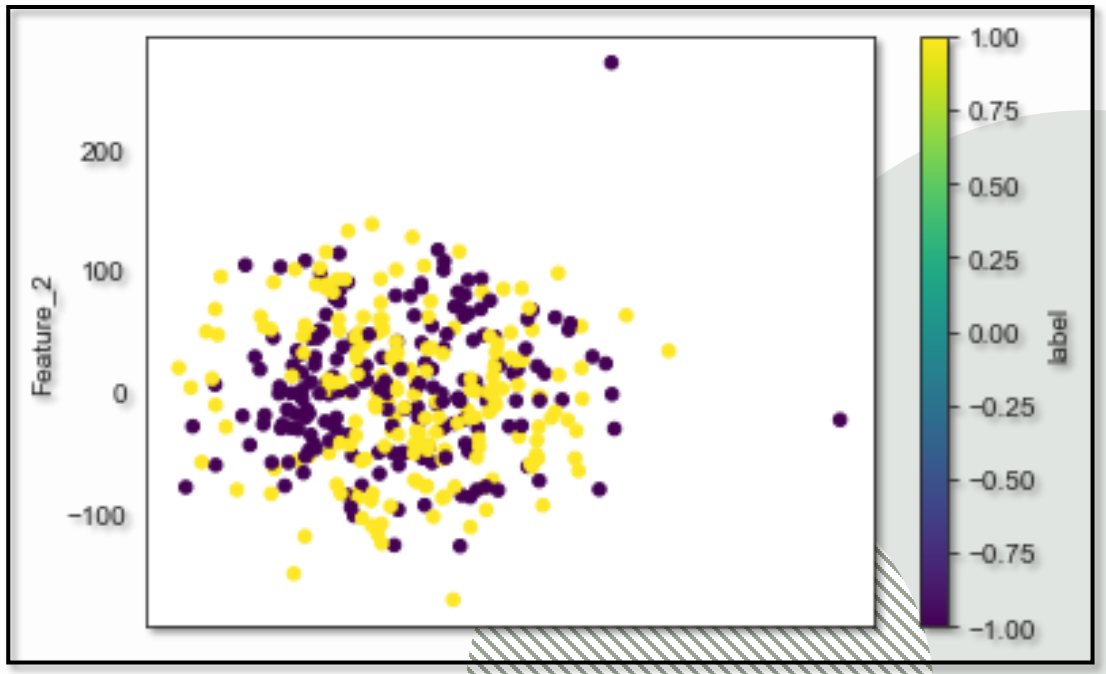


以為能分出特徵相似性，結果好像分不出來

推測是體檢這類型的特徵本來就很相似



T S N E



整理數據(轉類別 / NA)

- 先把類別變成數值(可處理)，再補缺失值

M-estimate Encoder

越大的 m 值會傾向讓整體平均較有影響力
encoding = weight * 類別平均 + (1 - weight) * 整體平均
weight = n / (n + m)

- **NA 視為同類別下去平均**

1. Label Encoding

有 order

&

兩種類的

2. Target Encoding

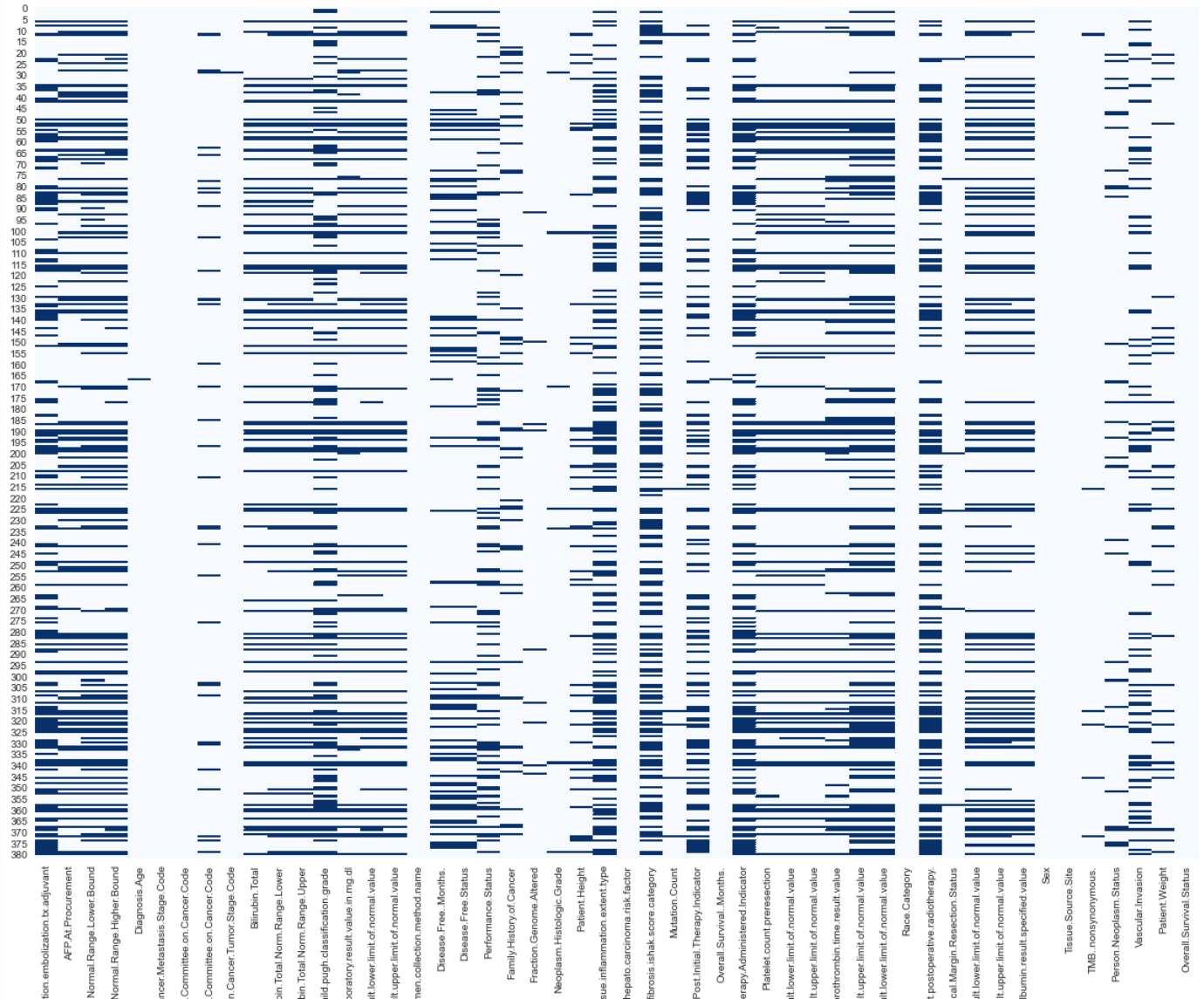
多類別

&

沒次序問題的特徵

目前剩餘缺失值

推測：缺失值純粹為無做單項試驗/人為漏填 等



1. Target Encoding

上頁補的 (4變數)

2. 最鄰近插值法

pandas 內建插值法

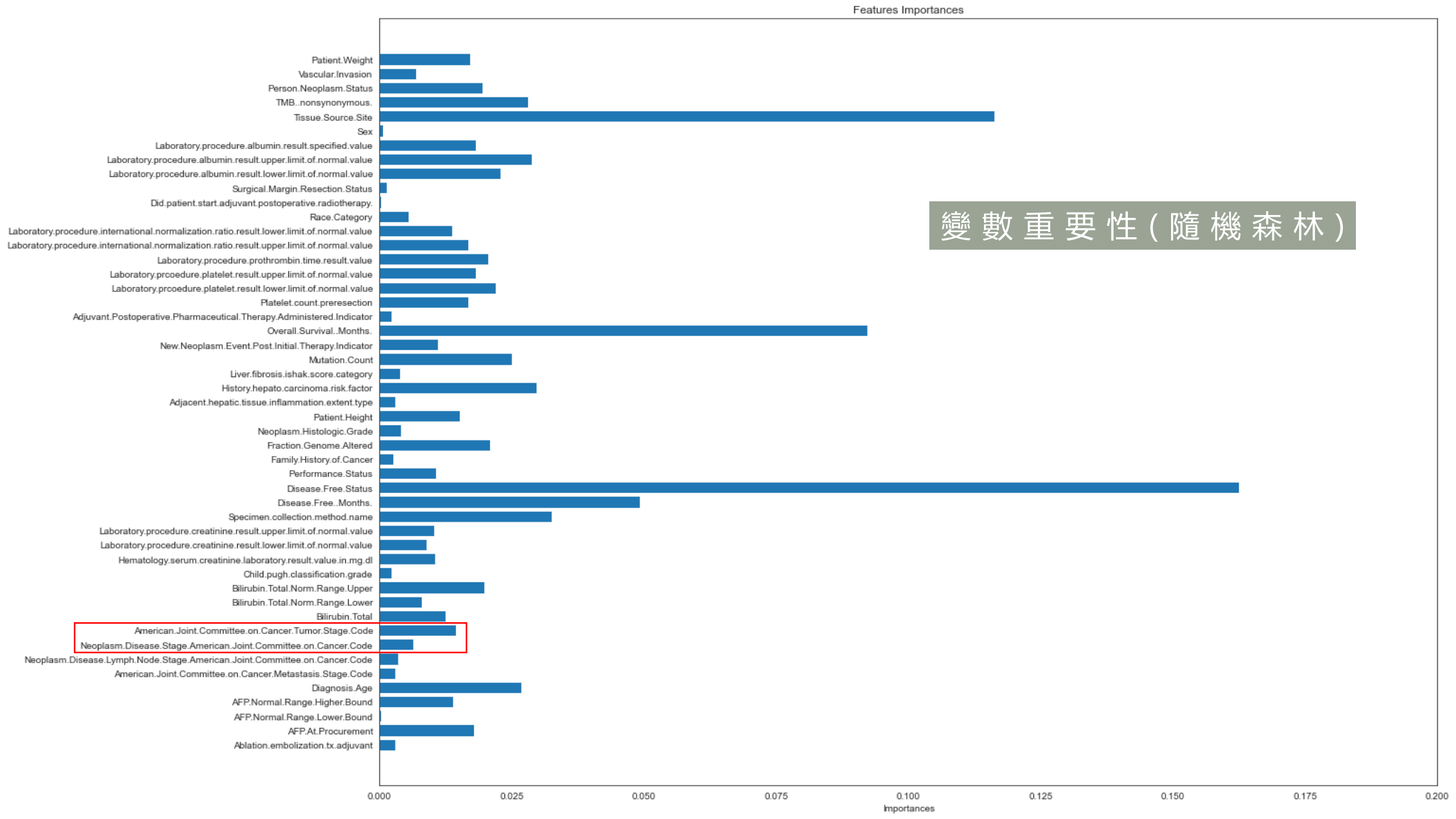
第二輪挑變數

隨機森林 +Backward Selection

382 rows × 50 columns

先挑出前 40 個重要變數
再利用 backward selection
看 ROC 決定最好的變數組合

直接 50 個變數下去
backward selection


Features Importances
Patient.Weight
Vascular.Invasion
Person.Neoplasm.Status
TMB..nonsynonymous.
Tissue.Source.Site
Sex
Laboratory.procedure.albumin.result.specified.value
Laboratory.procedure.albumin.result.upper.limit.of.normal.value
Laboratory.procedure.albumin.result.lower.limit.of.normal.value
Surgical.Margin.Resection.Status
Did.patient.start.adjuvant.postoperative.radiotherapy.
Race.Category
Laboratory.procedure.international.normalization.ratio.result.lower.limit.of.normal.value
Laboratory.procedure.international.normalization.ratio.result.upper.limit.of.normal.value
Laboratory.procedure.prothrombin.time.result.value
Laboratory.prcoedure.platelet.result.upper.limit.of.normal.value
Laboratory.prcoedure.platelet.result.lower.limit.of.normal.value
Platelet.count.preresection
Adjuvant.Postoperative.Pharmaceutical.Therapy.Administered.Indicator
Overall.Survival..Months.
New.Neoplasm.Event.Post.Initial.Therapy.Indicator
Mutation.Count
Liver.fibrosis.ishak.score.category
History.hepato.carcinoma.risk.factor
Adjacent.hepatic.tissue.inflammation.extent.type
Patient.Height
Neoplasm.Histologic.Grade
Fraction.Genome.Altered
Family.History.of.Cancer
Performance.Status
Disease.Free.Status
Disease.Free..Months.
Specimen.collection.method.name
Laboratory.procedure.creatinine.result.upper.limit.of.normal.value
Laboratory.procedure.creatinine.result.lower.limit.of.normal.value
Hematology.serum.creatinine.laboratory.result.value.in.mg.dl
Child.pugh.classification.grade
Bilirubin.Total.Norm.Range.Upper
Bilirubin.Total.Norm.Range.Lower
Bilirubin.Total
American.Joint.Committee.on.Cancer.Tumor.Stage.Code
Neoplasm.Disease.Stage.American.Joint.Committee.on.Cancer.Code
Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.Code
American.Joint.Committee.on.Cancer.Metastasis.Stage.Code
Diagnosis.Age
AFP.Normal.Range.Higher.Bound
AFP.Normal.Range.Lower.Bound
AFP.At.Procurement
Ablation.embolization.tx.adjuvant
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importances
變數重要性(隨機森林)

特徵選擇

- 因為資料不大，試試看原 50 個變數跑的結果，結果更好，變數也較直覺、好解釋

變數剩 9 個 ROC 分數 : 0.9287626962142198

```
['Laboratory.procedure.creatinine.result.upper.limit.of.normal.value',
'AFP.At.Procurement', 'Person.Neoplasm.Status', 'Laboratory.procedure.p
rothrombin.time.result.value', 'Mutation.Count', 'TMB..nonsynonymous.',
'History.hepato.carcinoma.risk.factor', 'Overall.Survival..Months.', 'T
issue.Source.Site']
```

變數 25 個 ROC 分數 : 0.9303714752468215

```
['Ablation.embolization.tx.adjuvant', 'AFP.Normal.Range.Lower.Bound',
'Neoplasm.Disease.Lymph.Node.Stage.American.Joint.Committee.on.Cancer.C
ode', 'Bilirubin.Total', 'Bilirubin.Total.Norm.Range.Upper', 'Child.pug
h.classification.grade', 'Specimen.collection.method.name', 'Disease.Fr
ee.Status', 'Family.History.of.Cancer', 'Fraction.Genome.Altered', 'Neo
plasm.Histologic.Grade', 'Adjacent.hepatic.tissue.inflammation.extent.t
ype', 'History.hepato.carcinoma.risk.factor', 'Mutation.Count', 'Overal
l.Survival..Months.', 'Adjuvant.Postoperative.Pharmaceutical.Therapy.Ad
ministered.Indicator', 'Platelet.count.preresection', 'Laboratory.prcoe
dure.platelet.result.lower.limit.of.normal.value', 'Laboratory.prcoedur
e.platelet.result.upper.limit.of.normal.value', 'Laboratory.procedure.p
rothrombin.time.result.value', 'Laboratory.procedure.international.norm
alization.ratio.result.lower.limit.of.normal.value', 'Laboratory.proced
ure.albumin.result.specified.value', 'Tissue.Source.Site', 'Person.Neop
lasm.Status', 'Patient.Weight']
```

Part 03

模　型

2022.12.19 統計學習與資料探勘

模型流程

01

運用10FoldCV找到樹的最適深度

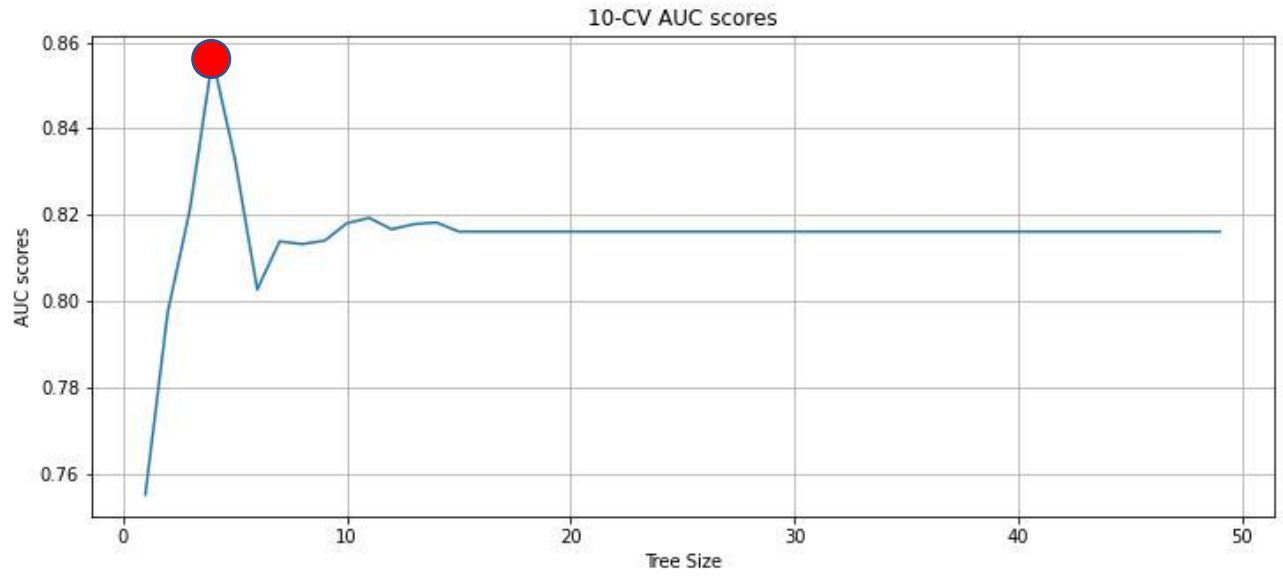
02

利用LightGBM, RandomForest, CatBoost

03

畫出混淆矩陣及ROC Curve

運用10FoldCV找到樹的最適深度



以AUC作為Scoring找到以tree_size=4為最適深度(以決策樹分類器+10FoldCV)

利用LightGBM, RandomForest,CatBoost

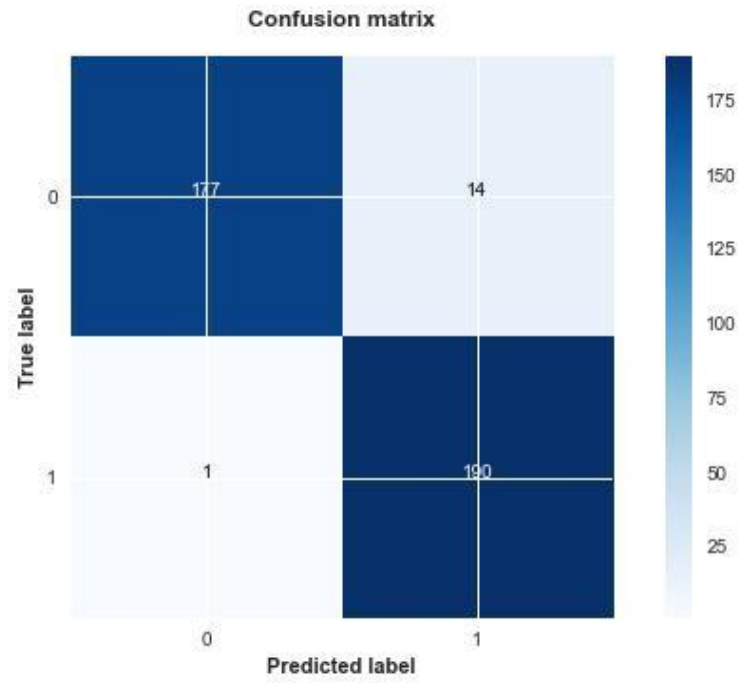
LightGBM	RandomForest	CatBoost
'boosting_type': 'gbdt' 'objective': 'binary' 'metric': 'auc' 'learning_rate':0.1 'num_leaves':30 'max_depth': 4	n_estimators:100, 'criterion': ['entropy'], 'max_depth':[4],	'depth' : [4], 'learning_rate' : [0.1], 'iterations':range(1,100)
n_estimators= 56	n_estimators= 100	n_estimators= 86

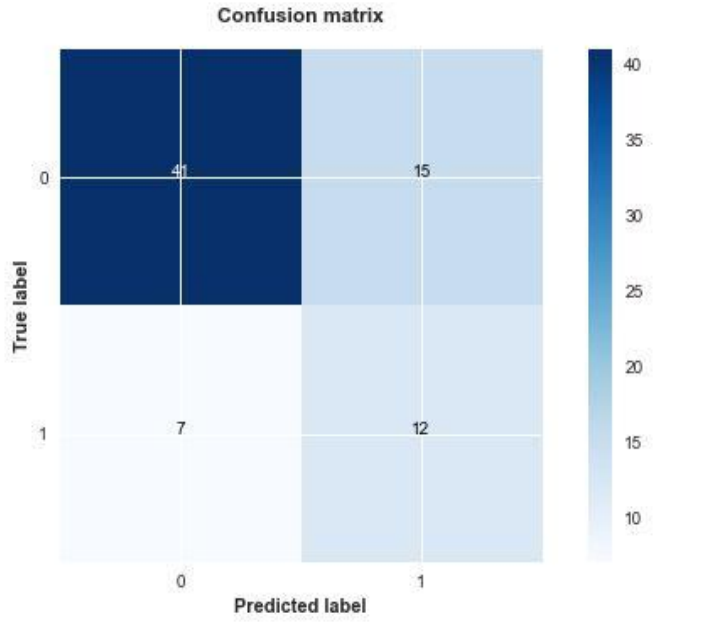
這裡的n_estimators為樹的顆數

LightGBM 在訓練集與測試集上的混淆矩陣

Threshold=0.4

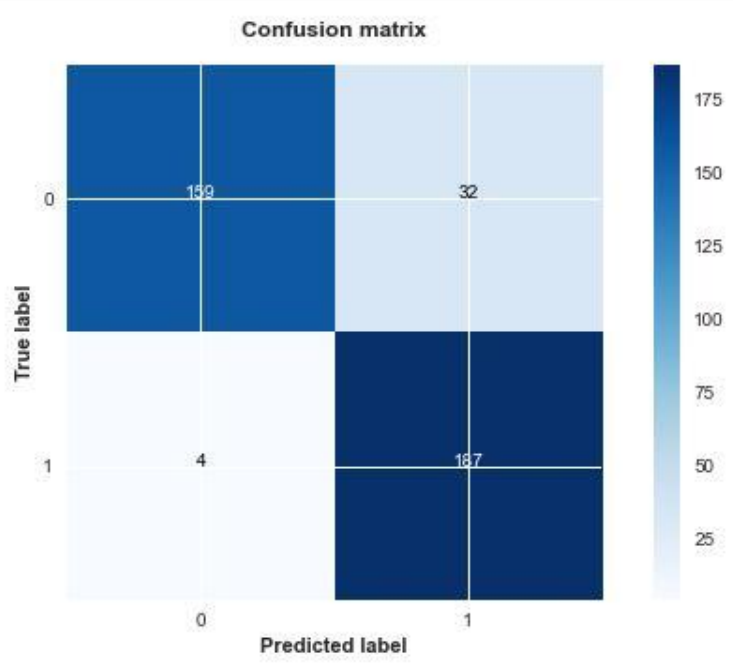
Training Set	Testing set
Recall: 0.994764	Recall: 0.631579

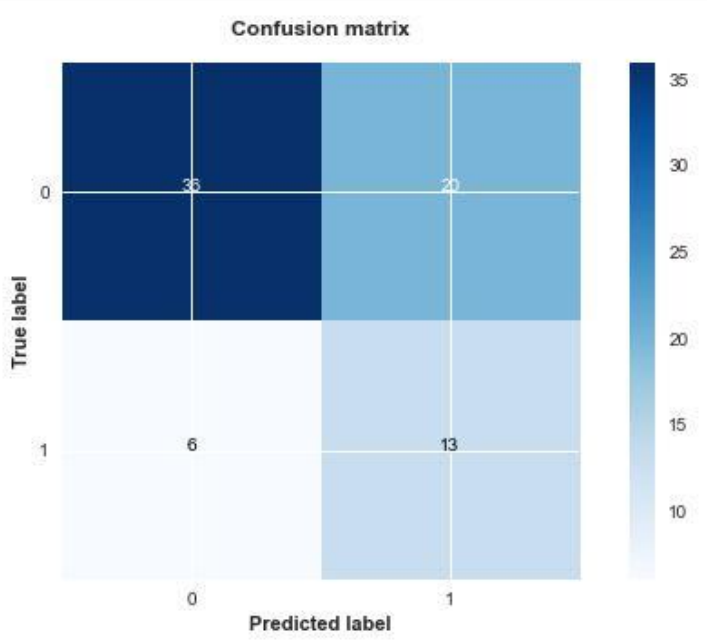




Random Forest 在訓練集與測試集上的混淆矩陣

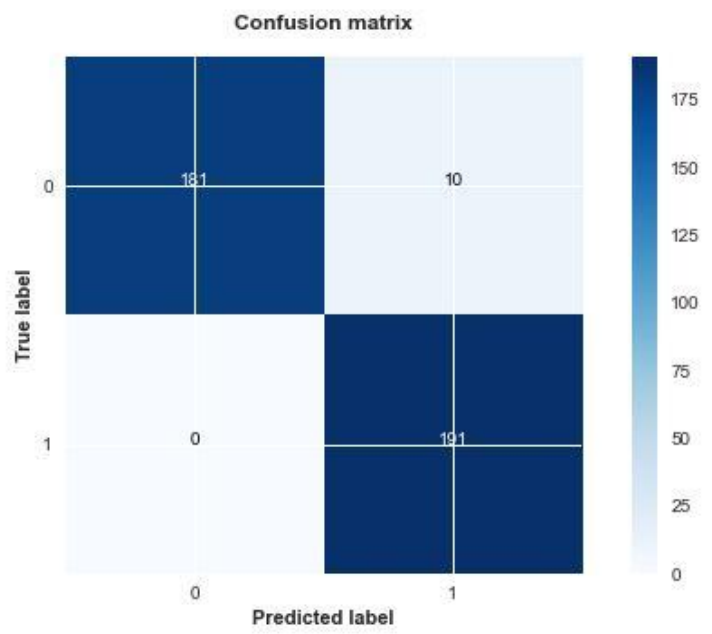
Training Set	Testing set
Recall: 0.979058	Recall: 0.684211



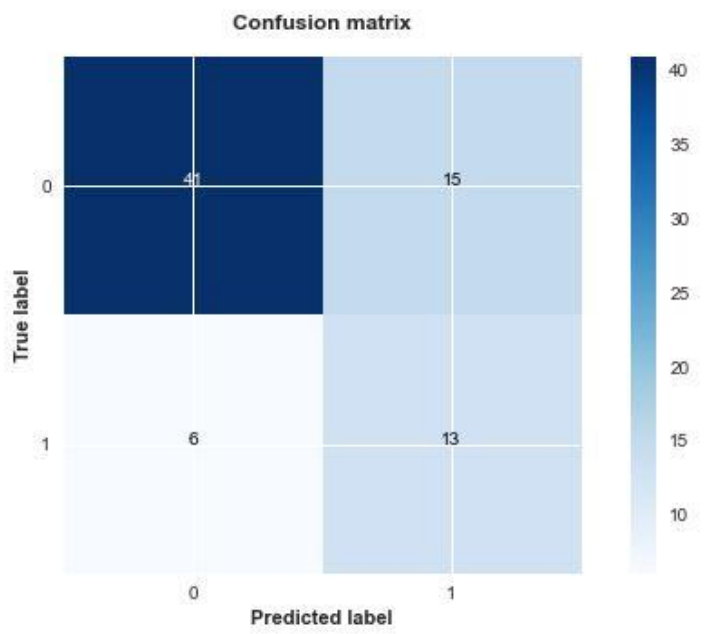


CatBoost 在訓練集與測試集上的混淆矩陣

Training Set
Recall: 1.0



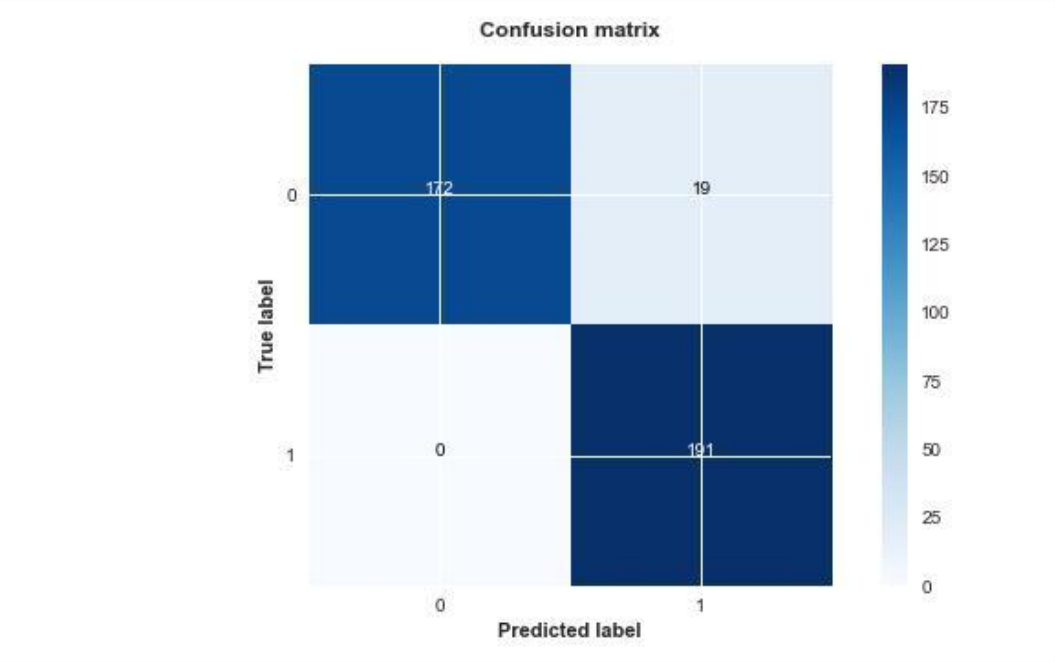
Testing set
Recall: 0.684211



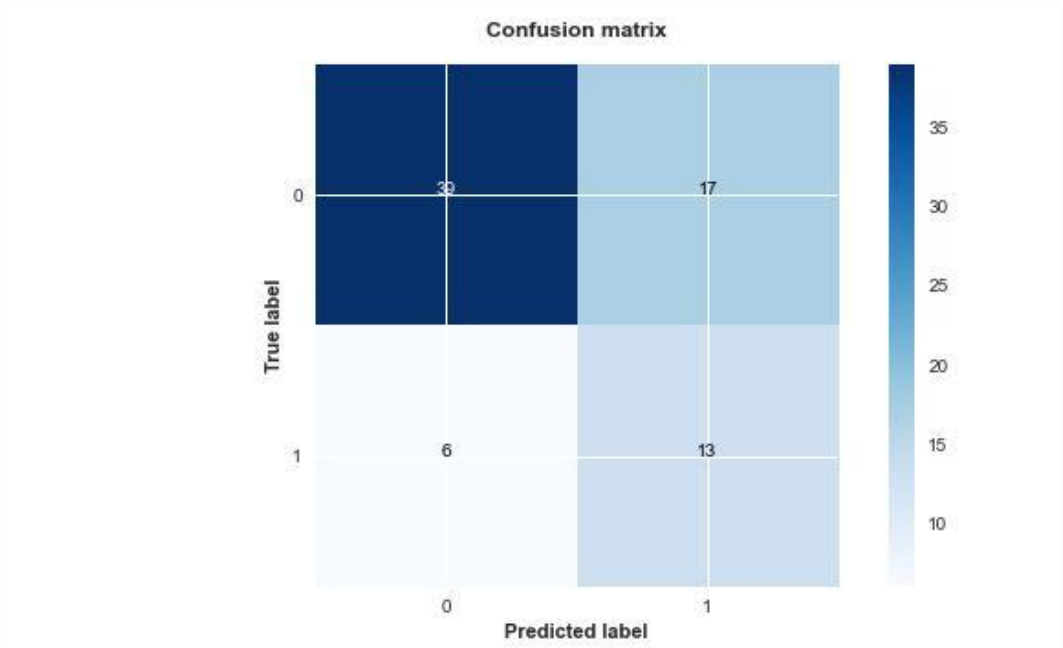
Ensemble 在訓練集與測試集上的混淆矩陣

三個模型取平均

Training Set
Recall: 1.0

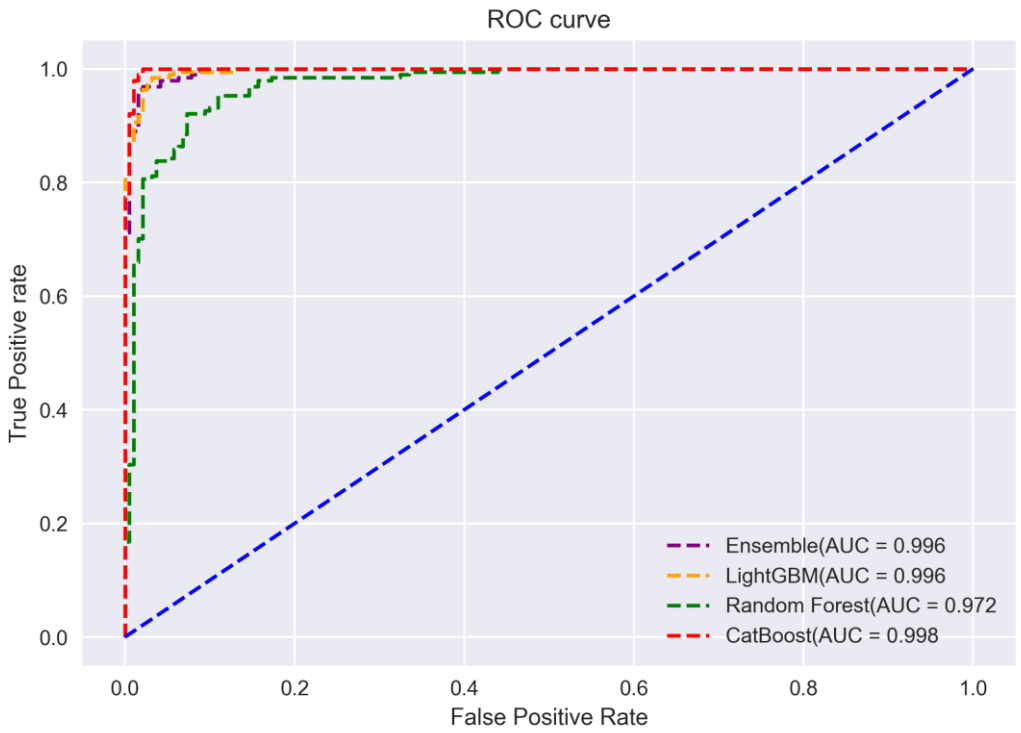


Testing set
Recall: 0.684211

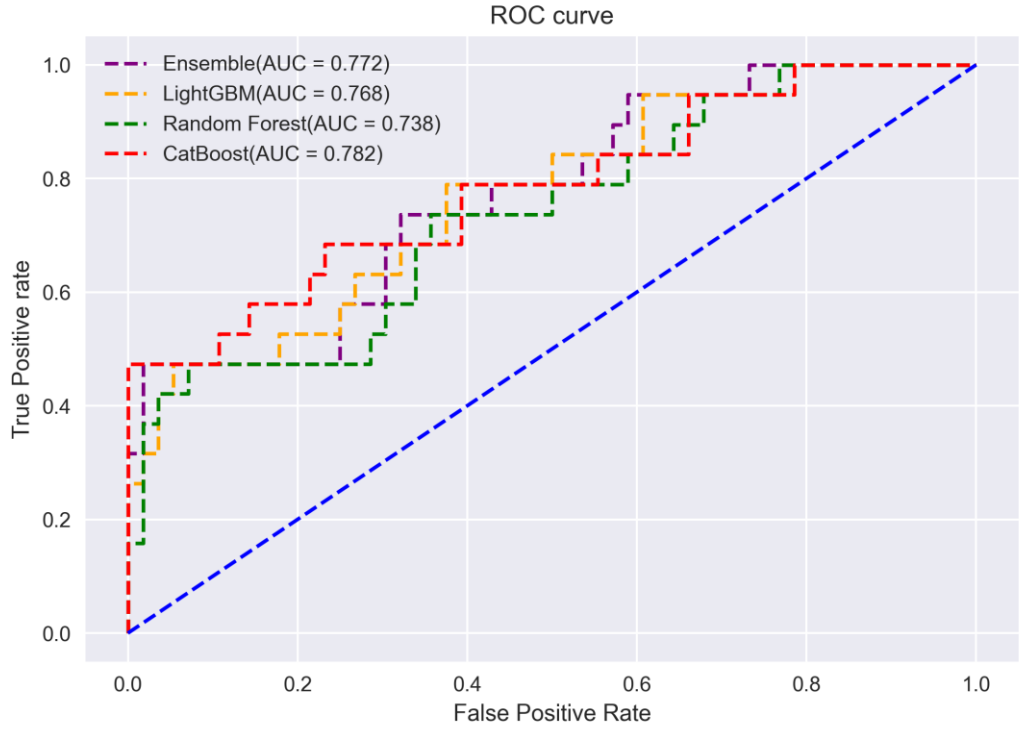


ROC Curve 在訓練集與測試集上

Training Set



Testing set



Part 04

總　結

2022.12.19 統計學習與資料探勘

比較所有模型的AUC score& Recall

AUC score

	Logistic	LightGBM	RandomForest	**CatBoost**	Ensemble
Training	0.686001	0.996327	0.972040	**0.998246**	0.995532
Testing	0.646617	0.767857	0.737782	**0.781955**	0.771617

recall

	LightGBM	RandomForest	**CatBoost**	Ensemble
Training	0.994764	0.979058	1.000000	1.000000
Testing	0.631579	0.684211	0.684211	0.684211

以CatBoost有最高的分數

Part 05

參考資料

2022.12.19 統計學習與資料探勘

2. 過採樣 — 版本 0.10.0 (imbalanced-learn.org)

隨機過採樣器 — 版本 0.10.0 (imbalanced-learn.org)

[機器學習二部曲] Python實作—資料預處理 : 如何將類別型特徵自動轉換成數值型? LabelEncoder ! – PyInvest (pyecontech.com)

機器學習筆記-Target Encoding.對於數值特徵（numerical... | by 黃柏竣 |中等 (medium.com)

M 估計 — 類別編碼器 2.5.1.post0 文檔 (scikit-learn.org)

熊貓。DataFrame.interpolate — pandas 1.5.2 文檔 (pydata.org)

[改善資料品質]Part-2 面對缺漏值的對策 - iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天 (ithome.com.tw)

後向特徵消除及其實現 (analyticsvidhya.com)

Python - 如何使用 t-SNE 進行降維 | Mortis

謝　謝　大　家

2022.12.19 統計學習與資料探勘 期末報告