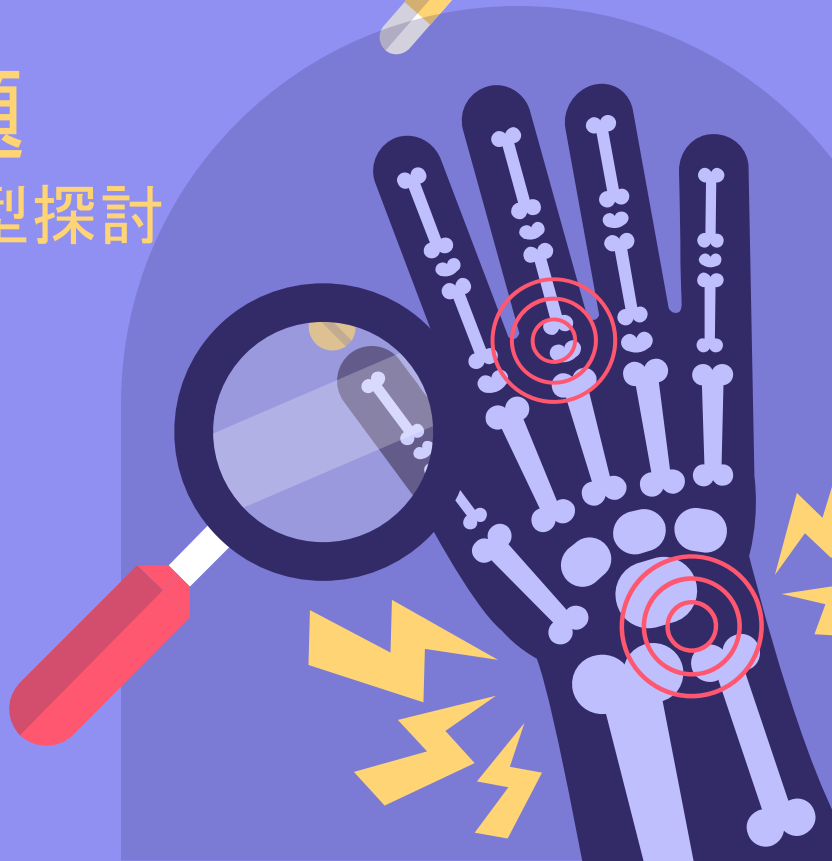
# 資料科學生物資訊專題

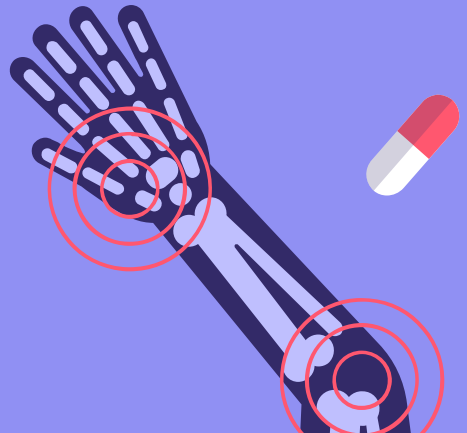
## 利用SNPs預測RA患病情形之模型探討


| | | |
|---|---|---|
| RE6111032 | 數據所碩一 | 曾以諾(組長) |
| N96104399 | 工科所碩二 | 楊育維 |
| L46111158 | 地科所碩一 | 郭人豪 |

# TABLE OF CONTENTS





# 01.

# 目標

# Identify problem

## 目標

嘗試建立一種用SNPs去預測是否罹患類風溼性關節炎的模型，我們主要關注模型recall與f1的表現。

# 02.

# 領域知識介紹

Identify problem

# 什麼是類風溼性關節炎?


健康2.0

類風濕性關節炎是一種**慢性且會不斷進展的發炎性關節疾病**，全身的關節都可能會受到影響。類風濕性關節炎可能發生在任何年紀，但又以中年女性罹病的可能性最高。

會得到這個病的原因是身體的免疫系統出了問題,產生了很多會破壞身體組織器官的自體抗體及發炎物質(如細胞激素)。這些抗體或發炎物質除了會攻擊全身的關節以外，還有可能侵犯其他的器官，如肝臟、脾臟、心臟、肺臟、血液系統、神經系統、淋巴系統等。

藉由抽血檢驗類風濕因子(RF) 與抗環瓜氨酸抗體(anti-CCP)，類風濕性關節炎病人可分成血清陰性及血清陽性兩群。

https://www.cmuh.cmu.edu.tw/HealthEdus/Detail?no=4769

Identify problem

# GWAS是什麼?

全基因組關聯研究(GWAS) 使用高通量基因組技術快速掃描大量受試者的整個基因組，以發現與特徵或疾病相關的遺傳變異。了解複雜疾病的遺傳結構在很大程度上依賴於疾病相關變異的發現和表徵，例如單核苷酸多態性(SNP) 和拷貝數變異(CNV)。

GWAS是什麼?

Identify problem

# 什麼是SNP基因分型?

SNP(單核苷酸多態性)基因分型是分析基因組序列變異的技術。

只靠一個鹼基替換為其他鹼基，並且必須至少存在於1% 的人群中，就稱為SNP。

人類的基因組中，大約500~1000個鹼基中有一個SNP，而個人或人種之間的序列會出現差距。平均而言，每一個人約有300萬個SNPs。

參考書:圖解人類基因組的介紹

Identify problem

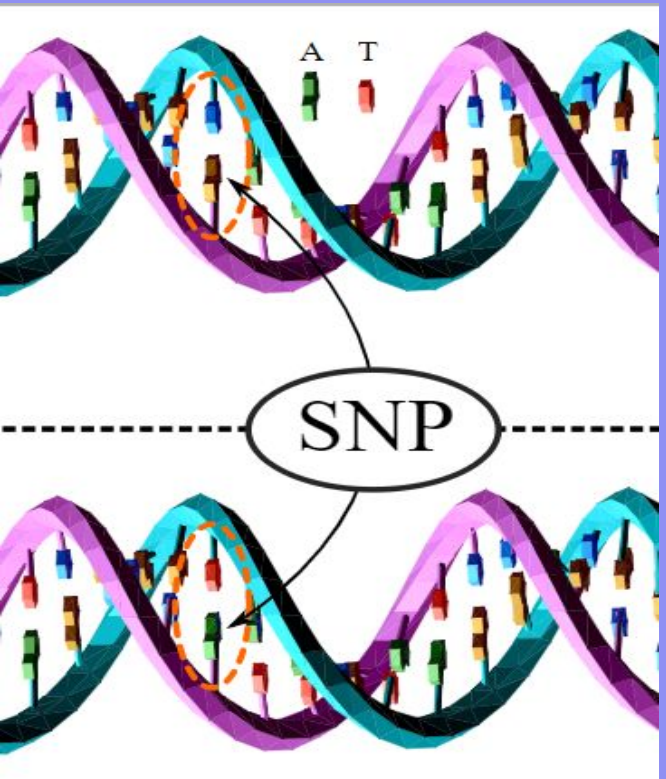
# 什麼是SNP基因分型?


| | | |
|---|---|---|
| cSNP | coding SNP | 在表現序列上，具有使蛋白質的氨基酸序列 產生變化的作用 |
| rSNP | regulatory SNP | 在基因的控制領域中，調節基因的作用 |
| iSNP | intron SNP | 在內子上 |
| sSNP | silent SNP | 在表現序列上，但是不具有使蛋白質的氨基酸序列 產生變化的作用 |
| gSNP | genome SNP | 在基因外的基因組的部分 |

參考書:圖解人類基因組的介紹

Identify problem

# 為什麼要對SNP進行分析?




單核苷酸多態性(single nucleotide polymorphism,SNP)指存在於某一(些)群體、正常個體的基因組DNA中單鹼基的序列差異，是生物體基因組中存在的最簡單也是最廣泛的多態形式，是人類遺傳基因的多態性在遺傳信息上的本質表現，90%以上的遺傳信息是由SNP所引起的。具有分佈廣、數目多，相對穩定的存在於染色體上等特點。SNP的產生可能會引起蛋白質表型的改變，所以**SNP也是影響人類體質的關鍵**，使人可能特別易於或特別不易患上某些疾病，或對於治療藥物的反應性有所差異。當前人們正致力於疾病基因背景的探索，SNP由於其在基因遺傳研究方面所具備的優勢，成為人類基因組計劃完成之後又一研究新熱點。

SNP

https://cdmd.cnki.com.cn/Article/CDMD-10299-2009102219.htm

Identify problem

# 為什麼要對SNP進行分析?

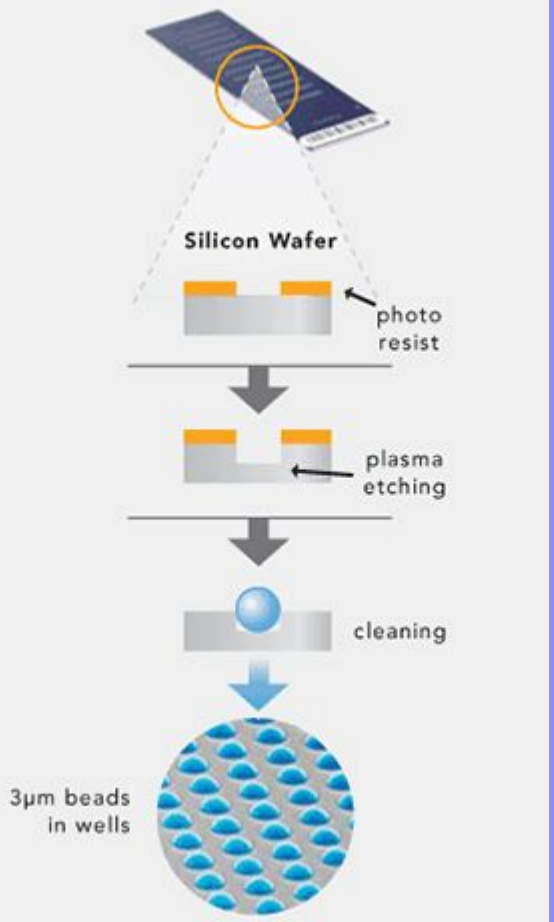
人類白細胞抗原(human leucocyte antigen,HLA)基因系統是人類主要組織相容性複合體(major histocompatibility complex, MHC)，是人類基因組中與免疫功能相關最密切的一個區域，由於該基因和免疫機制聯繫密切且具有豐富的多態性，決定了其成為探索免疫性疾病基因遺傳背景的一個重點研究靶位。目前，已有研究表明HLA-Ⅱ類基因中多態性最豐富的**HLA-DRB1和DQB1的等位基因與類風濕關節炎的發生相關**，但這兩個基因的SNP位點如何與RA相關聯並不清楚。

https://cdmd.cnki.com.cn/Article/CDMD-10299-2009102219.htm

Identify problem

使用微陣列從唾液樣本中讀取 DNA代碼

# 資料量測方式和設備介紹




利用 illumina 550K 晶片取得全基因體SNPs表現資料。

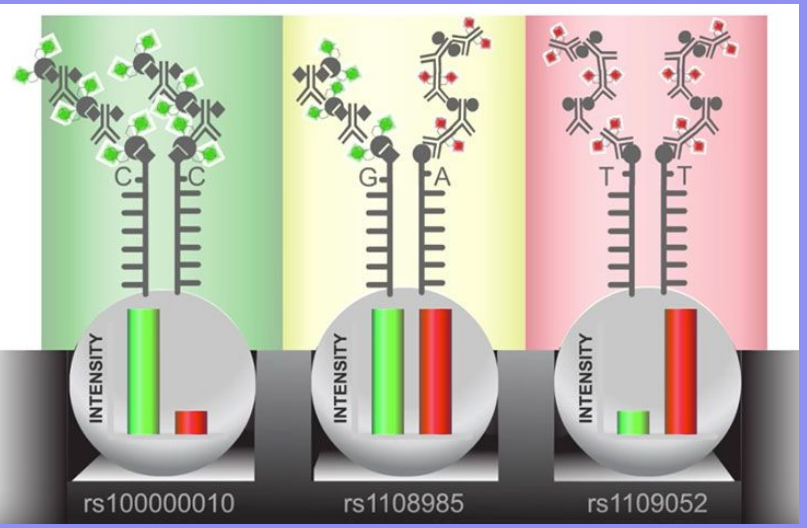
Illumina 晶片由2部分組成，玻璃基片和微珠。玻璃基片和普通的載玻片大小差不多，上面用光蝕刻出一個個小孔給微珠作為容器，一個小孔放一個微米級大小的微珠。**每個微珠表面都偶聯了一種序列的DNA片段**，每個微珠上有幾十萬個片段，一個珠子上的片段都是同一種序列。

Illumina Microarrays

https://youtu.be/FP4dsgwhygM

Identify problem

# 資料量測方式和設備介紹




利用 illumina 掃描儀掃描螢光標記，透過螢光標記強度的不同和探針設計不同可以區分出不同的SNPs。

https://youtu.be/FP4dsgwhygM

Identify problem

# 相關問題分析方法文獻回顧

1. 李怡然 et al. (2015), 基於SNP互作識別類風濕性關節炎的潛在致病基因
   a. 採用Logistic迴歸模型來檢測互作SNP與疾病的顯著相關性。結果分析疾病相關 SNP對，獲得34個潛在的類風濕性關節炎疾病風險基因，其中 5個為類風濕性關節炎已知疾病基因，16個已有文獻證實為疾病風險基因。結論通過對 SNP間上位效應的研究可獲得類風濕性關節炎新的潛在疾病基因。
   b. 數據來自WTCCC和UCSC
   c. 使用p-value判斷相關性，p-value<5x10^(-5)稱為與疾病顯著相關
2. Jawaheer et al. (2003), Screening the Genome for Rheumatoid Arthritis Susceptibility Genes
   a. 許多非 HLA 基因座已顯示與類風濕性關節炎（RA）相關的證據 (P < 0.05)。有證據顯示在1號(1p13, 1q43)、6號(6q21)、10號(10q21)、12號(12q12)、17號(17p13)、18號(18q21)染色體上存在基因連鎖。
   b. 數據來自NARAC招募的受試者
   c. 判斷標準用p-value
3. Newton et al. (2004), A review of the MHC genetics of rheumatoid arthritis
   a. 類風濕性關節炎是一種常見的複雜遺傳疾病，儘管有重要的遺傳因素，但 除 HLA-DRB1 外，沒有其他基因被明確證明與該疾病有關。

Identify problem

# 相關問題分析方法文獻回顧

4. Huizinga et al. (2005), Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins
   a. 目標:類風濕關節炎（RA）的主要遺傳風險因素 HLA 區域已被人們所知 25 年。先前的研究表明，在 RA 人群中，攜帶共享表位（SE）的 HLA-DRB1 等位基因與針對環狀瓜氨酸肽的抗體（Anti-CCP 抗體）之間存在關聯。作者進行了這項研究，首次比較了健康人群中的 SE 等位基因頻率與有或沒有 Anit-CCP 抗體的 RA 患者中的 SE 等位基因頻率。
   b. 結論：編碼 SE 的 HLA-DRB1 等位基因對以瓜氨酸肽抗體為特徵的疾病具有特異性，表明這些等位基因與 RA 本身不相關，而是與特定表型相關。
   c. 數據來自NARAC
   d. 判斷是否存在關聯之標準為 p-value

# 03.

# 資料前處理與視覺化

Prepare data

# 資料集介紹

1. 來自北美類風濕性關節炎聯盟(NARAC)

2. 總樣本數N=2062, 病例數N=868, 對照組(健康)N=1194

3. 資料欄位ID, Affection, Sex, DRB1_1, DRB1_2, SENum, SEStatus, AntiCCP, RFUW, SNP位點名稱(545080筆，包含粒線體DNA位點)

Prepare data

# 資料欄位介紹

**ID**

樣本流水編號

**Affection**

是否患有
類風溼性關節炎
0=unaffected,
1=affected

**Sex**

性別
F=Female, M=Male

**DRB1_1**

HLA-DRB1 allele 1

**DRB1_2**

HLA-DRB1 allele 2

**SENum**

共享表位的等位基因數量
NN=0, SN=1, SS=2

Prepare data

# 資料欄位介紹

## SEStatus

是否是共享表位的等位基因
yes or no

## AntiCCP

抗環瓜氨酸抗體

## RFUW

類風溼性關節炎因子
(IgM RF)

## SNPs

545080 SNP-genotype
from the Illumina 550K chip

Prepare data

# 資料前處理

1. **轉換類別型資料**
2. **NaN資料轉換成-1**
3. **SNP資料轉換成數字編碼**

| 鹼基 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 1 | 3 | 6 |
| C | 1 | 2 | 4 | 7 |
| G | 3 | 4 | 5 | 8 |
| T | 6 | 7 | 8 | 9 |

https://www.twblogs.net/a/5b8e03642b71771883423fda

Prepare data

# SNP位點介紹

每人平均含有的遺失值個數:每人含有的遺失值個數(共2062筆)相加再除以2062

每人平均遺失值比例:每人平均含有的遺失值個數除以SNP總數(545080)

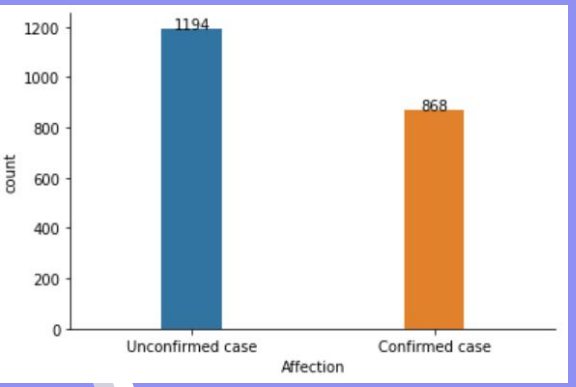
| 每人平均含有的遺失值個數 | 每人平均遺失值比例 |
| --- | --- |
| 3995.592628516004(個) | 0.007330286615755493 |

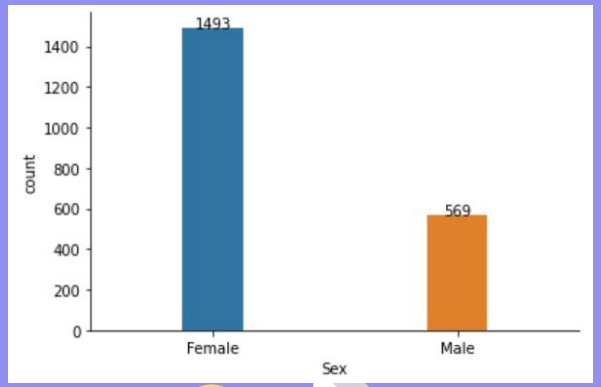
# Prepare data

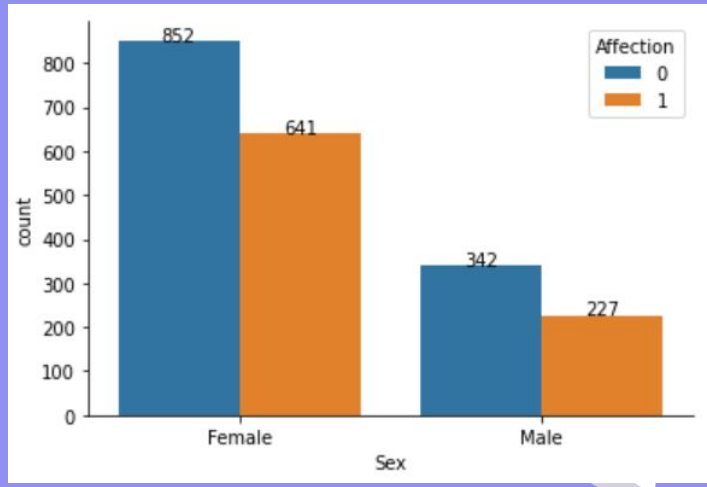
## 資料視覺化

1. **欄位[Sex]**
   **女生1493位，男生569位**

2. **欄位[Affection]**
   **確診類風溼性關節炎人數868人**
   **未確診類風溼性關節炎人數1194人**



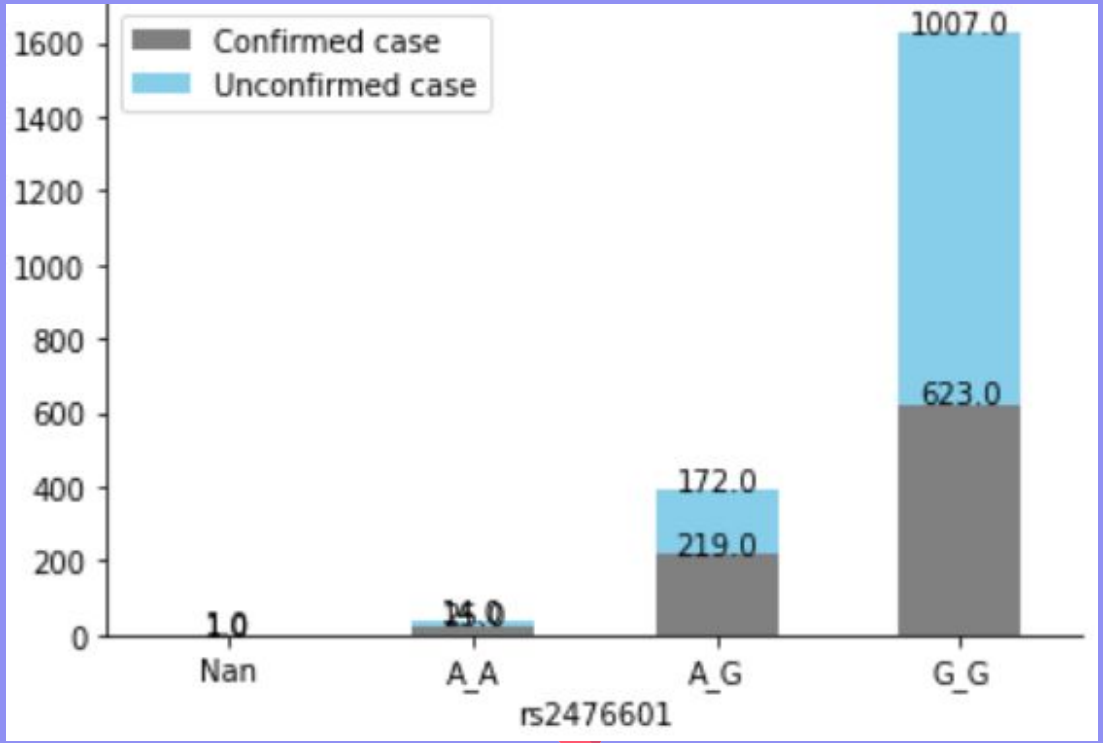



Prepare data

# 資料視覺化

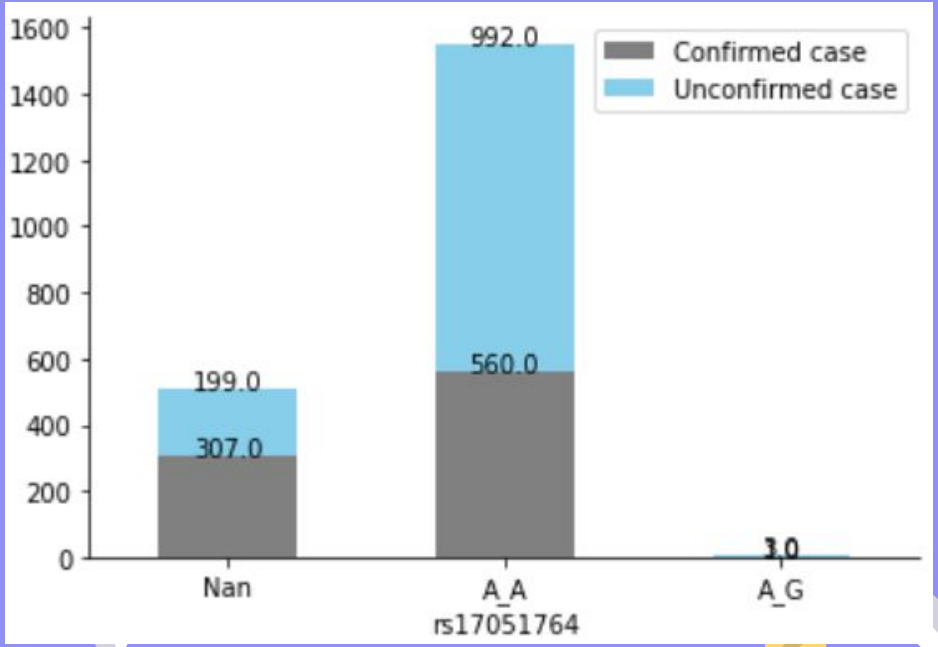
**由論文得知顯著SNP rs2476601其鹼基對分布情況**

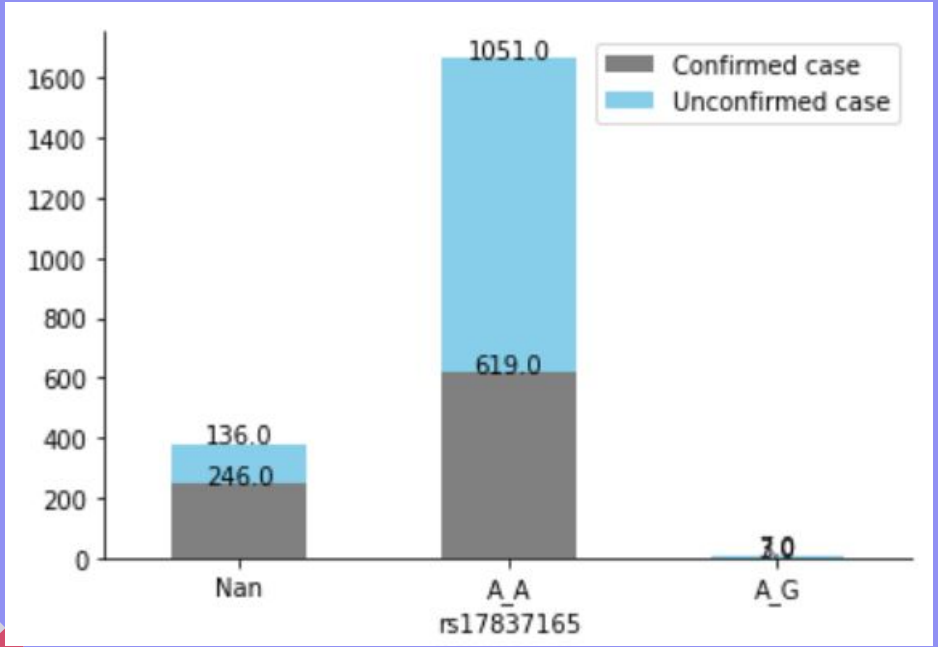


Prepare data

# 資料視覺化

**從顯著SNP中隨機抽取10個，觀察其鹼基對分布情況**

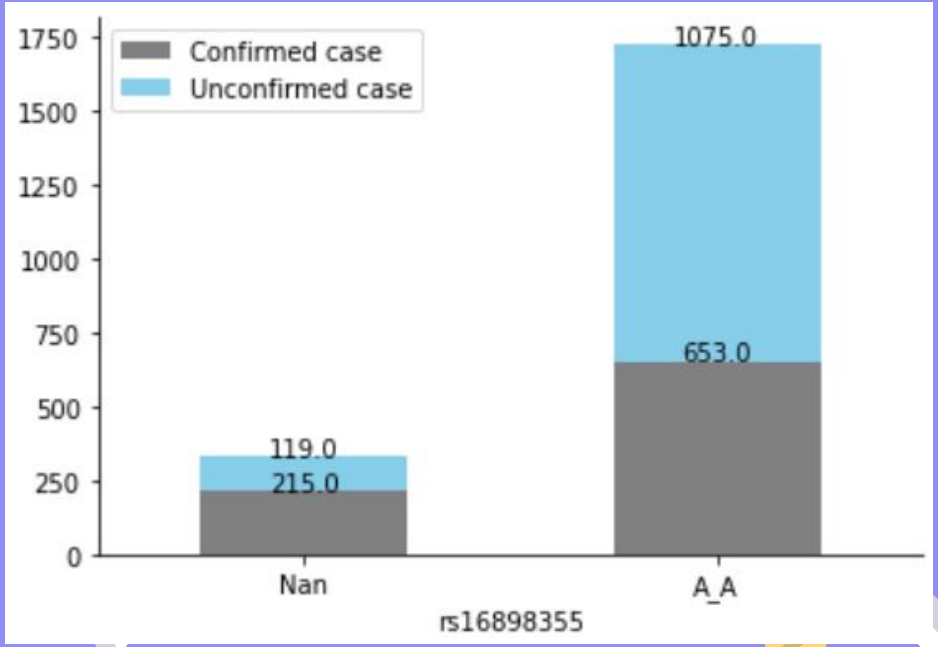


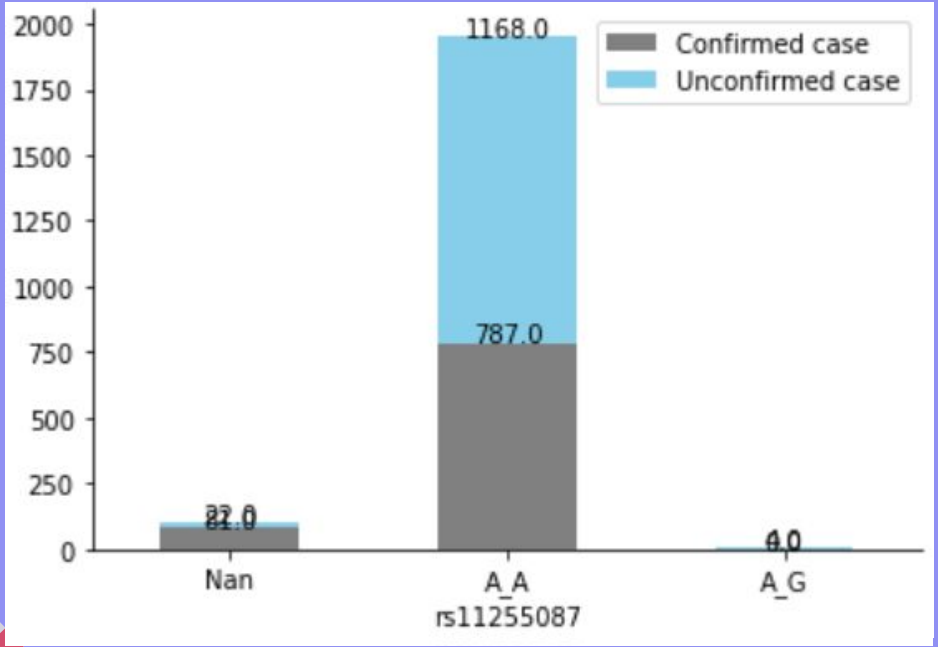



Prepare data

# 資料視覺化

**從顯著SNP中隨機抽取10個，觀察其鹼基對分布情況**

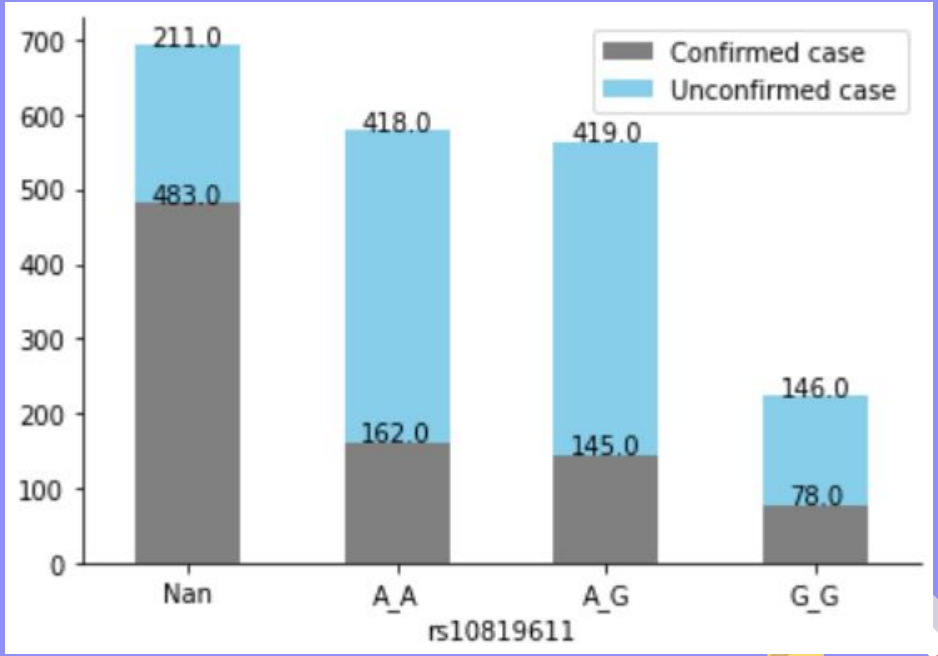


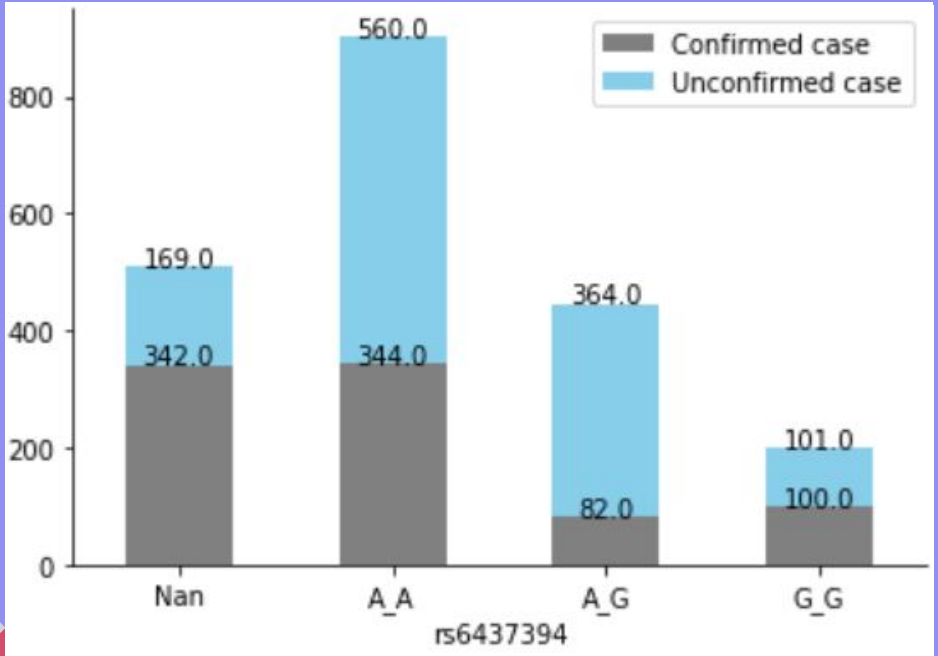



Prepare data

# 資料視覺化

**從顯著SNP中隨機抽取10個，觀察其鹼基對分布情況**

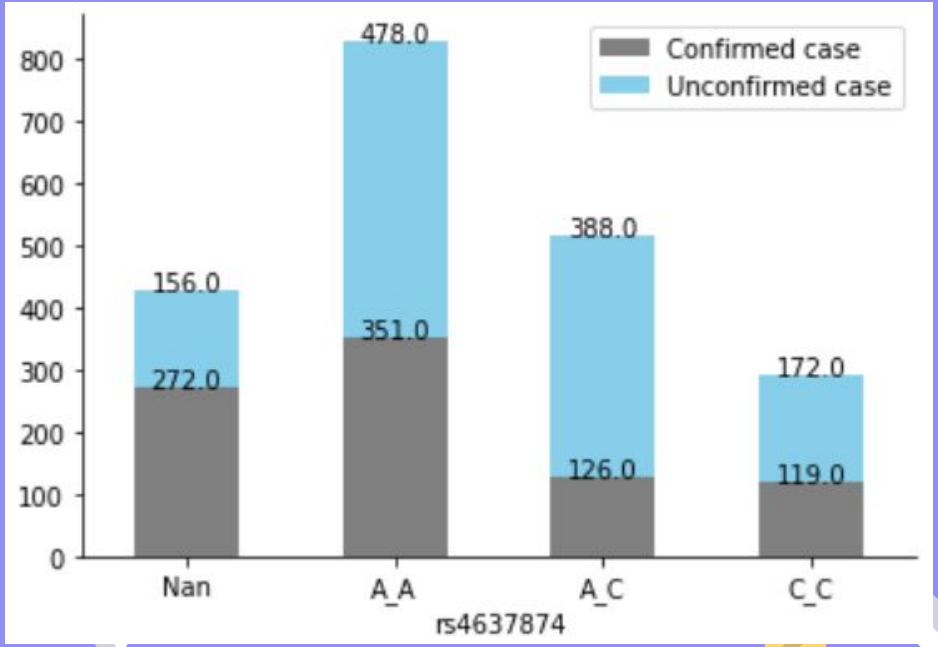


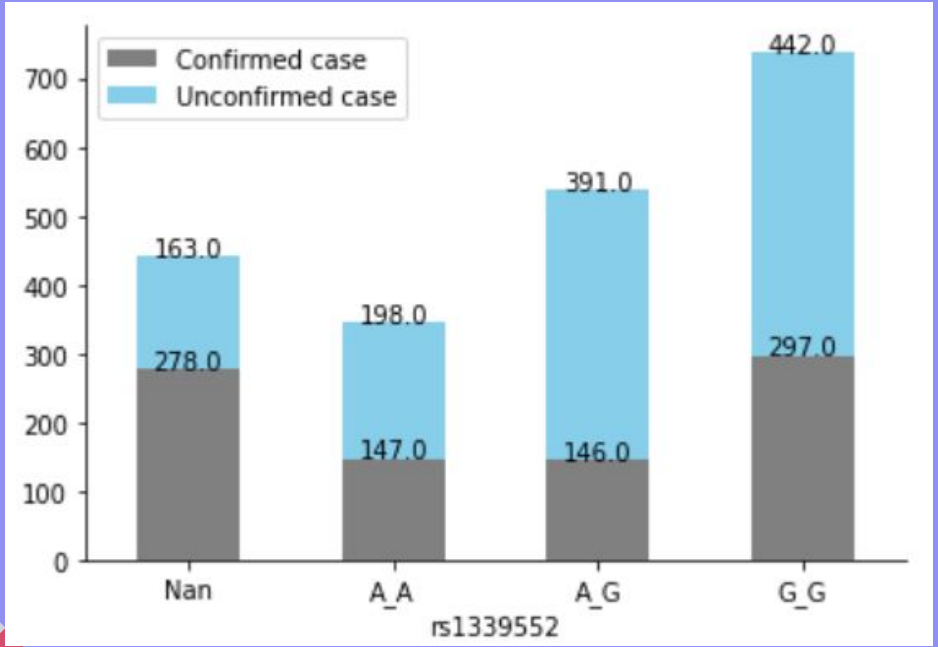



Prepare data

# 資料視覺化

**從顯著SNP中隨機抽取10個，觀察其鹼基對分布情況**

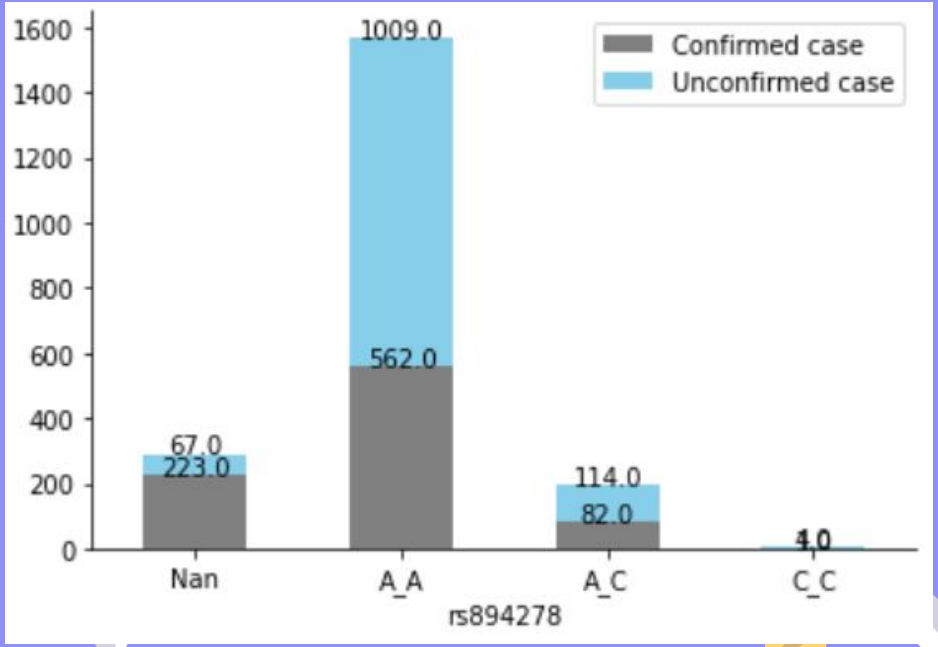


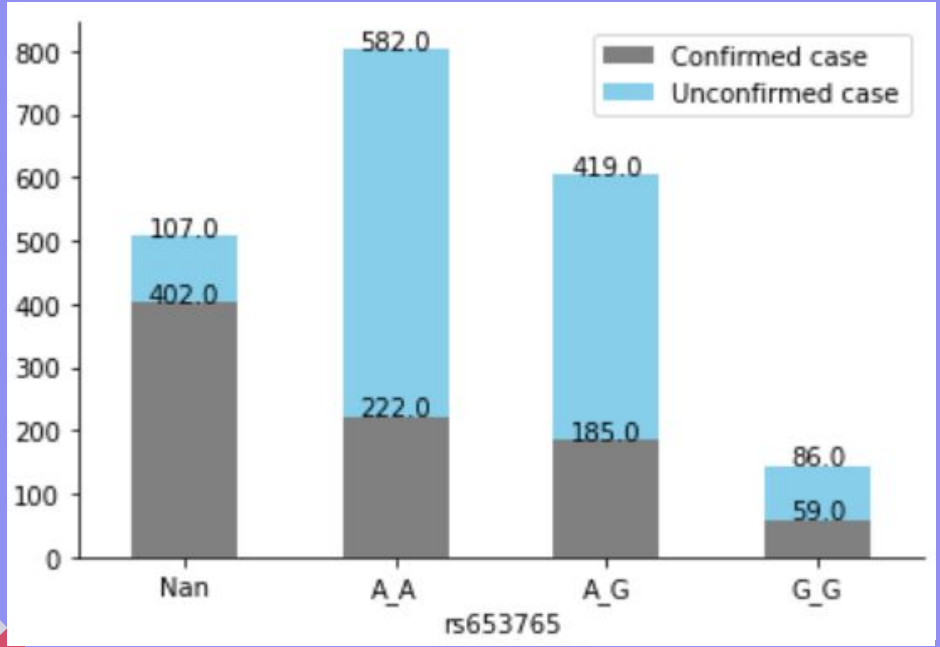



Prepare data

# 資料視覺化

**從顯著SNP中隨機抽取10個，觀察其鹼基對分布情況**



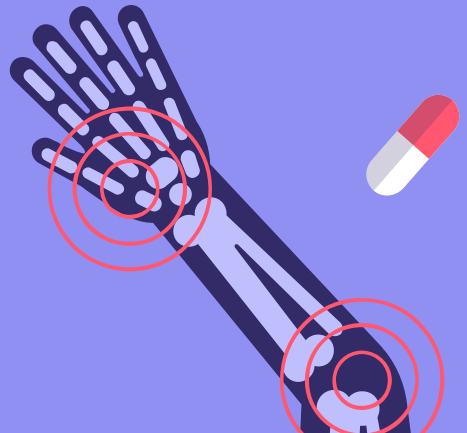
# 04.

# 模型建置與分析

# 模型探討

**01.**
挑選特徵多寡討論

**02.**
線性模型與非線性模型討論



# Build model

# 模型建置

1. 拆分訓練集與測試集
   a. 以8:2分配
   b. 亂數種子為666

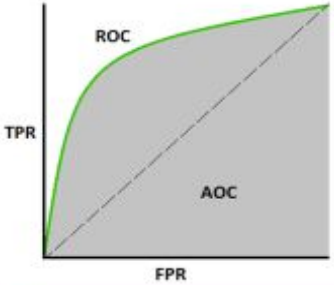
2. 特徵篩選(訓練集資料)
   a. 模型評估與基準
   b. 全部挑20個特徵
      i. DecisionTree
      ii. SelectFromModel_RandomForest
      iii. SelectFromModel_LGBM
      iv. SelectFromModel_XGB
   c. 特徵測試
      i. LogisticRegression
      ii. Random Forest
   d. 全部挑400個特徵
      i. SelectFromModel_LGBM
      ii. SelectFromModel_XGB
   e. 全部挑10個特徵
      i. SelectFromModel_LGBM
      ii. SelectFromModel_XGB

# Evaluate model

# 模型評估指標

**混淆矩陣(confusion matrix)**

| | 「模型預測」為真 (positive) | 「模型預測」為非 (negative) |
|---|---|---|
| 「真實情況」為真 | true positive (TP) | false negative (FN) |
| 「真實情況」為非 | false positive (FP) | true negative (TN) |




ROC曲線下方的面積(AUC ROC))
•因為是在1x1的方格裡求面積，AUC必在0~1之間。
•假設閾值以上是陽性，以下是陰性
•若隨機抽取一個陽性樣本和一個陰性樣本，分類器**正確判斷**陽性樣本的值高於陰性樣本之**機率**=AUC。

$$\text{Accuracy(準確率)} = \frac{TP+TN}{\text{全部資料總數}}$$

Model正確預測的機率值

在正向例子很少的狀況下，此種指標不適用

$$\text{Precision(精確率)} = \frac{TP}{TP+FP}$$

預測正向的狀況中，正確預測的機率值

較在意「預測正向」的答對數量

$$\text{Recall(召回率)} = \frac{TP}{TP+FN}$$

實際正向的狀況中，實際為真的機率值

較在意「實際正向」的答對數量

$$F_\beta = (1+\beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

① F1-source（$\beta = 1$）：precision和recall同等重要

② F2-source（$\beta = 2$）：recall比precision重要

③ F0.5-source（$\beta = 0.5$）：precision比recall重要

https://medium.com/@s716419/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E6%A8%A1%E5%9E%8B%E8%A9%95%E4%BC%B0%E6%8C%87%E6%A8%99-confusion-matrix-precision-and-recall-e9d64ff14d81

# Evaluate model

利用**Sex&DRB1&DRB2**當成基本的參考
以**CatBoost**模型為最好
約**0.80**

| Baseline(SEX&DRB1&DRB2) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
| 混淆矩陣 | [224, 0] [189, 0] | [13, 211] [4, 185] | [224, 0] [189, 0] | [13, 211] [4, 185] | [177, 47] [41, 148] | [182, 42] [61, 128] | [183, 41] [102, 87] | [200, 24] [69, 120] | [190, 34] [68, 121] | [190, 34] [54, 135] |
| accuracy | 0.54237 | 0.47942 | 0.54237 | 0.47942 | 0.78693 | 0.75061 | 0.65375 | 0.77482 | 0.75303 | 0.78693 |
| precision | 0.00000 | 0.46717 | 0.00000 | 0.46717 | 0.75897 | 0.75294 | 0.67969 | 0.83333 | 0.78065 | 0.79882 |
| recall | 0.00000 | 0.97884 | 0.00000 | 0.97884 | 0.78307 | 0.67725 | 0.46032 | 0.63492 | 0.64021 | 0.71429 |
| F1 | 0.00000 | 0.63248 | 0.00000 | 0.63248 | 0.77083 | 0.71309 | 0.54890 | 0.72072 | 0.70349 | 0.75419 |
| rocauc | 0.50000 | 0.51844 | 0.50000 | 0.51844 | 0.78662 | 0.74487 | 0.63864 | 0.76389 | 0.74421 | 0.78125 |

Train ,Test model Colab

# 模型數據分析

| SelectFromModel=DecisionTreeClassifier() | 20_feature | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
| 混淆矩陣 | [190, 34]<br>[36, 153] | [99, 125]<br>[0, 189] | [190, 34]<br>[32, 157] | [99, 125]<br>[0, 189] | [195, 29]<br>[46, 143] | [196, 28]<br>[26, 163] | [191, 33]<br>[35, 154] | [200, 24]<br>[46, 143] | [202, 22]<br>[41, 148] | [202, 22]<br>[33, 156] |
| accuracy | 0.830508 | 0.697337 | 0.840194 | 0.697337 | 0.818402 | 0.869249 | 0.835351 | 0.830508 | 0.847458 | 0.866828 |
| precision | 0.818182 | 0.601911 | 0.82199 | 0.601911 | 0.831395 | 0.853403 | 0.823529 | 0.856287 | 0.870588 | 0.876404 |
| recall | 0.809524 | 1.0 | 0.830688 | 1.0 | 0.756614 | 0.862434 | 0.814815 | 0.756614 | 0.783069 | 0.825397 |
| F1 | 0.81383 | 0.751491 | 0.826316 | 0.751491 | 0.792244 | 0.857895 | 0.819149 | 0.803371 | 0.824513 | 0.850136 |
| rocauc | 0.828869 | 0.720982 | 0.839451 | 0.720982 | 0.813575 | 0.868717 | 0.833747 | 0.824735 | 0.842427 | 0.863591 |
| 交叉驗證 (recall) | 0.809531 | 0.994737 | 0.835704 | 0.783073 | 0.756615 | 0.835989 | 0.814651 | 0.772262 | 0.841679 | 0.799004 |

**Features_select (CPU:56Cores RAM_MAX:192G Time:2m55.408s)**
**Train ,Test model Colab**

# 模型數據分析

| SelectFromModel=RandomForestClassifie() 20_feature | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
| 混淆矩陣 | [191 33]<br>[ 48 141] | [195 29]<br>[ 47 142] | [193 31]<br>[ 50 139] | [197 27]<br>[42 147] | [198 26]<br>[45 144] | [201 23]<br>[37 152] | [192 32]<br>[49 140] | [202 22]<br>[44 145] | [203 21]<br>[37 152] | [201 23]<br>[33 156] |
| accuracy | 0.803874 | 0.815981 | 0.803874 | 0.83293 | 0.828087 | 0.854722 | 0.803874 | 0.840194 | 0.859564 | 0.864407 |
| precision | 0.810345 | 0.830409 | 0.817647 | 0.844828 | 0.847059 | 0.868571 | 0.813953 | 0.868263 | 0.878613 | 0.871508 |
| recall | 0.746032 | 0.751323 | 0.73545 | 0.777778 | 0.761905 | 0.804233 | 0.740741 | 0.767196 | 0.804233 | 0.825397 |
| F1 | 0.77686 | 0.788889 | 0.774373 | 0.809917 | 0.802228 | 0.835165 | 0.775623 | 0.814607 | 0.839779 | 0.847826 |
| rocauc | 0.799355 | 0.810929 | 0.798528 | 0.828621 | 0.822917 | 0.850777 | 0.798942 | 0.834491 | 0.855241 | 0.861359 |
| 交叉驗證 (recall) | 0.772831 | 0.772546 | 0.762447 | 0.719915 | 0.788905 | 0.809531 | 0.783499 | 0.740967 | 0.804979 | 0.783215 |

Features_select (CPU:56Cores RAM_MAX:192G Time: 2m11.197s)
Train ,Test model Colab

# Evaluate model

# 模型數據分析

**Feature_selection_LGBM_20**
**lgb.LGBMClassifier(application='binary', boosting='gbdt', learning_rate=0.001, n_jobs=-1 ,max_depth=8)**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[191, 33], [41, 148]] | [[99, 125], [0, 189]] | [[190, 34], [38, 151]] | [[155, 69], [76, 113]] | [[185, 39], [45, 144]] | [[199, 25], [30, 159]] | [[194, 30], [47, 142]] | [[199, 25], [45, 144]] | [[202, 22], [42, 147]] | [[202, 22], [31, 158]] |
| accuracy | 0.82082 | 0.69734 | 0.82567 | 0.64891 | 0.79661 | 0.86683 | 0.81356 | 0.83051 | 0.84504 | 0.87167 |
| precision | 0.81768 | 0.60191 | 0.81622 | 0.62088 | 0.78689 | 0.86413 | 0.82558 | 0.85207 | 0.86982 | 0.87778 |
| recall | 0.78307 | 1.00000 | 0.79894 | 0.59788 | 0.76191 | 0.84127 | 0.75132 | 0.76191 | 0.77778 | 0.83598 |
| F1 | 0.80000 | 0.75149 | 0.80749 | 0.60916 | 0.77419 | 0.85255 | 0.78670 | 0.80447 | 0.82123 | 0.85637 |
| rocauc | 0.81787 | 0.72098 | 0.82358 | 0.64492 | 0.79390 | 0.86483 | 0.80870 | 0.82515 | 0.83978 | 0.86888 |

Features_select (CPU:56Cores RAM_MAX:192G Time:21m3.910s)

Train ,Test model Colab

# Evaluate model

# 模型數據分析

**Feature_selection_XGB_20**
**xgb.XGBClassifier(n_estimators=1000, learning_rate= 0.001,max_depth=8, n_jobs=-1, objective = 'binary:logitraw')**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[193, 31], [35, 154]] | [[99, 125], [0, 189]] | [[192, 32], [33, 156]] | [[99, 125], [0, 189]] | [[203, 21], [43, 146]] | [[199, 25], [24, 165]] | [[195, 29], [38, 151]] | [[200, 24], [29, 160]] | [[200, 24], [31, 158]] | [[201, 23], [22, 167]] |
| accuracy | 0.84019 | 0.69734 | 0.84262 | 0.69734 | 0.84504 | 0.88136 | 0.83777 | 0.87167 | 0.86683 | 0.89104 |
| precision | 0.83243 | 0.60191 | 0.82979 | 0.60191 | 0.87425 | 0.86842 | 0.83889 | 0.86957 | 0.86813 | 0.87895 |
| recall | 0.81482 | 1.00000 | 0.82540 | 1.00000 | 0.77249 | 0.87302 | 0.79894 | 0.84656 | 0.83598 | 0.88360 |
| F1 | 0.82353 | 0.75149 | 0.82759 | 0.75149 | 0.82023 | 0.87071 | 0.81843 | 0.85791 | 0.85175 | 0.88127 |
| rocauc | 0.83821 | 0.72098 | 0.84127 | 0.72098 | 0.83937 | 0.88070 | 0.83474 | 0.86971 | 0.86442 | 0.89046 |

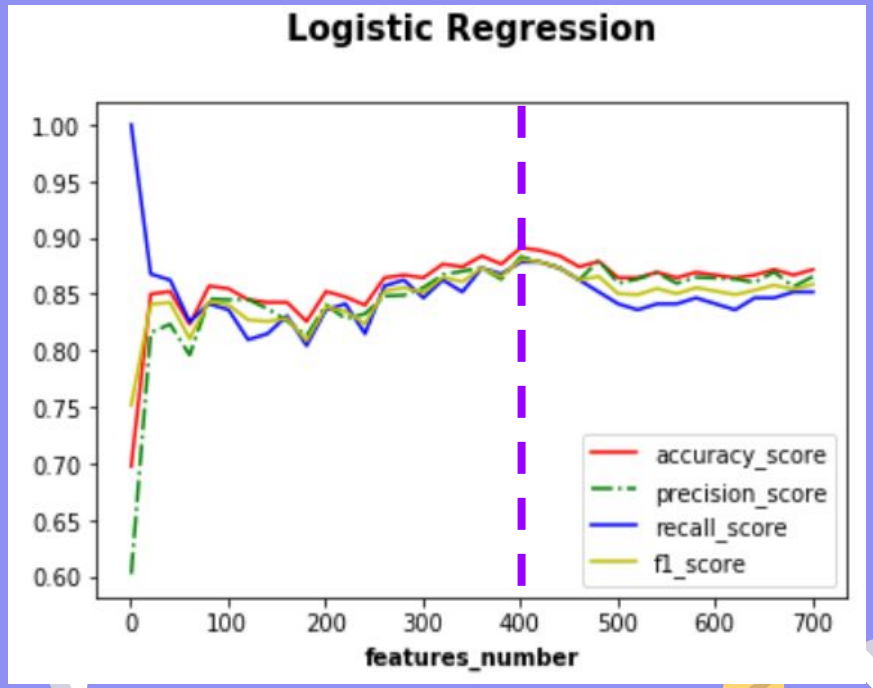
**Features_select (CPU:56Cores RAM_MAX:192G Time:52m44.273s)**

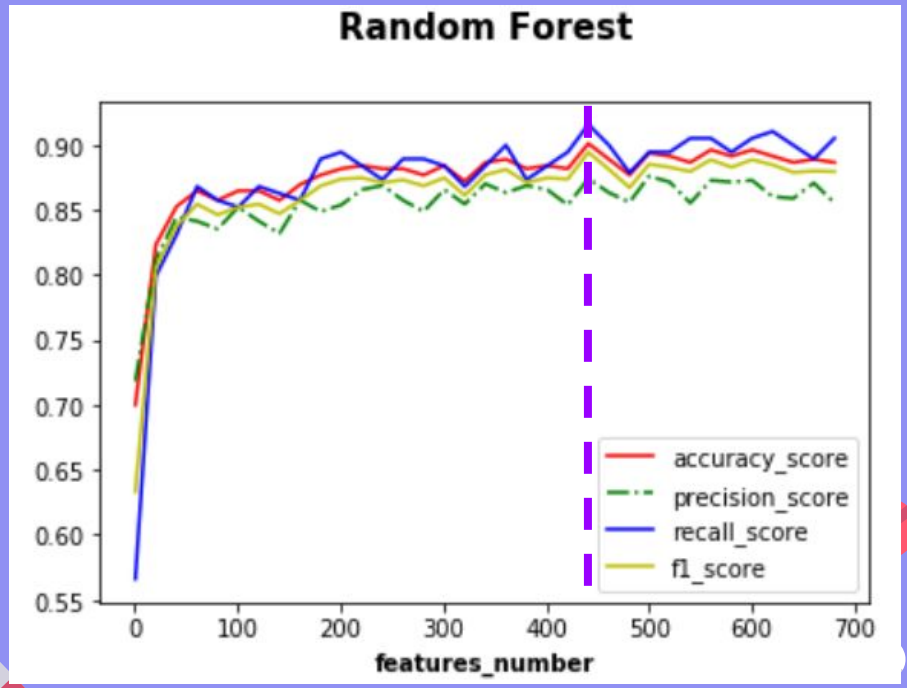
**Train ,Test model Colab**

# Evaluate model

## 模型數據分析

**Logistic Regression模型在特徵數量為400時，會有最佳的表現**
**Random Forest模型在特徵數量為435時，會有最佳的表現**



# Evaluate model

# 模型數據分析

**Feature_selection_LGBM_400**
**lgb.LGBMClassifier(application='binary', boosting='gbdt', learning_rate=0.001, n_jobs=-1 ,max_depth=8)**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[180, 44], [61, 128]] | [[109, 115], [1, 188]] | [[183, 41], [52, 137]] | [[220, 4], [178, 11]] | [[188, 36], [52, 137]] | [[195, 29], [28, 161]] | [[183, 41], [64, 125]] | [[203, 21], [31, 158]] | [[203, 21], [35, 154]] | [[202, 22], [25, 164]] |
| accuracy | 0.74576 | 0.71913 | 0.77482 | 0.55932 | 0.78693 | 0.86199 | 0.74576 | 0.87409 | 0.86441 | 0.88620 |
| precision | 0.74419 | 0.62046 | 0.76966 | 0.73333 | 0.79191 | 0.84737 | 0.75301 | 0.88268 | 0.88000 | 0.88172 |
| recall | 0.67725 | 0.99471 | 0.72487 | 0.05820 | 0.72487 | 0.85185 | 0.66138 | 0.83598 | 0.81482 | 0.86773 |
| F1 | 0.70914 | 0.76423 | 0.74659 | 0.10784 | 0.75691 | 0.84960 | 0.70423 | 0.85870 | 0.84615 | 0.87467 |
| rocauc | 0.74041 | 0.74066 | 0.77092 | 0.52017 | 0.78208 | 0.86119 | 0.73917 | 0.87111 | 0.86053 | 0.88476 |

Features_select (CPU:56Cores RAM_MAX:192G Time:21m9.439s)

Train ,Test model Colab

# Evaluate model

# 模型數據分析

**Feature_selection_XGB_400**
**xgb.XGBClassifier(n_estimators=1000, learning_rate= 0.001,max_depth=8, n_jobs=-1, objective = 'binary:logitraw')**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[183, 41], [37, 152]] | [[133, 91], [10, 179]] | [[189, 35], [35, 154]] | [[219, 5], [154, 35]] | [[199, 25], [45, 144]] | [[198, 26], [20, 169]] | [[181, 43], [37, 152]] | [[209, 15], [22, 167]] | [[202, 22], [24, 165]] | [[204, 20], [17, 172]] |
| accuracy | 0.81114 | 0.75545 | 0.83051 | 0.61501 | 0.83051 | 0.88862 | 0.80630 | 0.91041 | 0.88862 | 0.91041 |
| precision | 0.78757 | 0.66296 | 0.81482 | 0.87500 | 0.85207 | 0.86667 | 0.77949 | 0.91758 | 0.88235 | 0.89583 |
| recall | 0.80423 | 0.94709 | 0.81482 | 0.18519 | 0.76191 | 0.89418 | 0.80423 | 0.88360 | 0.87302 | 0.91005 |
| F1 | 0.79581 | 0.77996 | 0.81482 | 0.30568 | 0.80447 | 0.88021 | 0.79167 | 0.90027 | 0.87766 | 0.90289 |
| rocauc | 0.81060 | 0.77042 | 0.82928 | 0.58143 | 0.82515 | 0.88905 | 0.80613 | 0.90832 | 0.88740 | 0.91038 |

Features_select (CPU:56Cores RAM_MAX:192G Time:52m11.994s)

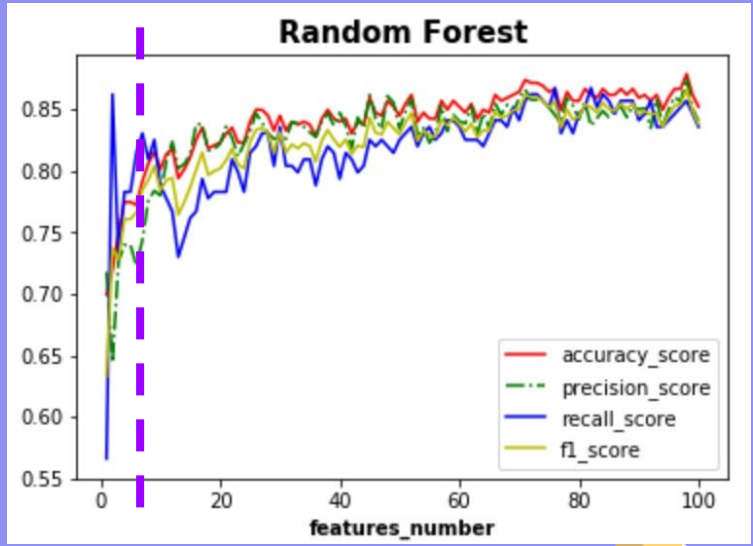
Train ,Test model Colab

# Evaluate model

# 模型數據分析

**Random Forest模型在特徵數量為10左右，就會到一定基準**

| | accuracy_score | precision_score | recall_score | f1_score | features_number |
|---|---|---|---|---|---|
| 0 | 0.699758 | 0.718121 | 0.566138 | 0.633136 | 1 |
| 1 | 0.719128 | 0.644269 | 0.862434 | 0.737557 | 2 |
| 2 | 0.750605 | 0.726316 | 0.730159 | 0.728232 | 3 |
| 3 | 0.774818 | 0.740000 | 0.783069 | 0.760925 | 4 |
| 4 | 0.774818 | 0.740000 | 0.783069 | 0.760925 | 5 |
| 5 | 0.772397 | 0.723005 | 0.814815 | 0.766169 | 6 |
| 6 | 0.791768 | 0.744076 | 0.830688 | 0.785000 | 7 |
| 7 | 0.806295 | 0.776650 | 0.809524 | 0.792746 | 8 |
| 8 | 0.815981 | 0.783920 | 0.825397 | 0.804124 | 9 |
| 9 | 0.801453 | 0.780105 | 0.788360 | 0.784211 | 10 |
| 10 | 0.813559 | 0.807692 | 0.777778 | 0.792453 | 11 |
| 11 | 0.818402 | 0.823864 | 0.767196 | 0.794521 | 12 |
| 12 | 0.794189 | 0.802326 | 0.730159 | 0.764543 | 13 |
| 13 | 0.801453 | 0.805714 | 0.746032 | 0.774725 | 14 |
| 14 | 0.811138 | 0.813559 | 0.761905 | 0.786885 | 15 |



# Evaluate model

# 模型數據分析

**Feature_selection_LGBM_10**
**lgb.LGBMClassifier(application='binary', boosting='gbdt', learning_rate=0.001, n_jobs=-1 ,max_depth=8)**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[189, 35], [42, 147]] | [[99, 125], [0, 189]] | [[186, 38], [39, 150]] | [[99, 125], [0, 189]] | [[193, 31], [39, 150]] | [[197, 27], [35, 154]] | [[189, 35], [43, 146]] | [[197, 27], [44, 145]] | [[202, 22], [47, 142]] | [[199, 25], [29, 160]] |
| accuracy | 0.81356 | 0.69734 | 0.81356 | 0.69734 | 0.83051 | 0.84988 | 0.81114 | 0.82809 | 0.83293 | 0.86925 |
| precision | 0.80769 | 0.60191 | 0.79787 | 0.60191 | 0.82873 | 0.85083 | 0.80663 | 0.84302 | 0.86585 | 0.86487 |
| recall | 0.77778 | 1.00000 | 0.79365 | 1.00000 | 0.79365 | 0.81482 | 0.77249 | 0.76720 | 0.75132 | 0.84656 |
| F1 | 0.79245 | 0.75149 | 0.79576 | 0.75149 | 0.81081 | 0.83243 | 0.78919 | 0.80332 | 0.80453 | 0.85562 |
| rocauc | 0.81076 | 0.72098 | 0.81200 | 0.72098 | 0.82763 | 0.84714 | 0.80812 | 0.82333 | 0.82655 | 0.86748 |

Features_select (CPU:56Cores RAM_MAX:192G Time:21m17.850s)

Train ,Test model Colab

# Evaluate model

# 模型數據分析

**Feature_selection_XGB_10**
**xgb.XGBClassifier(n_estimators=1000, learning_rate= 0.001,max_depth=8, n_jobs=-1, objective = 'binary:logitraw')**

| estimator | LR | NB | LDA | QDA | DT | RF | SVM | XGB | LGBM | CAT |
|---|---|---|---|---|---|---|---|---|---|---|
| 混淆矩陣 | [[189, 35], [42, 147]] | [[99, 125], [0, 189]] | [[186, 38], [39, 150]] | [[99, 125], [0, 189]] | [[193, 31], [39, 150]] | [[197, 27], [35, 154]] | [[189, 35], [43, 146]] | [[197, 27], [44, 145]] | [[202, 22], [47, 142]] | [[199, 25], [29, 160]] |
| accuracy | 0.81356 | 0.69734 | 0.81356 | 0.69734 | 0.83051 | 0.84988 | 0.81114 | 0.82809 | 0.83293 | 0.86925 |
| precision | 0.80769 | 0.60191 | 0.79787 | 0.60191 | 0.82873 | 0.85083 | 0.80663 | 0.84302 | 0.86585 | 0.86487 |
| recall | 0.77778 | 1.00000 | 0.79365 | 1.00000 | 0.79365 | 0.81482 | 0.77249 | 0.76720 | 0.75132 | 0.84656 |
| F1 | 0.79245 | 0.75149 | 0.79576 | 0.75149 | 0.81081 | 0.83243 | 0.78919 | 0.80332 | 0.80453 | 0.85562 |
| rocauc | 0.81076 | 0.72098 | 0.81200 | 0.72098 | 0.82763 | 0.84714 | 0.80812 | 0.82333 | 0.82655 | 0.86748 |

Features_select (CPU:56Cores RAM_MAX:192G Time:52m1.194s)

Train ,Test model Colab

Evaluate model

# 模型數據分析

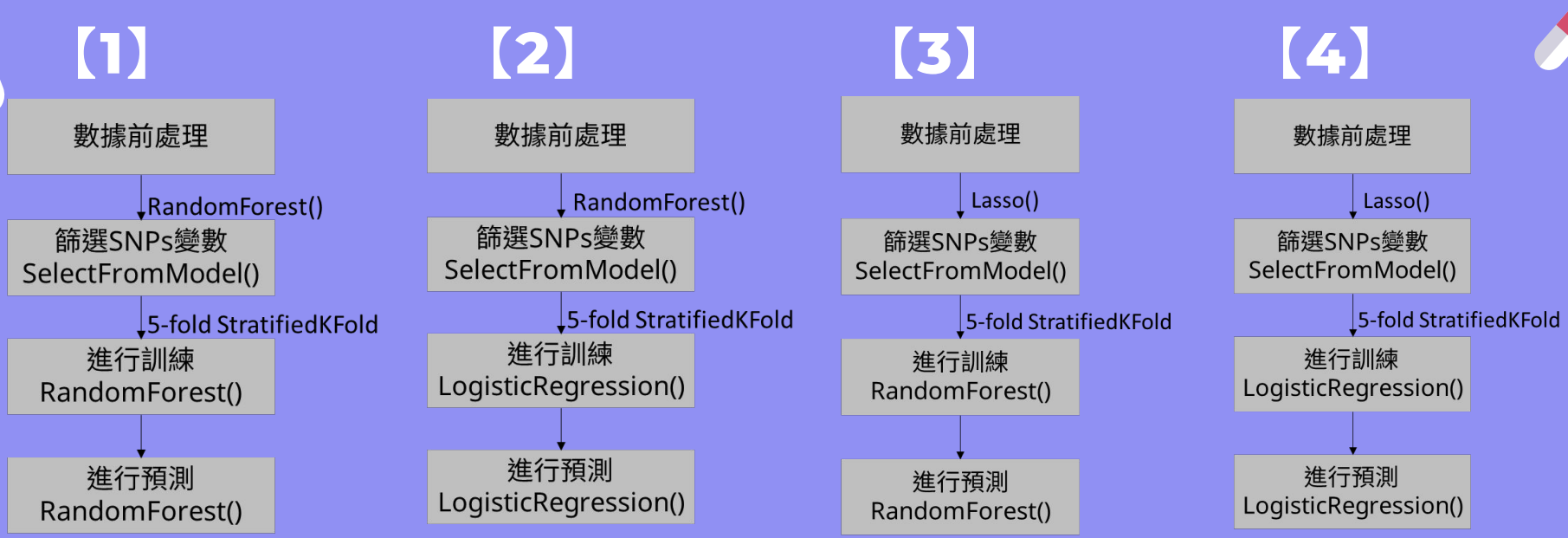
**就算挑出來的feature不一樣，模型訓練出來成果也很好**

LGBM
feature = [,'rs660895', 'rs9275388', 'rs12525276', 'rs12530984', 'rs10508741',
'rs12249377', 'rs8181424', 'rs1704094', 'rs281440' ,'MitoG1440A']

XGB
feature = [,'rs187454', 'rs11746584', 'rs2395175', 'rs660895' ,'rs9275388',
'rs910603','rs12530984', 'rs12249377' ,'rs1704094', 'rs4919838']

# Evaluate model

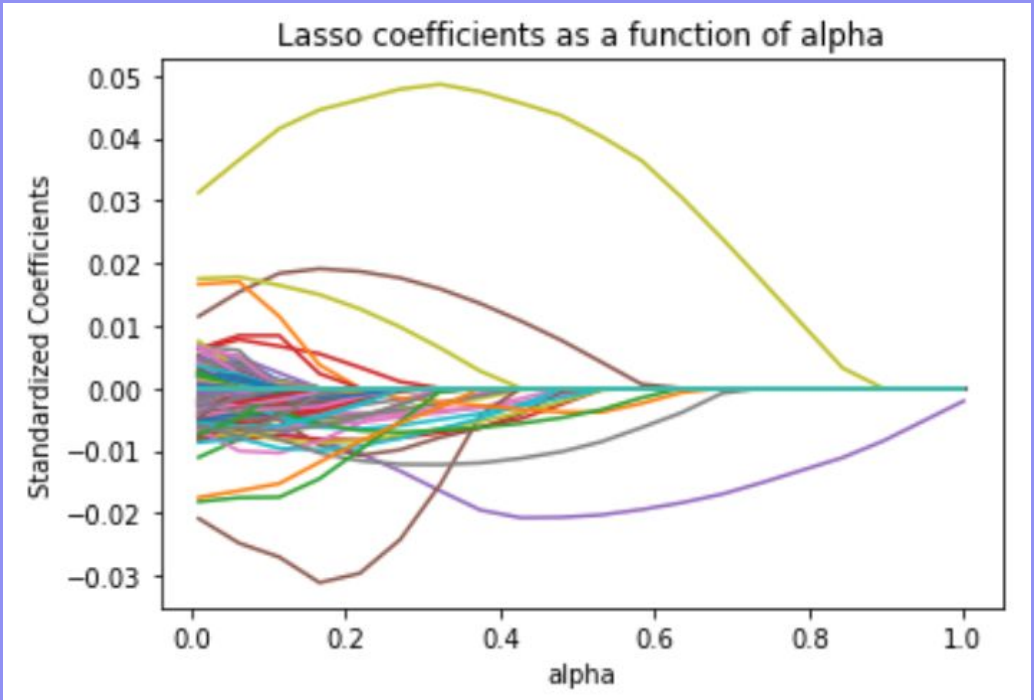
# 模型數據分析



Evaluate model

# 模型數據分析

**Lasso模型在α為0.1時，會有最佳的表現**



# Evaluate model

# 模型數據分析

| SelectFromModel(estimator=RandomForest, max_features=435) | | |
|---|---|---|
| | RF Time:20.9s | LR Time:20.3s |
| 混淆矩陣 | [[202 22]<br>[ 27 162]] | [[202 22]<br>[ 27 162]] |
| accuracy | 0.8814 | 0.8814 |
| precision | 0.8804 | 0.8804 |
| recall | 0.8571 | 0.8571 |
| f1 | 0.8686 | 0.8686 |

Features_select (CPU:Intel(R) Xeon(R) Gold 6138 CPU(64 cores) RAM_MAX:37.6G/1.5T)

# Evaluate model

# 模型數據分析

| SelectFromModel(estimator=Lasso(alpha=0.1)) features=51 | | |
|---|---|---|
| | RF Time:1m41.5s | LR Time:1m47.8s |
| 混淆矩陣 | [[201 23]<br>[ 27 162]] | [[191 33]<br>[ 28 161]] |
| accuracy | 0.8789 | 0.8523 |
| precision | 0.8757 | 0.8299 |
| recall | 0.8571 | 0.8519 |
| f1 | 0.8663 | 0.8407 |

Features_select (CPU:Intel(R) Xeon(R) Gold 6138 CPU(64 cores) RAM_MAX:37.6G/1.5T)

# Evaluate model

# 模型數據分析

| SelectFromModel(estimator=LassoCV(n_alpha=1, alphas=[0.1], cv=StratifiedKFold(n_splits=5))) | | features=51 |
|---|---|---|
| | RF Time:2m0.2s | LR Time:2m10.5s |
| 混淆矩陣 | [[201 23]<br>[ 27 162]] | [[191 33]<br>[ 28 161]] |
| accuracy | 0.8789 | 0.8523 |
| precision | 0.8757 | 0.8299 |
| recall | 0.8571 | 0.8519 |
| f1 | 0.8663 | 0.8407 |

Features_select (CPU:Intel(R) Xeon(R) Gold 6138 CPU(64 cores) RAM_MAX:37.6G/1.5T)

Communicate results

# 結果與討論

- 透過上面的模型結果，我們總結了以下幾點
  - 特徵值的多寡(10個與400個)與模型表現影響不大(5%以內)
  - 非線性模型挑選出來的變數(SNP)送進非線性模型中，表現比線性模型的結果要來得好

# THANKS!

# 討論紀錄 20221104

時間:2022/11/04　18:00-20:00
人員:楊育維、郭人豪、曾以諾
會議記錄:曾以諾
討論內容:

1. 挑選SNP的方法
    a. 隨機挑選、挑選論文提出的高風險 SNP
2. 資料預處理、視覺化細節(郭人豪)
3. 決定用logistic regression和naive Bayes為初步模型(楊育維)
4. 相關文獻回顧(曾以諾)
5. 決定下次討論時間

# 討論紀錄 20221107

時間:2022/11/07　18:00-20:00
人員:楊育維、郭人豪、曾以諾
會議記錄:楊育維
討論內容:

1. 簡報個人部分
2. 討論資料視覺化圖形
3. 模型評估結果討論

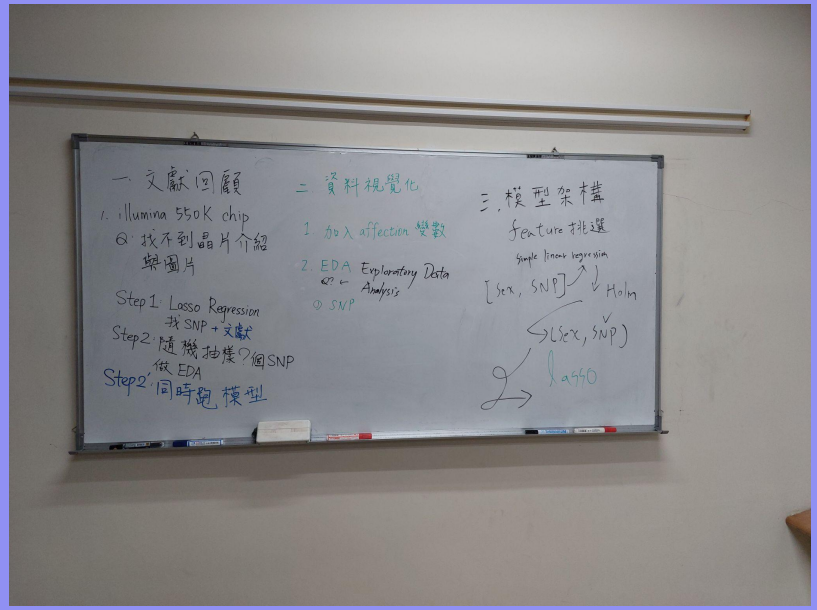
# 討論紀錄 20221202

時間:2022/12/02　16:00-18:00
人員:楊育維、郭人豪、曾以諾
會議記錄:曾以諾
討論內容:

1. 修改文獻回顧
2. 修改資料視覺化
3. 修改模型之輸入變數
4. 變數(SNP)選擇

分工內容:

1. 優化參考文獻回顧:曾以諾
2. EDA與資料視覺化:郭人豪
3. 建立模型與結果分析:楊育維



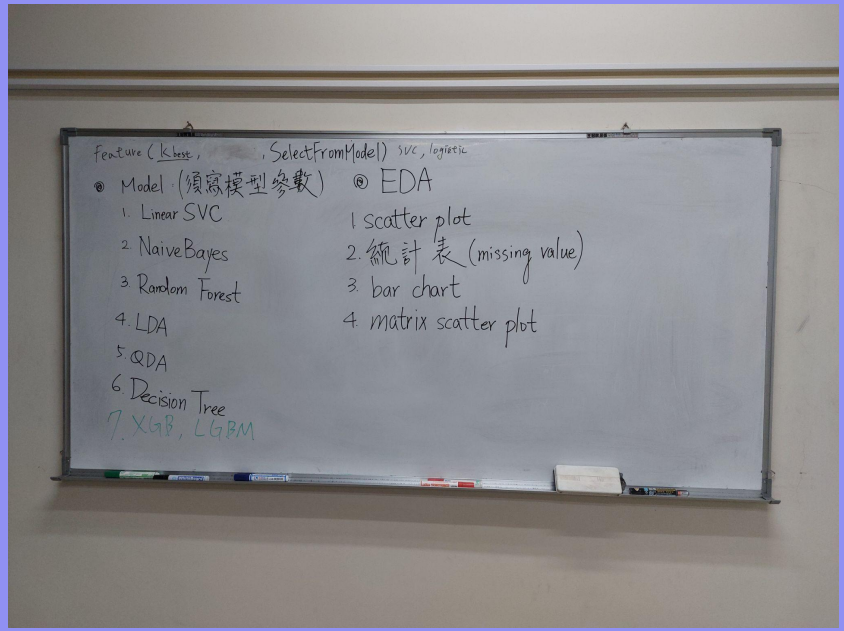
# 討論紀錄 20221205

時間:2022/12/05　17:00~19:00
人員:楊育維、郭人豪、曾以諾
會議記錄:曾以諾
討論內容:

1. 修改文獻回顧
2. 修改資料視覺化
3. 修改模型之輸入變數
4. 變數(SNP)選擇
5. 模型選擇

分工內容:

1. 優化參考文獻回顧、跑模型:曾以諾
2. EDA與資料視覺化:郭人豪
3. 跑模型與結果分析:楊育維



# 討論紀錄 20221206

時間:2022/12/05　17:30~18:30
人員:楊育維、郭人豪、曾以諾
會議記錄:曾以諾
討論內容:

1. 討論簡報、報告內容

分工內容:

1. 優化參考文獻回顧、跑模型:曾以諾
2. EDA與資料視覺化:郭人豪
3. 跑模型與結果分析:楊育維

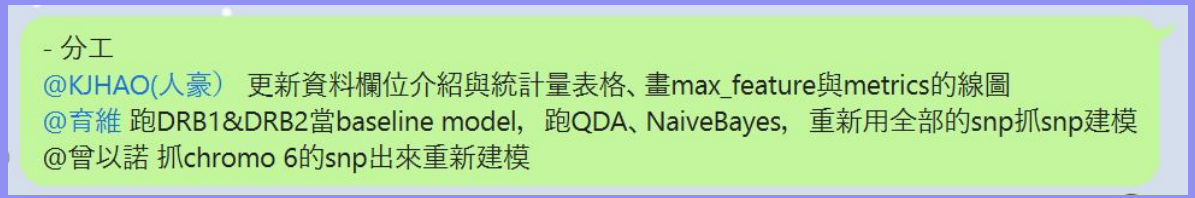
# 討論紀錄 20221224

**時間:2022/12/24　18:00~20:00**
**人員:楊育維、郭人豪、曾以諾**
**會議記錄:曾以諾**
**討論內容:**

1. **討論簡報、報告內容**

**分工內容:**

1. **跑Random Forest, Lasso模型:曾以諾**
2. **EDA與資料視覺化修改、跑模型:郭人豪**
3. **跑Random Forest, Decision Tree模型與結果分析:楊育維**



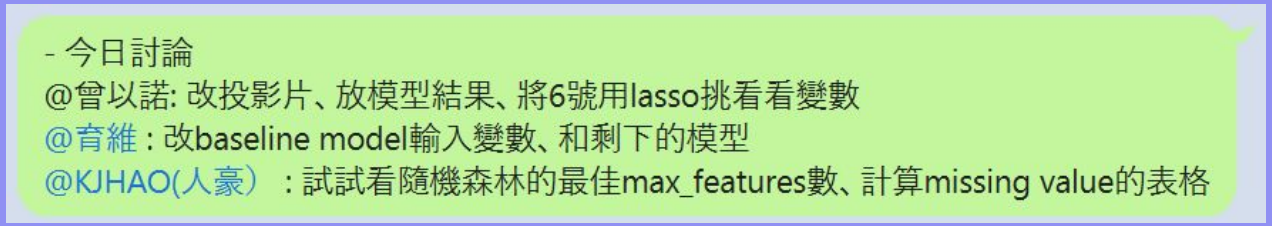
# 討論紀錄 20221226

**時間:2022/12/26　18:00~19:00**
**人員:楊育維、郭人豪、曾以諾**
**會議記錄:曾以諾**
**討論內容:**

1. **討論簡報、報告內容**

**分工內容:**

1. **跑Lasso模型、修改投影片(增加上次討論模型的結果):曾以諾**
2. **修改計算missing values的表格、尋找應該挑選多少參數的理想 值:郭人豪**
3. **跑XGBoost、LightGBM、CATboost等模型與結果分析:楊育維**



# 討論紀錄 20221226

時間:2022/12/26　18:00~19:00
人員:楊育維、郭人豪、曾以諾
會議記錄:曾以諾
討論內容:

1. 討論簡報、報告內容

分工內容:

1. 跑Lasso模型、修改投影片(增加上次討論模型的結果):曾以諾
2. 修改計算missing values的表格、尋找應該挑選多少參數的理想 值:郭人豪
3. 跑XGBoost、LightGBM、CATboost等模型不同特徵數量討論:楊育維