
美國國家糖尿病與消化和腎臟疾病研究所糖尿病數
據與身體各項指標關聯探討與患病預測

Author

109550074 吳秉澍

109550178 黃昱翰

109550121 溫柏萱

June 21, 2024

Contents

1 研究動機	5
1.1 糖尿病簡介	5
1.2 各國糖尿病統計	5
2 資料分析	7
2.1 資料來源	7
2.2 欄位解釋	7
2.3 資料分布	8
2.3.1 類別型資料分佈	8
2.3.2 數值型資料分佈	9
2.4 異常值檢測	10
2.4.1 缺失值	10
2.5 資料前處理	10
2.5.1 資料前處理對 AIC 的影響	10
3 變數分析	13
3.1 One-way ANOVA	13
3.1.1 8 個解釋變數	13
3.1.2 將有缺失值的欄位根據「是否為缺失值」作為 5 個解釋變數 . . .	14
3.1.3 將以上結果合併為 13 個解釋變數	15
3.2 Two-way ANOVA	16
3.2.1 Pregnancies vs. Glucose	16
3.2.2 Pregnancies vs. BloodPressure	17
3.2.3 Pregnancies vs. SkinThickness	18
3.2.4 Pregnancies vs. Insulin	19
3.2.5 Pregnancies vs. BMI	19
3.2.6 Pregnancies vs. DiabetesPedigreeFunction	20
3.2.7 Pregnancies vs. Age	21
3.2.8 Glucose vs. BloodPressure	23
3.2.9 Glucose vs. SkinThickness	24
3.2.10 Glucose vs. Insulin	24
3.2.11 Glucose vs. BMI	26
3.2.12 Glucose vs. DiabetesPedigreeFunction	27
3.2.13 Glucose vs. Age	28

3.2.14	BloodPressure vs. SkinThickness	30
3.2.15	BloodPressure vs. Insulin	30
3.2.16	BloodPressure vs. BMI	31
3.2.17	BloodPressure vs. DiabetesPedigreeFunction	32
3.2.18	BloodPressure vs. Age	33
3.2.19	SkinThickness vs. Insulin	34
3.2.20	SkinThickness vs. BMI	35
3.2.21	SkinThickness vs. DiabetesPedigreeFunction	35
3.2.22	SkinThickness vs. Age	36
3.2.23	Insulin vs. BMI	38
3.2.24	Insulin vs. DiabetesPedigreeFunction	39
3.2.25	Insulin vs. Age	40
3.2.26	BMI vs. DiabetesPedigreeFunction	41
3.2.27	BMI vs. Age	42
3.2.28	DiabetesPedigreeFunction vs. Age	43
3.3	AIC	43
3.3.1	不考慮交互作用	44
3.3.1.1	只用原始的 8 個解釋變數的模型	44
3.3.1.2	使用原始的 8 個解釋變數以及額外的 5 個解釋變數的結果	45
3.3.1.3	將 13 個解釋變數中不具統計顯著性的解釋變數刪除後 的結果	46
3.3.2	考慮交互作用	47
3.3.2.1	考慮全部交互作用項	47
3.3.2.2	從 8 個原始解釋變數中取出具統計顯著性的 4 個解釋變 數相乘	48
3.3.2.3	加入 two-way ANOVA 中， p-value 小於 0.1 的交互作用 項的結果	50
3.3.2.4	加入 two-way ANOVA 中， p-value 小於 0.05 的交互作用項	51
3.3.2.5	考慮 Age(X8) 與 4 個重要解釋變數的所有交互作用項 . .	53
3.3.2.6	缺失值處理-加入 5 個額外解釋變數	55
3.3.2.7	缺失值處理-中位數填充	56
3.3.2.8	最終模型	59
3.4	模型精確度	61

4 機器學習模型比較	65
4.1 方法介紹	65
4.2 模型訓練與評估	65
4.3 模型準確率和分類報告	66
4.3.1 使用 Raw Data 預測結果	66
4.3.2 使用 Filtered Data 預測結果	67
4.3.3 使用 Imputed Data 預測結果	68
4.4 總結	68
4.5 與自定義統計模型 3.3.2.8 比較	69
5 結論	69
6 心得	70

1 研究動機

根據衛福部統計，糖尿病位居十大死因之一，每年近萬人因糖尿病死亡，根據國民健康署統計，全國約有 200 多萬名糖尿病患者，且每年以 25,000 名的速度持續增加。截至 2019 年，全球有一億九千萬名糖尿病患，至 2025 年世界衛生組織預估有三億三千萬名病患，但截至 2021 年的糖尿病患數量為 5.37 億人 [1]，遠遠超過 2019 年的預估值。其死亡率增加速度為在過去二十年中最快的一種疾病，而糖尿病所帶來的醫療負擔沈重，美國的糖尿病患者佔總人口數 2.8%，當年花在糖尿病的總費用約為 920 億美元，其中直接的醫藥支出為 452 億美元，而間接的費用則為 466 億美元 [2]，若在輕度甚至未達糖尿病標準時及早預防及控制，不僅可避免諸多併發症與提高患者生活質量，更能減輕醫療的沈重負擔。

故，我們希望通過深入探討身體各項指標與糖尿病發病之間的關聯，建立一個準確的預測模型，從而為及早診斷和治療提供科學依據。

1.1 糖尿病簡介

糖尿病是一種慢性代謝性疾病，特徵是血糖水平持續升高，主要分為 1 型糖尿病、2 型糖尿病和妊娠糖尿病三種類型。1 型糖尿病是由於胰島素生成不足所致，常見於兒童和青少年；2 型糖尿病則與胰島素抵抗和相對胰島素不足有關，通常發生在成年人中；妊娠糖尿病則是孕期發生的高血糖現象，對母嬰健康有重大影響 [3]。

糖尿病的併發症可分為急性與慢性，常見急性併發症包括低血糖、高血糖和糖尿病酮症酸中毒，嚴重者可能導致昏迷或死亡。慢性併發症涉及多器官系統，包括心血管疾病、腎臟病、視網膜病變、足部潰傷難以癒合，嚴重者可能需要截肢 [4]。

1.2 各國糖尿病統計

根據國際糖尿病聯盟統計 [1]，截至 2021 年全球有 5.37 億糖尿病病患 [1]，預估於 2030 年將增至 6.43 億，2045 年將進一步增至 7.83 億。



Figure 1: 2021 世界各地糖尿病統計

糖尿病的高發病率和高死亡率促使科學界不斷探索其病因、預防和治療方法。現有研究表明，通過健康飲食、規律運動和體重管理等方式，可以有效預防 2 型糖尿病的發生。同時，早期診斷和及時治療對於控制病情和減少併發症具有關鍵作用。

2 資料分析

2.1 資料來源

資料源自美國國家糖尿病與消化和腎臟疾病研究所數據，共 768 筆資料，資料蒐集時限定範圍為 21 歲以上具有印度 Pima 血統的女性，故應變數分佈並非全人口患有糖尿病的真實分佈。

2.2 欄位解釋

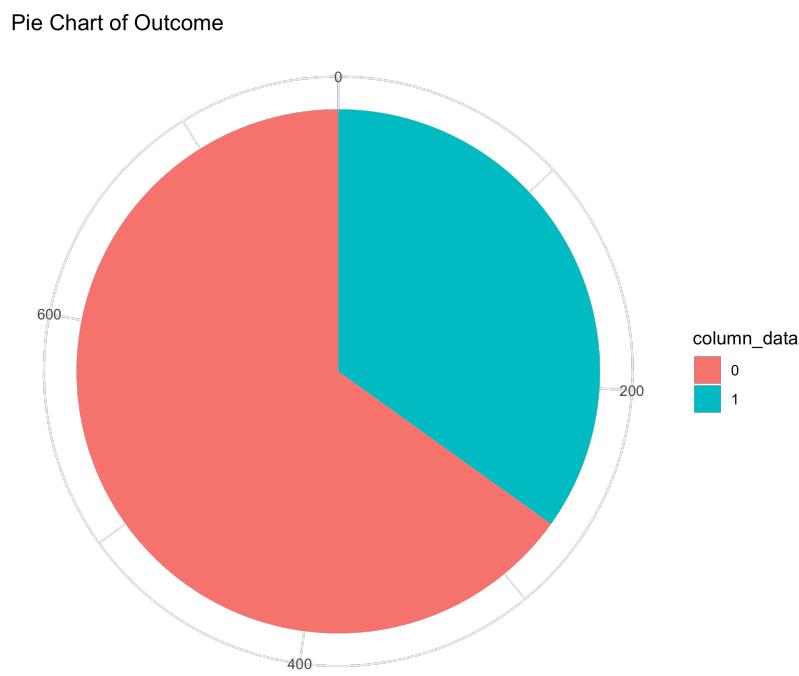
欄位名稱	解釋
Pregnancies	妊娠次數 (次)
Age	年齡 (歲)
BloodPressure	血壓 (mmHg)
BMI	身體質量指數 ($\frac{\text{體重(kg)}}{\text{身高(m)}^2}$)
SkinThickness	三頭肌皮褶厚度 (mm)，用於評估體脂肪
Glucose	口服葡萄糖耐量測試 (OGTT) 中的 2 小時血漿葡萄糖濃度 (mg/dL)，用於評估胰島素的分泌情況及其對葡萄糖負荷的反應
DiabetesPedigreeFunction	糖尿病家族病史指標，根據家族成員中的糖尿病情況以及不同成員之間的親緣關係，量化家族史對於個人糖尿病風險的影響
Insulin	2 小時血清胰島素濃度 ($\mu\text{U}/\text{mL}$)，在進行口服葡萄糖耐量測試 (Oral Glucose Tolerance Test, OGTT) 後，受試者在攝入葡萄糖溶液後 2 小時所測量的血清胰島素濃度。這個測量值以微國際單位每毫升 ($\mu\text{U}/\text{mL}$) 為單位，用來評估胰島素的分泌情況及其對葡萄糖負荷的反應。

Table 1: 資料集欄位解釋

2.3 資料分布

2.3.1 類別型資料分佈

Outcome 為 0 代表未罹患糖尿病，1 代表罹患糖尿病。資料集中，無糖尿病與有糖尿病的資料筆數分別為 500 筆與 268 筆。



(a) Response Variable 分佈

Figure 2: 類別型資料圓餅圖

2.3.2 數值型資料分佈



2.4 異常值檢測

2.4.1 缺失值

BMI(身體質量指數)、BloodPressure(血壓)、SkinThickness(皮膚厚度)、Glucose(血糖濃度)、Insulin(血液中胰島素)有0顯然不合理，判斷為缺失值。

2.5 資料前處理

去除資料集中顯不合理的資料，如身體質量指數、血壓、皮膚厚度、血糖濃度、血液中胰島素為零者。

2.5.1 資料前處理對AIC的影響

```
1 > filtered_data <- data %>%
2 +   filter(Glucose != 0, BloodPressure != 0, SkinThickness != 0, Insulin != 0, BMI != 0)
3 >
4 > model <- glm(Outcome ~ ., data = filtered_data, family = binomial)
5 > summary(model)
6
7 Call:
8 glm(formula = Outcome ~ ., family = binomial, data = filtered_data)
9
10 Coefficients:
11                               Estimate Std. Error z value Pr(>|z|)
12 (Intercept)           -1.004e+01  1.218e+00 -8.246  < 2e-16 ***
13 Pregnancies            8.216e-02  5.543e-02  1.482  0.13825
14 Glucose                3.827e-02  5.768e-03  6.635 3.24e-11 ***
15 BloodPressure          -1.420e-03  1.183e-02 -0.120  0.90446
16 SkinThickness           1.122e-02  1.708e-02  0.657  0.51128
17 Insulin                -8.253e-04  1.306e-03 -0.632  0.52757
18 BMI                     7.054e-02  2.734e-02  2.580  0.00989 **
19 DiabetesPedigreeFunction 1.141e+00  4.274e-01  2.669  0.00760 **
20 Age                      3.395e-02  1.838e-02  1.847  0.06474 .
21 ---
22 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
23
24 (Dispersion parameter for binomial family taken to be 1)
25
26 Null deviance: 498.10  on 391  degrees of freedom
27 Residual deviance: 344.02  on 383  degrees of freedom
28 AIC: 362.02
29
30 Number of Fisher Scoring iterations: 5
31
```

```

32 > raw_model <- glm(Outcome ~ ., data = data, family = binomial)
33 > summary(raw_model)
34
35 Call:
36 glm(formula = Outcome ~ ., family = binomial, data = data)
37
38 Coefficients:
39                         Estimate Std. Error z value Pr(>|z|)
40 (Intercept)           -8.4046964  0.7166359 -11.728 < 2e-16 ***
41 Pregnancies            0.1231823  0.0320776   3.840 0.000123 ***
42 Glucose                 0.0351637  0.0037087   9.481 < 2e-16 ***
43 BloodPressure          -0.0132955  0.0052336  -2.540 0.011072 *
44 SkinThickness           0.0006190  0.0068994   0.090 0.928515
45 Insulin                -0.0011917  0.0009012  -1.322 0.186065
46 BMI                     0.0897010  0.0150876   5.945 2.76e-09 ***
47 DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 **
48 Age                     0.0148690  0.0093348   1.593 0.111192
49 ---
50 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
51
52 (Dispersion parameter for binomial family taken to be 1)
53
54 Null deviance: 993.48 on 767 degrees of freedom
55 Residual deviance: 723.45 on 759 degrees of freedom
56 AIC: 741.45
57
58 Number of Fisher Scoring iterations: 5

```

在未進行資料過濾的模型 `raw_model` 中，有六個顯著影響判斷結果的解釋變數，赤池訊息量 AIC 值為 741.45，根據 Omnibus 係數測試，此模型的 Significant code 小於顯著水準 0.05，模型可以拒絕虛無假設 H_0 ，此模型至少有一個變數對結果有影響力。

		Chi-square	df	Sig.
Step 1	Step	270.039	8	<.001
	Block	270.039	8	<.001
	Model	270.039	8	<.001

Table 2: Omnibus Tests of Model Coefficients

在過濾不合理資料點的模型 `model` 中，有五個顯著影響結果的解釋變數，赤池訊息量為 344.02，遠低於使用全部資料點的 AIC 值，證明資料篩選有助於提高模型的適配度，但因為 AIC 亦與資料筆數有關，我們無法判定是刪除資料造成一致性增加抑或單純因為資料減少而使 AIC 下降。

在兩個模型中，葡萄糖、BMI 和糖尿病家族病史指標均保持顯著，表明它們與糖尿病的強烈相關性。然而，過濾零值後，模型擬合度有所改善，並且減少了潛在的虛假零值對預測變量的影響，但亦有可能因為 Insulin 欄位為 0 的資料點為 374 筆，全部刪除可能造成資料缺失嚴重，故我們分析交互作用時採預分析該項交互作用時才刪除，也實驗不同資料缺失處理方法以提升模型的準確度。

3 變數分析

3.1 One-way ANOVA

3.1.1 8 個解釋變數

以下是由全部的資料得出的結果

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > model1 <- aov(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
+                   Insulin + BMI + DiabetesPedigreeFunction + Age, data=raw_data)
4 > summary(model1)
5
6              Df Sum Sq Mean Sq F value    Pr(>F)
7 Pregnancies      1   8.59    8.59  53.638 6.16e-13 ***
8 Glucose          1  34.02   34.02 212.406 < 2e-16 ***
9 BloodPressure    1   0.12    0.12   0.771  0.380213
10 SkinThickness    1   0.86    0.86   5.393  0.020481 *
11 Insulin          1   0.26    0.26   1.594  0.207108
12 BMI              1   6.78    6.78  42.331 1.40e-10 ***
13 DiabetesPedigreeFunction 1   1.82    1.82  11.349  0.000793 ***
14 Age              1   0.46    0.46   2.865  0.090922 .
15 Residuals        759 121.57   0.16
16 ---
17 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18 >
```

以下是將有缺失值的資料移除後得到的結果

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age
13 + )
14 >
15 > diabetes_filter = diabetes[diabetes$X2!=0 & diabetes$X3!=0 & diabetes$X4!=0 &
+   diabetes$X5!=0 & diabetes$X6!=0,]
16 > model <- aov(response ~ X1+X2+X3+X4+X5+X6+X7+X8, data=diabetes_filter)
17 > summary(model)
```

```

18          Df Sum Sq Mean Sq F value    Pr(>F)
19 X1          1   5.72   5.719  38.536  1.4e-09 ***
20 X2          1  19.54  19.542 131.667 < 2e-16 ***
21 X3          1   0.30   0.303   2.040  0.154013
22 X4          1   1.68   1.683  11.338  0.000836 ***
23 X5          1   0.00   0.000   0.003  0.955436
24 X6          1   0.91   0.915   6.164  0.013464 *
25 X7          1   1.22   1.221   8.228  0.004354 **
26 X8          1   0.66   0.660   4.448  0.035591 *
27 Residuals   383  56.84   0.148
28 ---
29 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
30 >

```

從以上兩份統計報表可以看出，BloodPressure(X3) 和 Insulin(X5) 都是不重要的單一解釋變數，但他們與其他解釋變數是否存在交互作用留待3.2分析。

3.1.2 將有缺失值的欄位根據「是否為缺失值」作為 5 個解釋變數

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes_missing <- data.frame(
4 +   response = raw_data$Outcome,
5 +   Glucose = raw_data$Glucose == 0,
6 +   BloodPressure = raw_data$BloodPressure == 0,
7 +   SkinThickness = raw_data$SkinThickness == 0,
8 +   Insulin = raw_data$Insulin == 0,
9 +   BMI = raw_data$BMI == 0
10 + )
11 >
12 > model1 <- aov(response ~ Glucose + BloodPressure + SkinThickness + Insulin + BMI,
13 +                     data=diabetes_missing)
14 > summary(model1)
15
16          Df Sum Sq Mean Sq F value Pr(>F)
17 Glucose      1   0.01   0.0131   0.058 0.8102
18 BloodPressure 1   0.43   0.4320   1.902 0.1682
19 SkinThickness 1   0.27   0.2747   1.210 0.2717
20 Insulin       1   0.01   0.0059   0.026 0.8722
21 BMI           1   0.72   0.7249   3.192 0.0744 .
22 Residuals    762 173.03   0.2271
23 ---
24 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
25 >

```

3.1.3 將以上結果合併為 13 個解釋變數

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 + )
19 >
20 > model1 <- aov(response ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13, data=diabetes)
21 > summary(model1)
22
23   Df Sum Sq Mean Sq F value    Pr(>F)
24 X1      1   8.59    8.59  55.392 2.70e-13 ***
25 X2      1  34.02   34.02 219.352  < 2e-16 ***
26 X3      1   0.12    0.12   0.796 0.372541
27 X4      1   0.86    0.86   5.569 0.018531 *
28 X5      1   0.26    0.26   1.646 0.199846
29 X6      1   6.78    6.78  43.716 7.19e-11 ***
30 X7      1   1.82    1.82  11.720 0.000652 ***
31 X8      1   0.46    0.46   2.959 0.085814 .
32 X9      1   3.46    3.46  22.293 2.79e-06 ***
33 X10     1   0.36    0.36   2.316 0.128459
34 X11     1   0.06    0.06   0.387 0.534173
35 X12     1   0.08    0.08   0.524 0.469282
36 X13     1   0.67    0.67   4.299 0.038485 *
37 Residuals 754 116.94    0.16
38 ---
39 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  '  1
```

加入是否有缺失值作為解釋變數後的模型的殘差由 121.57 降至 116.94，由 p-value 為 2.79×10^{-6} 與 0.038485 可知血糖濃度的缺失值與 BMI 缺失值亦為重要變數，推論可能因為血糖濃度與 BMI 可以解釋大量變異，故是否有此二變數的資訊成為重要判斷依據。

單就 One-way ANOVA test，重要的解釋變數為懷孕次數、血糖濃度、皮膚厚度、BMI、糖尿病家族病史、年齡、與是否缺少 BMI 與血糖濃度資訊為重要解釋變數。懷孕次數增加可能會增加女性罹患妊娠糖尿病的風險，雖然通常產後自癒，但會增加女性未來罹患 2 型糖尿病的風險 [5]。

血糖濃度明顯與判定是否罹患糖尿病高度相關 [6]，皮膚厚度與 BMI 皆為衡量肥胖程度的指標，皮膚厚度主要衡量體脂肪而 BMI 衡量體重是否超標，兩者皆與罹患糖尿病風險高度相關。體脂肪、特別是內臟脂肪會導致胰島素抗阻，故肥胖指標可作為判斷罹患糖尿病風險的依據 [7]。

糖尿病家族病史亦為影響罹患糖尿病風險的重要變數，因為遺傳基因可能會影響胰島素的分泌和作用 [8]。人體對胰島素的敏感度可能隨年齡增長而降低，加之生活方式積累的後果如飲食、缺乏運動習慣等，亦可能導致罹患糖尿病的風險增加 [9]。因為 BMI 與血糖濃度為重要解釋變數，故是否缺失亦會影響判斷結果。

3.2 Two-way ANOVA

我們採用 Forward Selection 逐一檢測每個解釋變數可能的交互作用。另外，為了避免分析過程過於複雜，我們在此不考慮「是否為缺失值」的 5 個額外變數。

3.2.1 Pregnancies vs. Glucose

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose
7 + )
8 >
9 > diabetes_filter = diabetes[X2 != 0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14 Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1    8.30     8.30  49.813 3.82e-12 ***
16 X2          1  38.35    38.35 230.159 < 2e-16 ***
17 X1:X2       1    0.13     0.13   0.807    0.369
18 Residuals   759 126.48    0.17
19 ---
20 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  '  1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
```

```

23          Df Sum Sq Mean Sq F value    Pr(>F)
24 X1          1   8.30    8.30  49.83 3.79e-12 ***
25 X2          1 38.35   38.35 230.22 < 2e-16 ***
26 Residuals  760 126.61    0.17
27 ---
28 Signif. codes:  0  '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
29 >

```

懷孕次數與血糖濃度為顯著影響分類結果的重要解釋變數，但其交乘項的 p-value 為 0.369，遠大於 5% 的顯著水準，顯示懷孕次數與血糖濃度的交互作用對於判斷是否罹患糖尿病無明顯關聯，故我們可以不考慮 Pregnancies 與 Glucose 的交互作用。

3.2.2 Pregnancies vs. BloodPressure

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$BloodPressure
7 + )
8 >
9 > diabetes_filter = diabetes[X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13          Df Sum Sq Mean Sq F value    Pr(>F)
14 X1          1   8.79    8.795  41.654 1.99e-10 ***
15 X2          1   2.55    2.546  12.057 0.000546 ***
16 X1:X2       1   0.10    0.100   0.475  0.490809
17 Residuals  729 153.92    0.211
18 ---
19 Signif. codes:  0  '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
20 >
21 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
22 > summary(model2)
23          Df Sum Sq Mean Sq F value    Pr(>F)
24 X1          1   8.79    8.795  41.68  1.96e-10 ***
25 X2          1   2.55    2.546   12.06  0.000544 ***
26 Residuals  730 154.02    0.211
27 ---
28 Signif. codes:  0  '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
29 >

```

血壓與懷孕次數為判斷是否罹患糖尿病的重要變數。懷孕次數與血壓的交乘項 p-value 為 0.490809，遠大於 5% 顯著水準，顯示懷孕次數與血壓的交互作用對判定罹患糖尿病 無明顯關聯。

3.2.3 Pregnancies vs. SkinThickness

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$SkinThickness
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   8.39    8.394  43.107 1.22e-10 ***
16 X2       1   6.59    6.586  33.826 1.04e-08 ***
17 X1:X2    1   0.57    0.568   2.916   0.0883 .
18 Residuals 537 104.56    0.195
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   8.39    8.394  42.95 1.31e-10 ***
27 X2       1   6.59    6.586  33.71 1.10e-08 ***
28 Residuals 538 105.13    0.195
29 ---
```

皮膚厚度為衡量身體肥胖程度的指標，與糖尿病有高度關聯，此推論可由統計分析結果的 p-value 為 1.04×10^{-8} ，遠小於 5% 的顯著水準印證。而懷孕次數與皮膚厚度的 p-value 為 0.0883，已達 10% 的顯著水準，未達 5% 顯著水準，可由此推論懷孕較多次對婦女的 肥胖程度有些微影響，此推論亦符合常理，但由兩個模型的殘差差異極小 (0.57)，可推 定二者的交互作用項對於是否罹患糖尿病僅有些微影響，留待以整體模型的檢定判定 去留。

3.2.4 Pregnancies vs. Insulin

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Insulin
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14 Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1  5.84   5.836  30.798 5.30e-08 ***
16 X2          1  6.98   6.983  36.851 3.03e-09 ***
17 X1:X2       1  0.38   0.382   2.015    0.157
18 Residuals  390 73.91   0.190
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25 Df Sum Sq Mean Sq F value    Pr(>F)
26 X1          1  5.84   5.836  30.72 5.50e-08 ***
27 X2          1  6.98   6.983  36.76 3.16e-09 ***
28 Residuals  391 74.29   0.190
29 ---
```

因為胰島素為 0 的資料點有 374 筆，但無法判斷胰島素為 0 是資料缺失或無法量測到胰島素，使用中位數補值容易造成偏誤，故我們選擇刪除胰島素為 0 的資料點。懷孕次數與胰島素濃度皆為判斷是否罹患糖尿病的重要變數，但懷孕次數並不影響胰島素濃度，其交乘項的 p-value 為 0.157 且殘差差異極小，故建模時可忽略兩者交乘項。

3.2.5 Pregnancies vs. BMI

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
```

```

6 +     X2 = raw_data$BMI
7 +
8 >
9 > diabetes_filter = diabetes[diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14 Df Sum Sq Mean Sq F value    Pr(>F)
15 X1      1   8.08   8.085  41.143 2.5e-10 ***
16 X2      1  16.48  16.479  83.861 < 2e-16 ***
17 X1:X2   1   0.00   0.000   0.002   0.966
18 Residuals 753 147.97   0.197
19 ---
20 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25 Df Sum Sq Mean Sq F value    Pr(>F)
26 X1      1   8.08   8.085  41.20  2.43e-10 ***
27 X2      1  16.48  16.479   83.97 < 2e-16 ***
28 Residuals 754 147.97   0.196
29 ---
30 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

```

懷孕次數與身體質量指數 (BMI) 皆為重要變數，但懷孕次數與 BMI 值交互作用不直接影響判定罹患糖尿病與否，且兩個模型殘差相同，故可忽略兩者交互作用。

3.2.6 Pregnancies vs. DiabetesPedigreeFunction

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
8 >
9 > model1 <- aov(response ~ X1 * X2, data=diabetes)
10 > summary(model1)
11
12 Df Sum Sq Mean Sq F value    Pr(>F)
13 X1      1   8.59   8.591  41.021 2.63e-10 ***
14 X2      1   5.74   5.740  27.409 2.13e-07 ***
15 X1:X2   1   0.14   0.141   0.672   0.413
16 Residuals 764 160.01   0.209

```

```

16 ---
17 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18 >
19 > model2 <- aov(response ~ X1 + X2, data=diabetes)
20 > summary(model2)
21
22 Df Sum Sq Mean Sq F value    Pr(>F)
23 X1      1   8.59   8.591   41.04 2.61e-10 ***
24 X2      1   5.74   5.740   27.42 2.12e-07 ***
25 Residuals 765 160.15   0.209
26 ---
27 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

懷孕次數與糖尿病家族史皆為影響判定罹患糖尿病的重要因素，但由常理推論，兩者並無直接關聯，亦或兩者交互作用並未影響判定是否罹患糖尿病，此推論可由統計結果印證，兩者的 p-value 為 0.413，遠大於 5% 顯著水準，故在預測模型可刪除其交乘項。

3.2.7 Pregnancies vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > model1 <- aov(response ~ X1 * X2, data=diabetes)
10 > summary(model1)
11
12 Df Sum Sq Mean Sq F value    Pr(>F)
13 X1      1   8.59   8.591   40.800 2.93e-10 ***
14 X2      1   3.43   3.427   16.276 6.03e-05 ***
15 X1:X2    1   1.59   1.587    7.539  0.00618 **
16 Residuals 764 160.87   0.211
17 ---
18 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19 >
20 > model2 <- aov(response ~ X1 + X2, data=diabetes)
21 > summary(model2)
22
23 Df Sum Sq Mean Sq F value    Pr(>F)
24 X1      1   8.59   8.591   40.45 3.46e-10 ***
25 X2      1   3.43   3.427   16.14 6.47e-05 ***
26 Residuals 765 162.46   0.212
27 ---
28 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

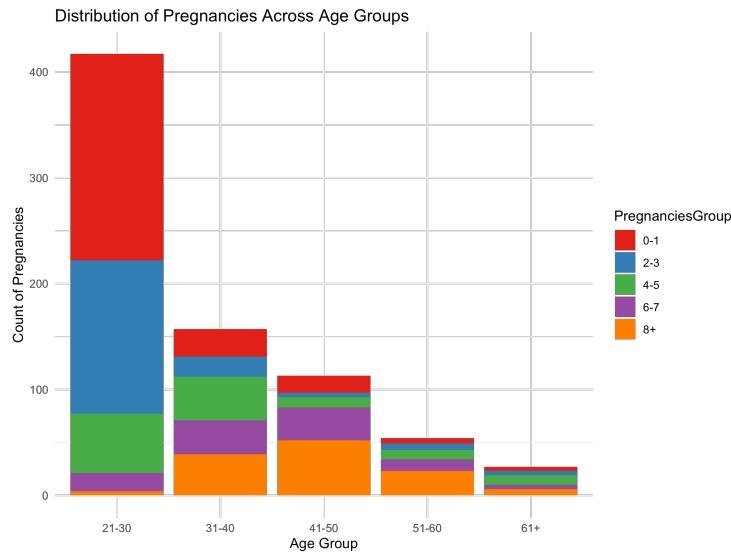


Figure 4: 懷孕次數與年齡分佈

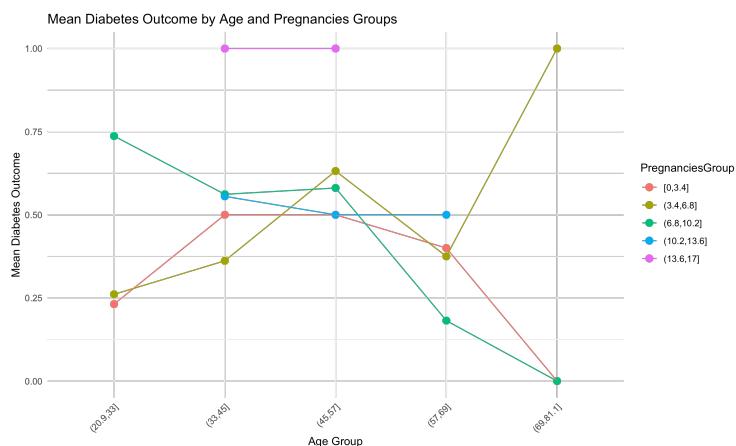


Figure 5: 不同年齡範圍和不同懷孕次數範圍對糖尿病發病率的影響

懷孕次數與年齡皆顯著影響罹患糖尿病的判斷，值得注意的是，年齡與懷孕次數的交互作用顯著，代表年齡對糖尿病的影響會因懷孕次數的不同而改變，反之亦然。圖 4, 5 為懷孕次數與年齡的分佈圖，在較年輕的組別（21 歲-30 歲）中，懷孕次數相對較多，且糖尿病的發病率可能會因懷孕次數的增加而增加。這表明對年輕女性來說，懷孕次數對糖尿病風險有較大的影響。對於中年至中老年女性，懷孕次數的分佈較為均勻，懷孕次數對糖尿病風險的影響不如年輕組明顯，原因可能為隨年齡增長，其他因素（如生活方式、飲食）對糖尿病的影響變得更加顯著。老年組中，懷孕次數相對較少，這可能是

因為大多數女性在這個年齡段已經不再懷孕。這組的糖尿病發病率更多地受到年齡本身的影響，而非懷孕次數，但亦有較極端的值存在。極高懷孕次數的糖尿病發病率在各年齡段皆為 1，代表多次懷孕會顯著增加罹患糖尿病的風險。

此圖因為將年齡與懷孕次數分成五群，造成圖片較雜亂難以判定整體交互作用，故在 3.3.2.8 中僅分成三群分析。

3.3.2.8 Glucose vs. BloodPressure

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$BloodPressure
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0 & diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14 Df Sum Sq Mean Sq F value Pr(>F)
15 X1 1 39.01 39.01 226.841 <2e-16 ***
16 X2 1 0.63 0.63 3.649 0.0565 .
17 X1:X2 1 0.00 0.00 0.001 0.9728
18 Residuals 724 124.51 0.17
19 ---
20 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25 Df Sum Sq Mean Sq F value Pr(>F)
26 X1 1 39.01 39.01 227.154 <2e-16 ***
27 X2 1 0.63 0.63 3.654 0.0563 .
28 Residuals 725 124.51 0.17
29 ---
30 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
31 >
```

血糖濃度是判定是否罹患糖尿病的重要因素 (p -value 小於 2×10^{-6})，而血壓的 p -value 為 0.0565，在 10% 顯著水準下顯著，但位於 5% 顯著水準的臨界值，可保留至討論決定去留。兩者的交互作用 p -value 為 0.97，遠超過顯著水準的 0.05，故可忽略交互作用影響。

3.2.9 Glucose vs. SkinThickness

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$SkinThickness
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0 & diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1 30.23  30.230 186.674 < 2e-16 ***
16 X2          1  2.51   2.505  15.469 9.5e-05 ***
17 X1:X2       1  0.00   0.000   0.002   0.964
18 Residuals  532  86.15   0.162
19
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1          1 30.23  30.230 187.0 < 2e-16 ***
27 X2          1  2.51   2.505   15.5 9.35e-05 ***
28 Residuals  533  86.15   0.162
29
```

血糖濃度與皮膚厚度皆為判定是否罹患糖尿病的重要指標，但是兩者交乘項的 p-value 為 0.964，顯示兩者的交互作用項對於判定是否罹患糖尿病影響較小，可忽略。

3.2.10 Glucose vs. Insulin

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$Insulin
7 + )
8 >
```

```

9 > diabetes_filter = diabetes[diabetes$X1 != 0 & diabetes$X2 != 0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value Pr(>F)
15 X1       1  23.13  23.131 141.312 <2e-16 ***
16 X2       1    0.00    0.001    0.005   0.945
17 X1:X2    1    0.19    0.193    1.179   0.278
18 Residuals 389  63.67    0.164
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value Pr(>F)
26 X1       1  23.13  23.131 141.247 <2e-16 ***
27 X2       1    0.00    0.001    0.005   0.945
28 Residuals 390  63.87    0.164
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
31 >

```

在此資料集中，血糖對於判定糖尿病患者有顯著性的影響，但是在已知血糖濃度的情況下，胰島素水平對糖尿病結果並無顯著影響。表示在控制了血糖變數後，胰島素對結果的影響不顯著，因為血糖的解釋力非常強，幾乎完全解釋了結果的變異，導致胰島素變數在這個情境下顯得不顯著。

我們懷疑可能是刪掉胰島素為 0 的資料造成的偏誤，故我們同時檢測了原始資料血糖濃度與胰島素的交互作用，得到結果如下。

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$Insulin
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value Pr(>F)
15 X1       1  42.39   42.39 246.252 <2e-16 ***
16 X2       1    0.20    0.20   1.167   0.280
17 X1:X2    1    0.00    0.00   0.010   0.919

```

```

17 Residuals    759 130.67     0.17
18 ---
19 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
20 >
21 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
22 > summary(model2)
23
24      Df Sum Sq Mean Sq F value Pr(>F)
25 X1          1   42.39    42.39 246.573 <2e-16 ***
26 X2          1    0.20     0.20   1.169   0.28
27 Residuals  760 130.67     0.17
28 ---
29 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

```

胰島素的 p-value 大幅下降，但是仍未達成為顯著變數的標準，更遑論其交乘項。

3.2.11 Glucose vs. BMI

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$BMI
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0 & diabetes$X2!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value Pr(>F)
15 X1          1   41.84    41.84 255.464 < 2e-16 ***
16 X2          1    6.88     6.88  41.989 1.66e-10 ***
17 X1:X2       1    0.09     0.09   0.572     0.45
18 Residuals  748 122.51     0.16
19 ---
20 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value Pr(>F)
26 X1          1   41.84    41.84 255.61 < 2e-16 ***
27 X2          1    6.88     6.88  42.01 1.64e-10 ***
28 Residuals  749 122.60     0.16
29 ---
30 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

```

血糖濃度與 BMI 皆為判定是否罹患糖尿病的重要變數，但是兩者的交互作用並未造成顯著影響，且兩者已經解釋幾乎所有變異，故可以不考慮兩者交互作用。

3.2.12 Glucose vs. DiabetesPedigreeFunction

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14 Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1  42.39   42.39 250.857 < 2e-16 ***
16 X2          1   1.92    1.92  11.390 0.000776 ***
17 X1:X2       1   0.68    0.68   4.003 0.045788 *
18 Residuals  759 128.27    0.17
19 ---
20 Signif. codes:  0  ***
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25 Df Sum Sq Mean Sq F value    Pr(>F)
26 X1          1  42.39   42.39 249.87 < 2e-16 ***
27 X2          1   1.92    1.92   11.35 0.000795 ***
28 Residuals  760 128.95    0.17
29 ---
30 Signif. codes:  0  ***
31 >
```

血糖濃度與糖尿病家族病史皆為影響判定罹患糖尿病的重要因素，其各自 p-value 遠小於 5% 顯著水準，且兩者存在強烈交互作用，可能解釋為具有糖尿病家族史的人較容易有較高血糖，導致兩者關聯度高。圖 6 為不同血糖濃度下的糖尿病譜系功能（Diabetes Pedigree Function）平均值，隨著血糖濃度增加，DiabetesPedigreeFunction 的平均值亦增長，血糖較高者可能具有較高的遺傳糖尿病風險，因為糖尿病譜系功能是一個考量遺傳

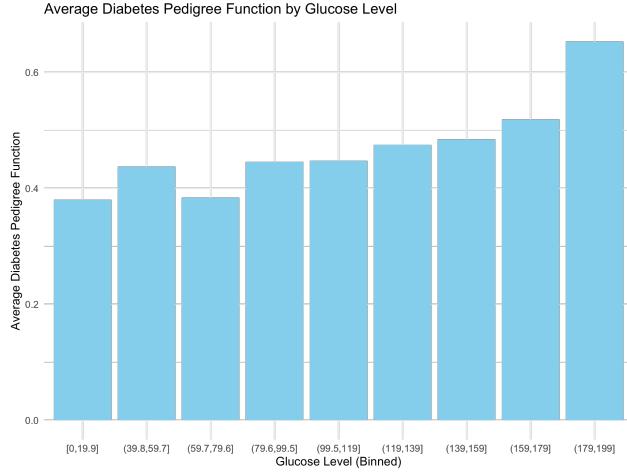


Figure 6: 不同範圍的血糖濃度 (Glucose) 水平對糖尿病譜系功能 (Diabetes Pedigree Function) 平均值的影響

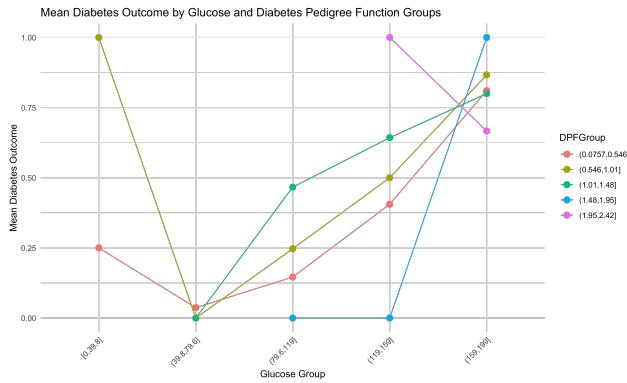


Figure 7: 不同血糖濃度範圍和糖尿病家族史功能 (Diabetes Pedigree Function, DPF) 範圍對糖尿病發病率的影響

和家族史的指標。高血糖與高糖尿病譜系功能的正相關，反映出這些個體的糖尿病風險可能更高。

圖 7 為不同血糖濃度範圍和家族病史對發病率的影響，在低血糖組中，有較多糖尿病家族史的人比較容易罹患糖尿病，但是在高血糖群組中，擁有較多糖尿病家族史者反而罹患糖尿病的比例較低，我們推測可能因為家族內糖尿病患者較多、且高血糖已引起個體對自身健康狀況的重視而開始控制飲食、生活型態等。

3.2.13 Glucose vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Glucose,

```

```

6 +     X2 = raw_data$Age
7 +
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1  42.39   42.39 250.938 < 2e-16 ***
16 X2       1   1.98    1.98 11.718 0.000652 ***
17 X1:X2    1   0.66    0.66  3.925 0.047948 *
18 Residuals 759 128.23    0.17
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1  42.39   42.39 249.98 < 2e-16 ***
27 X2       1   1.98    1.98 11.67 0.000668 ***
28 Residuals 760 128.89    0.17
29 ---
30 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

血糖與年齡皆為影響判定是否為糖尿病患者的重要因素，且年齡與血糖的交乘項 p-value 為 0.000668，顯示其在判定是否罹患糖尿病的重要性。

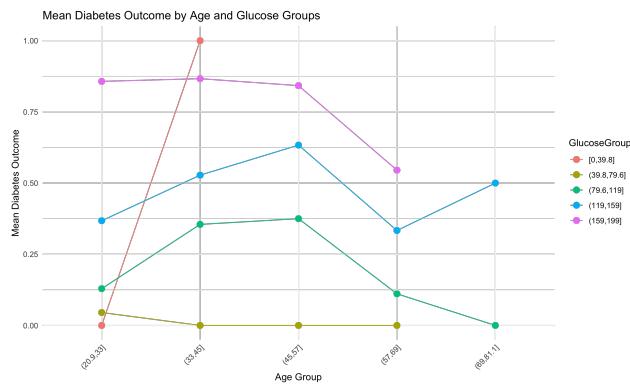


Figure 8: 根據年齡群組和葡萄糖群組計算糖尿病結果的平均值

圖 8 為不同年齡範圍和不同葡萄糖範圍對糖尿病發病率的影響，y 軸為糖尿病結果 (Outcome) 的平均值，圖中各血糖範圍所繪出的線不平行，代表血糖與年齡對是否罹患糖尿病有交互作用。

在所有年齡段中，高血糖的糖尿病發病率相對較高。在 $(33, 45]$ 年齡群組中，高葡萄糖水平的糖尿病發病率接近 0.75，並且隨著年齡的增長，發病率呈下降趨勢。中低血糖的個體在各個年齡段的發病率均低，代表可能是一般健康個體的血糖值。

3.2.14 BloodPressure vs. SkinThickness

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BloodPressure,
6 +   X2 = raw_data$SkinThickness
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0 & diabetes$X2 != 0, ]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1      1  4.12   4.119  20.159 8.73e-06 ***
16 X2      1  5.94   5.940  29.072 1.05e-07 ***
17 X1:X2   1  0.18   0.179   0.878     0.349
18 Residuals 535 109.32   0.204
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value    Pr(>F)
26 X1      1  4.12   4.119  20.16 8.71e-06 ***
27 X2      1  5.94   5.940  29.08 1.04e-07 ***
28 Residuals 536 109.50   0.204
29 ---
```

如上述，血壓與皮膚厚度在沒有更顯著、直接影響糖尿病判斷的變數下為顯著變數，但是血壓與皮膚厚度的 p-value 為 0.349，遠大於 5% 顯著水準，因此須接受虛無假設 H_0 ，兩者並無顯著交互作用。

3.2.15 BloodPressure vs. Insulin

```
> raw_data <- read.csv("diabetes.csv")
```

```

2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BloodPressure,
6 +   X2 = raw_data$Insulin
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0 & diabetes$X2 != 0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   3.24   3.239  16.505 5.87e-05 ***
16 X2       1   7.12   7.119  36.284 3.95e-09 ***
17 X1:X2    1   0.23   0.225   1.148     0.285
18 Residuals 390  76.52   0.196
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   3.24   3.239  16.50 5.88e-05 ***
27 X2       1   7.12   7.119  36.27 3.96e-09 ***
28 Residuals 391  76.75   0.196
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
31 >

```

僅用血壓與胰島素預測是否罹患糖尿病時，兩者皆為解釋變異的重要變數，但兩者的交互作用並未對判斷是否罹患糖尿病有顯著影響，p-value 0.285 大於顯著水準 0.05，因此虛無假設 H_0 成立，兩個變數的交互作用對 Outcome 無明顯作用。

3.2.16 BloodPressure vs. BMI

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BloodPressure,
6 +   X2 = raw_data$BMI
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0 & diabetes$X2 != 0,]
10 >

```

```

11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1  4.66   4.659  22.825 2.15e-06 ***
16 X2       1 11.66  11.658  57.117 1.24e-13 ***
17 X1:X2    1  0.29   0.290   1.422     0.234
18 Residuals 725 147.97   0.204
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1  4.66   4.659  22.81 2.16e-06 ***
27 X2       1 11.66  11.658  57.08 1.26e-13 ***
28 Residuals 726 148.26   0.204
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

僅以血壓與 BMI 值判定罹患糖尿病與否時，兩者皆為顯著影響判斷的重要變數，但兩者的交互作用對判定是否罹患糖尿病影響甚微，p-value 0.234 大於顯著水準 0.05，故建模時可忽略不計。

3.2.17 BloodPressure vs. DiabetesPedigreeFunction

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BloodPressure,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1  4.81   4.812  22.59 2.41e-06 ***
16 X2       1  5.28   5.284  24.81 7.92e-07 ***
17 X1:X2    1  0.00   0.000   0.00     0.991
18 Residuals 729 155.27   0.213
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

```

20 >
21 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
22 > summary(model2)
23
24   Df Sum Sq Mean Sq F value    Pr(>F)
25 X1       1  4.81    4.812  22.62 2.37e-06 ***
26 X2       1  5.28    5.284  24.84 7.78e-07 ***
27 Residuals 730 155.27   0.213
28 ---
29 Signif. codes:  0  ***
30

```

血壓與家族病史在判斷罹患糖尿病佔顯著影響，其 p-value 遠小於顯著水準 0.05，但兩者交乘項的 p-value 為 0.991，遠大於顯著水準 0.05，無法拒絕虛無假設 H_0 ，故在判斷是否罹患糖尿病時可忽略不計交乘項的影響。

3.2.18 BloodPressure vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BloodPressure,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1  4.81    4.812  22.853 2.12e-06 ***
16 X2       1  6.85    6.846  32.511 1.72e-08 ***
17 X1:X2    1  0.20    0.198   0.938     0.333
18 Residuals 729 153.51   0.211
19 ---
20 Signif. codes:  0  ***
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1  4.81    4.812  22.86 2.11e-06 ***
27 X2       1  6.85    6.846  32.51 1.72e-08 ***
28 Residuals 730 153.71   0.211
29 ---
30 Signif. codes:  0  ***

```

血壓與年齡皆為影響判斷罹患糖尿病的重要解釋變數，p-value 皆遠小於顯著水準 0.05，但其交互作用項 p-value 為 0.333，大於顯著水準 0.05，故我們無法拒絕虛無假設 H_0 ，兩者在判斷是否罹患糖尿病不存在明顯交互作用。

3.2.19 SkinThickness vs. Insulin

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$SkinThickness,
6 +   X2 = raw_data$Insulin
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0 & diabetes$X2!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   5.79   5.792  29.981 7.83e-08 ***
16 X2       1   5.90   5.901  30.545 5.98e-08 ***
17 X1:X2    1   0.08   0.077   0.398     0.528
18 Residuals 390  75.34   0.193
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   5.79   5.792  30.03 7.65e-08 ***
27 X2       1   5.90   5.901  30.59 5.84e-08 ***
28 Residuals 391  75.41   0.193
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
31 >

```

皮膚厚度與胰島素濃度在單獨判定是否罹患糖尿病時為重要解釋變數，能解釋大部分變異，但其交乘項對判定是否罹患糖尿病的貢獻甚微 (p-value 為 0.528，大於顯著水準 0.05，接受虛無假設)，顯示皮膚厚度與胰島素濃度的交互作用在判定是否罹患糖尿病獨立，可忽略交乘項。

3.2.20 SkinThickness vs. BMI

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$SkinThickness,
6 +   X2 = raw_data$BMI
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0 & diabetes$X2!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   8.00   8.002  39.883 5.67e-10 ***
16 X2       1   4.12   4.125  20.557 7.15e-06 ***
17 X1:X2    1   0.42   0.418   2.082     0.15
18 Residuals 535 107.34   0.201
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   8.00   8.002  39.80 5.89e-10 ***
27 X2       1   4.12   4.125  20.52 7.29e-06 ***
28 Residuals 536 107.76   0.201
29 ---
```

皮膚厚度與 BMI 皆為判斷是否罹患糖尿病的重要變數，但是兩者的交互作用對於判斷罹患糖尿病與否貢獻較小，p-value 為 0.15，高於顯著水準 0.05，可以忽略其交互作用項。

3.2.21 SkinThickness vs. DiabetesPedigreeFunction

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$SkinThickness,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
```

```

8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   8.09   8.088  40.454 4.31e-10 ***
16 X2       1   4.66   4.658  23.300 1.81e-06 ***
17 X1:X2    1   0.00   0.005   0.024    0.876
18 Residuals 537 107.36   0.200
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25   Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   8.09   8.088  40.53 4.16e-10 ***
27 X2       1   4.66   4.658  23.34 1.77e-06 ***
28 Residuals 538 107.36   0.200
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
31 >

```

皮膚厚度與糖尿病家族病史在判斷罹患糖尿病具有顯著影響，但是其交乘項 p-value 為 0.876，高於顯著水準 0.05，故無法拒絕虛無假設 H_0 ，交互作用影響可以忽略不計。

3.2.22 SkinThickness vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$SkinThickness,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14   Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   8.09   8.088  42.718 1.47e-10 ***
16 X2       1   9.74   9.736  51.421 2.48e-12 ***
17 X1:X2    1   0.62   0.617   3.259   0.0716 .
18 Residuals 537 101.67   0.189

```

```

18 ---  

19 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

20 >  

21 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)  

22 > summary(model2)  

23  

24   Df Sum Sq Mean Sq F value    Pr(>F)  

25 X1       1  8.09   8.088  42.54 1.60e-10 ***  

26 X2       1  9.74   9.736  51.21 2.74e-12 ***  

27 Residuals 538 102.29   0.190  

28 ---  

29 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

>

```

皮膚厚度和年紀皆為影響判定罹患糖尿病的重要因素。結果顯示，這兩者之間的交互作用項 p-value 為 0.0716，略大於 5% 的顯著水準，但小於 10% 的顯著水準，表明這兩者之間存在一定的交互作用。兩個模型的殘差差異極小（約為 0.62），說明交互作用項對於是否罹患糖尿病的影響雖然顯著，但整體影響有限。由於皮膚厚度資料分布不均，

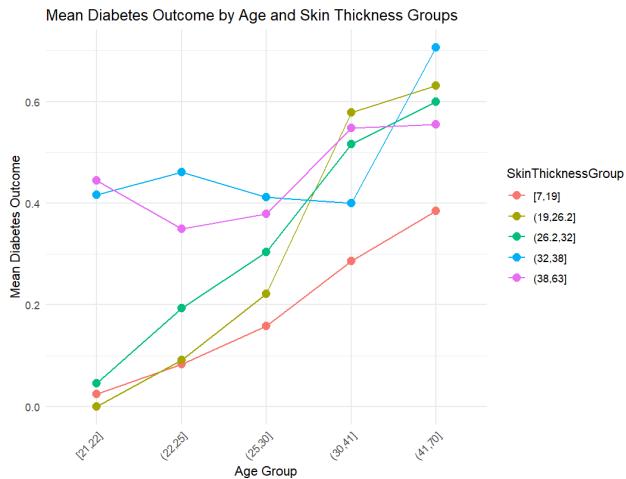


Figure 9: 過濾皮膚厚度為零的資料後，使用等頻分群皮膚厚度和年紀，不同皮膚厚度範圍和年紀範圍對糖尿病發病率的影響。

有 227 筆皆為零，因此圖 9 使用去除零值並等頻分群皮膚厚度和年紀後的資料來觀察有無交互作用。可以看到圖中各皮膚厚度組別雖然並不平行，但罹患糖尿病的機率大致上都隨著年齡增長而提升，表示存在一定的交互作用。可以看出皮膚厚度較高的兩組隨年齡增加罹患糖尿病的機率變化相對小，基本都維持在較高的得病機率，而皮膚厚度最薄的組別得病機率均較低。根據衛福部的參考標準，女性正常範圍落在 12.0 到 25.5 之間，由圖 9，正常偏肥的個體因年齡而改變的患病機率最多，代表此群體是否罹患糖尿病對年齡較敏感，在年輕的群組中，皮膚厚度較厚的兩組患病機率明顯高於較薄的三組，而

隨年齡增長，患病機率逐漸靠近。

3.2.23 Insulin vs. BMI

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Insulin,
6 +   X2 = raw_data$BMI
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0 & diabetes$X2 != 0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   7.97   7.975  41.781 3.06e-10 ***
16 X2       1   3.75   3.747  19.633 1.22e-05 ***
17 X1:X2    1   1.03   1.028   5.384   0.0208 *
18 Residuals 389  74.25   0.191
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   7.97   7.975  41.32 3.78e-10 ***
27 X2       1   3.75   3.747  19.41 1.36e-05 ***
28 Residuals 390  75.28   0.193
29 ---
```

胰島素水平和 BMI 皆為影響判定罹患糖尿病的重要因素。結果顯示，這兩者之間的交互作用項 p-value 為 0.0208，小於 5% 的顯著水準，表明它們之間存在顯著的交互作用。兩個模型的殘差差異約為 1.03，代表交互作用項對於是否罹患糖尿病的影響雖然顯著，但整體影響有限。

由於胰島素濃度的資料分布大多集中於 $300(\mu\text{U/mL})$ 以下，並且有 374 筆胰島素濃度資料為零，因此圖 10 使用去除零值並等頻分群胰島素濃度後的資料來觀察有無交互作用，至於 BMI 則按照 WHO 的標準分為過輕、正常、過重和肥胖。可以看到圖中各 BMI 組別所繪出的線彼此並不平行，說明交互作用的存在。對於正常體重的人，胰島素

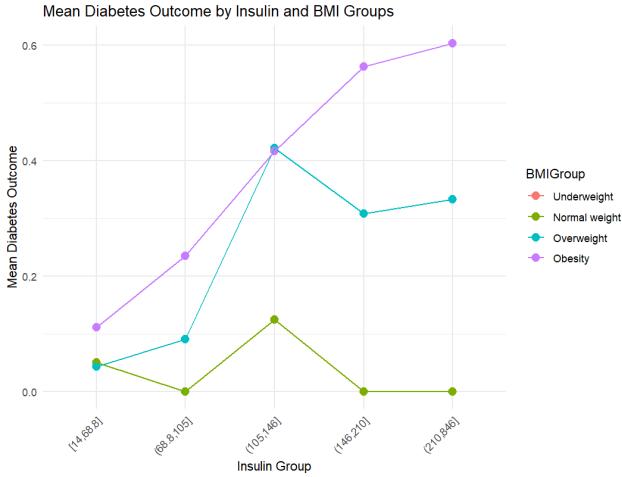


Figure 10: 過濾胰島素濃度為零的資料後，使用等頻分群胰島素濃度，不同胰島素濃度範圍和 BMI 範圍對糖尿病發病率的影響。

濃度高低對罹患糖尿病並沒有太大影響，而過重和肥胖的人罹患糖尿病的機率基本上是隨胰島素濃度上升而提高。

3.2.24 Insulin vs. DiabetesPedigreeFunction

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Insulin,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1   8.02   8.021  41.781 3.05e-10 ***
16 X2       1   2.30   2.302  11.993 0.000593 ***
17 X1:X2    1   1.91   1.911   9.955 0.001729 **
18 Residuals 390  74.87   0.192
19
20 Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1   8.02   8.021   40.85 4.69e-10 ***
27 X2       1   2.30   2.302   11.72 0.000682 ***

```

```

26 Residuals   391  76.78    0.196
27 ---
28 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
29 >

```

胰島素水平和糖尿病家族史在僅以此二變數判斷是否罹患糖尿病時有顯著影響，且兩者交乘項的 p-value 為 0.001729，顯著小於 5% 的顯著水準，表明這兩者之間存在強烈的交互作用。然而，兩個模型的殘差差異約為 1.91，說明交互作用項對於是否罹患糖尿病的影響雖然顯著，但整體影響有限。

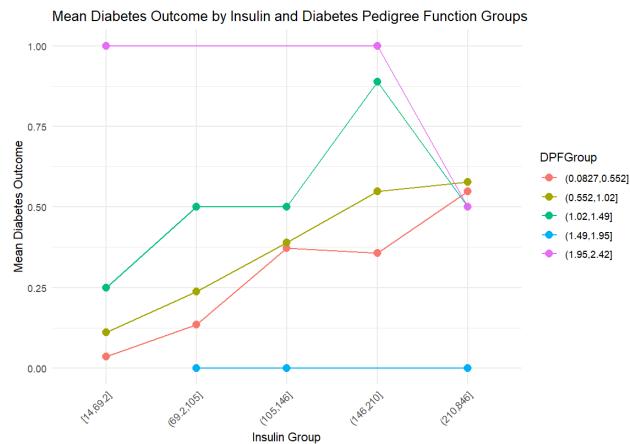


Figure 11: 過濾胰島素濃度為零的資料後，使用等頻分群胰島素濃度，不同胰島素濃度範圍和 DPF 範圍對糖尿病發病率的影響。

圖11一樣使用去除零值並等頻分群胰島素濃度的資料來觀察有無交互作用。可以看到圖中各 DPF 範圍所繪出的線彼此並不平行，說明交互作用的存在。對於低胰島素濃度的人來說，有較多糖尿病家族史的人比較容易罹患糖尿病，但是對高胰島素濃度 (210, 846] 的人，糖尿病家族史相較之下對罹患糖尿病的機率影響較小。

3.2.25 Insulin vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Insulin,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0,]
10 >

```

```

11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14          Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1   8.02   8.021  43.699 1.26e-10 ***
16 X2          1   7.50   7.496  40.837 4.72e-10 ***
17 X1:X2       1   0.00   0.004   0.022     0.883
18 Residuals  390  71.59   0.184
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25          Df Sum Sq Mean Sq F value    Pr(>F)
26 X1          1   8.02   8.021  43.81 1.20e-10 ***
27 X2          1   7.50   7.496  40.94 4.49e-10 ***
28 Residuals  391  71.59   0.183
29 ---
30 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
31 >

```

胰島素水平和年紀都對判定是否罹患糖尿病有顯著影響，但兩者交乘項的 p-value 為 0.883，遠大於 5% 顯著水準且殘差差異極小，故預測模型可忽略兩者交乘項。

3.2.26 BMI vs. DiabetesPedigreeFunction

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BMI,
6 +   X2 = raw_data$DiabetesPedigreeFunction
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1 != 0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14          Df Sum Sq Mean Sq F value    Pr(>F)
15 X1          1 16.98   16.976  83.862 < 2e-16 ***
16 X2          1   3.03    3.028  14.960 0.000119 ***
17 X1:X2       1   0.10    0.096   0.473 0.491684
18 Residuals  753 152.43   0.202
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
21 >

```

```

21 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
22 > summary(model2)
23
24      Df Sum Sq Mean Sq F value    Pr(>F)
25 X1       1 16.98  16.976   83.92 < 2e-16 ***
26 X2       1  3.03   3.028   14.97 0.000119 ***
27 Residuals 754 152.53   0.202
28 ---
29 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

BMI 和糖尿病家族史皆為影響判定罹患糖尿病的重要因素，兩者交乘項的 p-value 為 0.492，遠大於 5% 顯著水準，因此不能拒絕無交互作用的虛無假設。

3.2.27 BMI vs. Age

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$BMI,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > diabetes_filter = diabetes[diabetes$X1!=0,]
10 >
11 > model1 <- aov(response ~ X1 * X2, data=diabetes_filter)
12 > summary(model1)
13
14      Df Sum Sq Mean Sq F value    Pr(>F)
15 X1       1 16.98  16.976   87.423 < 2e-16 ***
16 X2       1  8.97   8.969   46.187 2.19e-11 ***
17 X1:X2    1  0.37   0.365   1.881   0.171
18 Residuals 753 146.22   0.194
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21 >
22 > model2 <- aov(response ~ X1 + X2, data=diabetes_filter)
23 > summary(model2)
24
25      Df Sum Sq Mean Sq F value    Pr(>F)
26 X1       1 16.98  16.976   87.32 < 2e-16 ***
27 X2       1  8.97   8.969   46.13 2.24e-11 ***
28 Residuals 754 146.59   0.194
29 ---
30 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

BMI 和年紀皆對判定罹患糖尿病有顯著影響，兩者交乘項的 p-value 為 0.171，大於 5% 顯著水準，表示兩者並無明顯交互作用，因此視為獨立。

3.2.28 DiabetesPedigreeFunction vs. Age

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$DiabetesPedigreeFunction,
6 +   X2 = raw_data$Age
7 + )
8 >
9 > model1 <- aov(response ~ X1 * X2, data=diabetes)
10 > summary(model1)
11
12      Df Sum Sq Mean Sq F value    Pr(>F)
13 X1          1  5.27   5.273  25.248 6.28e-07 ***
14 X2          1  9.44   9.444  45.220 3.45e-11 ***
15 X1:X2       1  0.20   0.201   0.964    0.326
16 Residuals  764 159.56   0.209
17 ---
18 Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1
19 >
20 > model2 <- aov(response ~ X1 + X2, data=diabetes)
21 > summary(model2)
22
23      Df Sum Sq Mean Sq F value    Pr(>F)
24 X1          1  5.27   5.273  25.25 6.28e-07 ***
25 X2          1  9.44   9.444  45.22 3.44e-11 ***
26 Residuals  765 159.76   0.209
27 ---
```

糖尿病家族史和年紀皆為影響判定罹患糖尿病的重要因素，兩者交乘項的 p-value 為 0.326，遠大於 5% 顯著水準，故在預測模型可刪除其交乘項。

3.3 AIC

我們比較不考慮交互作用所擬合出來的 Logistic Regression Model 與考慮不同交互作用下的 Logistic Regression Model。

3.3.1 不考慮交互作用

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 + )
19 >
```

3.3.1.1 只用原始的 8 個解釋變數的模型

```
1 > model1 <- glm(response ~ X1+X2+X3+X4+X5+X6+X7+X8, data=diabetes, family=binomial)
2 > summary(model1)
3
4 Call:
5 glm(formula = response ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,
6     family = binomial, data = diabetes)
7
8 Coefficients:
9             Estimate Std. Error z value Pr(>|z|)
10 (Intercept) -8.4046964  0.7166359 -11.728 < 2e-16 ***
11 X1           0.1231823  0.0320776   3.840 0.000123 ***
12 X2           0.0351637  0.0037087   9.481 < 2e-16 ***
13 X3          -0.0132955  0.0052336  -2.540 0.011072 *
14 X4           0.0006190  0.0068994   0.090 0.928515
15 X5          -0.0011917  0.0009012  -1.322 0.186065
16 X6           0.0897010  0.0150876   5.945 2.76e-09 ***
17 X7           0.9451797  0.2991475   3.160 0.001580 **
18 X8           0.0148690  0.0093348   1.593 0.111192
19 ---
20 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```

21
22 (Dispersion parameter for binomial family taken to be 1)
23
24 Null deviance: 993.48 on 767 degrees of freedom
25 Residual deviance: 723.45 on 759 degrees of freedom
26 AIC: 741.45
27
28 Number of Fisher Scoring iterations: 5
29
30 >

```

3.3.1.2 使用原始的 8 個解釋變數以及額外的 5 個解釋變數的結果

```

1 > model2 <- glm(response ~ ., data=diabetes, family=binomial)
2 > summary(model2)
3
4 Call:
5 glm(formula = response ~ ., family = binomial, data = diabetes)
6
7 Coefficients:
8             Estimate Std. Error z value Pr(>|z|)
9 (Intercept) -9.3826661  0.8313109 -11.287 < 2e-16 ***
10 X1          0.1244084  0.0325203   3.826  0.00013 ***
11 X2          0.0378306  0.0039461   9.587 < 2e-16 ***
12 X3         -0.0104368  0.0087623  -1.191  0.23361
13 X4          0.0040094  0.0134387   0.298  0.76544
14 X5         -0.0006452  0.0011822  -0.546  0.58526
15 X6          0.0959924  0.0180916   5.306 1.12e-07 ***
16 X7          0.9765315  0.3059045   3.192  0.00141 **
17 X8          0.0121485  0.0096162   1.263  0.20647
18 X9TRUE      5.0512899  1.1653165   4.335 1.46e-05 ***
19 X10TRUE     0.2594184  0.7981229   0.325  0.74515
20 X11TRUE     0.0800522  0.4974750   0.161  0.87216
21 X12TRUE     0.2369177  0.3280111   0.722  0.47012
22 X13TRUE     2.2086153  1.0449557   2.114  0.03455 *
23 ---
24 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
25
26 (Dispersion parameter for binomial family taken to be 1)
27
28 Null deviance: 993.48 on 767 degrees of freedom
29 Residual deviance: 704.02 on 754 degrees of freedom
30 AIC: 732.02
31

```

```

32 Number of Fisher Scoring iterations: 5
33
34 >

```

僅用 8 個解釋變數擬合 Logistic Regression 時，Residual Deviance 為 723.45，而新增 5 個缺失值的指標後 Residual Deviance 略降為 704.02，代表模型的擬合效果有提升。

3.3.1.3 將 13 個解釋變數中不具統計顯著性的解釋變數刪除後的結果

```

1 > model3 <- glm(response ~ X1+X2+X6+X7+X9+X13, data=diabetes, family=binomial)
2 > summary(model3)
3
4 Call:
5 glm(formula = response ~ X1 + X2 + X6 + X7 + X9 + X13, family = binomial,
6      data = diabetes)
7
8 Coefficients:
9             Estimate Std. Error z value Pr(>|z|)
10 (Intercept) -9.182605   0.706620 -12.995 < 2e-16 ***
11 X1           0.144023   0.027579   5.222 1.77e-07 ***
12 X2           0.036880   0.003497  10.547 < 2e-16 ***
13 X6           0.088448   0.014718   6.009 1.86e-09 ***
14 X7           0.888753   0.295444   3.008  0.00263 **
15 X9TRUE       5.043238   1.117864   4.511 6.44e-06 ***
16 X13TRUE      2.516103   0.973756   2.584  0.00977 **
17 ---
18 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
19
20 (Dispersion parameter for binomial family taken to be 1)
21
22 Null deviance: 993.48 on 767 degrees of freedom
23 Residual deviance: 715.70 on 761 degrees of freedom
24 AIC: 729.7
25
26 Number of Fisher Scoring iterations: 5
27
28 >

```

刪除不顯著的解釋變數後，模型的 AIC 相較於使用 13 個解釋變數的模型有些微下降，且在僅使用七個變數的情況下模型的 Residual Deviance 比使用了八個變數的模型 (3.3.1.1) 小，顯示選擇顯著影響的變數有助提升模型的預測能力且與複雜度稍稍取得平衡。

3.3.2 考慮交互作用

3.3.2.1 考慮全部交互作用項

以下為直接將 8 個原始解釋變數相乘的結果，可看出此模型無法使用，AIC 的數值極大，模型複雜度過高，因此選擇變數有其必要性。

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age
13 + )
14 >
15 > model1 <- glm(response ~ X1*X2*X3*X4*X5*X6*X7*X8, data=diabetes, family=binomial)
16 警告訊息：
17 glm.fit: 擬合機率算出來是數值零或一
18 > summary(model1)

19
20 Call:
21 glm(formula = response ~ X1 * X2 * X3 * X4 * X5 * X6 * X7 * X8,
22       family = binomial, data = diabetes)
23
24 Coefficients:
25                               Estimate Std. Error     z value Pr(>|z|)
26 (Intercept)           -2.302e+17  3.017e+09 -76305063 <2e-16 ***
27 X1                      7.748e+16  9.902e+08  78250547 <2e-16 ***
28 X2                      1.381e+15  2.325e+07  59421613 <2e-16 ***
29 X3                      2.535e+15  4.340e+07  58407889 <2e-16 ***
30 X4                      8.038e+15  5.128e+08  15675498 <2e-16 ***
31 X5                     -3.053e+16  5.550e+08 -55018180 <2e-16 ***
32 X6                      7.876e+15  1.084e+08  72651881 <2e-16 ***
33 X7                      8.594e+17  8.880e+09  96770541 <2e-16 ***
34 X8                      9.888e+15  1.160e+08  85278278 <2e-16 ***
35 X1:X2                  -5.515e+14  8.253e+06 -66824425 <2e-16 ***
36 X1:X3                  -8.355e+14  1.322e+07 -63222012 <2e-16 ***
37 X2:X3                  -1.419e+13  3.385e+05 -41933421 <2e-16 ***
38 X1:X4                  -8.980e+15  1.704e+08 -52696978 <2e-16 ***
39 X2:X4                  -1.050e+14  4.317e+06 -24325837 <2e-16 ***
40 X3:X4                  -1.828e+13  6.845e+06 -2670422 <2e-16 ***
41 X1:X5                  -1.582e+15  1.374e+08 -11516184 <2e-16 ***
```

```

42 X2:X5          2.815e+14  4.473e+06   62929180 <2e-16 ***
43 X3:X5          3.718e+14  7.704e+06   48260289 <2e-16 ***
44 X4:X5          8.399e+14  1.821e+07   46131614 <2e-16 ***
45 X1:X6          -2.682e+15  3.479e+07  -77097323 <2e-16 ***
46 X2:X6          -4.646e+13  8.397e+05  -55331147 <2e-16 ***
47 X3:X6          -8.892e+13  1.485e+06  -59886293 <2e-16 ***
48 X4:X6          -4.577e+14  1.584e+07  -28896327 <2e-16 ***
49 X5:X6          1.070e+15  1.732e+07   61803057 <2e-16 ***
50 X1:X7          -3.582e+17  3.508e+09  -102088159 <2e-16 ***
51 X2:X7          -6.593e+15  7.434e+07  -88680725 <2e-16 ***
52 X3:X7          -1.036e+16  1.217e+08  -85182105 <2e-16 ***
53 X4:X7          1.089e+16  9.966e+08   10928455 <2e-16 ***
54 X5:X7          5.895e+16  1.143e+09   51566914 <2e-16 ***
55 X6:X7          -3.403e+16  3.516e+08  -96765171 <2e-16 ***
56 X1:X8          -3.024e+15  2.949e+07  -102538183 <2e-16 ***
57 X2:X8          -6.540e+13  8.493e+05  -77004456 <2e-16 ***
58 X3:X8          -1.079e+14  1.598e+06  -67495059 <2e-16 ***
59 X4:X8          -2.791e+14  2.104e+07  -13262555 <2e-16 ***
60 X5:X8          1.481e+15  2.284e+07   64837220 <2e-16 ***
61 X6:X8          -3.299e+14  3.935e+06  -83857759 <2e-16 ***
62 X7:X8          -3.067e+16  3.053e+08  -100470693 <2e-16 ***
63 (以下省略)
64 ---
65 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
66
67 (Dispersion parameter for binomial family taken to be 1)
68
69 Null deviance: 993.48 on 767 degrees of freedom
70 Residual deviance: 8290.04 on 512 degrees of freedom
71 AIC: 8802
72
73 Number of Fisher Scoring iterations: 15

```

3.3.2.2 從 8 個原始解釋變數中取出具統計顯著性的 4 個解釋變數相乘

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,

```

```

10 +   X6 = raw_data$BMI,
11 +   X7 = raw_data$DiabetesPedigreeFunction,
12 +   X8 = raw_data$Age,
13 +   X9 = raw_data$Glucose == 0,
14 +   X10 = raw_data$BloodPressure == 0,
15 +   X11 = raw_data$SkinThickness == 0,
16 +   X12 = raw_data$Insulin == 0,
17 +   X13 = raw_data$BMI == 0
18 +
19 >
20 > model1 <- glm(response ~ X1*X2*X6*X7, data=diabetes, family=binomial)
21 > summary(model1)
22
23 Call:
24 glm(formula = response ~ X1 * X2 * X6 * X7, family = binomial,
25      data = diabetes)
26
27 Coefficients:
28              Estimate Std. Error z value Pr(>|z|)
29 (Intercept) -1.910e+01 6.226e+00 -3.068 0.00215 ***
30 X1           3.241e+00 1.201e+00  2.699 0.00696 ***
31 X2           1.042e-01 4.782e-02  2.179 0.02930 *
32 X6           3.733e-01 1.710e-01  2.182 0.02909 *
33 X7           6.971e+00 8.781e+00  0.794 0.42723
34 X1:X2        -2.058e-02 9.390e-03 -2.192 0.02837 *
35 X1:X6        -1.032e-01 3.567e-02 -2.893 0.00381 ***
36 X2:X6        -1.891e-03 1.302e-03 -1.452 0.14638
37 X1:X7        -4.775e+00 1.951e+00 -2.448 0.01436 *
38 X2:X7        -2.986e-02 6.999e-02 -0.427 0.66964
39 X6:X7        -2.240e-01 2.333e-01 -0.960 0.33695
40 X1:X2:X6     6.811e-04 2.760e-04  2.468 0.01358 *
41 X1:X2:X7     3.107e-02 1.494e-02  2.080 0.03752 *
42 X1:X6:X7     1.834e-01 5.864e-02  3.127 0.00177 ***
43 X2:X6:X7     1.165e-03 1.816e-03  0.642 0.52108
44 X1:X2:X6:X7 -1.205e-03 4.363e-04 -2.762 0.00575 **
45 ---
46 Signif. codes:  0  ***
47
48 (Dispersion parameter for binomial family taken to be 1)
49
50 Null deviance: 993.48  on 767  degrees of freedom
51 Residual deviance: 700.19  on 752  degrees of freedom
52 AIC: 732.19
53
54 Number of Fisher Scoring iterations: 6
55
56 >

```

僅取四個顯著的單一解釋變數並考慮所有四個變數的交互作用後 AIC 大幅下降至與不考慮交互作用相差無幾，且 Residual Deviance 比只考慮顯著變數(3.3.1.2)小，顯示取顯著變數與其交互作用的擬合度與複雜度截至目前為最佳。此模型中，四個顯著單一解釋變數為懷孕次數、血糖濃度、BMI 值和糖尿病家族病史，符合3.1.3中的文獻探討結果。

3.3.2.3 加入 two-way ANOVA 中，p-value 小於 0.1 的交互作用項的結果

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 + )
19 >
20 > model1 <- glm(response ~ X1+X2+X6+X7+X1*X4+X1*X8+X2*X7+X2*X8+X4*X8+X5*X6+X5*X7,
21 +                     data=diabetes, family=binomial)
22 > summary(model1)
23
24 Call:
25   glm(formula = response ~ X1 + X2 + X6 + X7 + X1 * X4 + X1 * X8 +
26       X2 * X7 + X2 * X8 + X4 * X8 + X5 * X6 + X5 * X7, family = binomial,
27       data = diabetes)
28
29 Coefficients:
30             Estimate Std. Error z value Pr(>|z|)
31 (Intercept) -1.451e+01  1.843e+00 -7.874 3.45e-15 ***
32 X1          4.355e-01  1.204e-01  3.616 0.000299 ***
33 X2          6.867e-02  1.283e-02  5.353 8.66e-08 ***
34 X6          9.966e-02  1.855e-02  5.373 7.76e-08 ***
35 X7          3.876e+00  1.251e+00  3.097 0.001952 **

```

```

35 X4          -2.785e-02  1.938e-02  -1.437  0.150688
36 X8          1.128e-01  3.946e-02   2.858  0.004268 ** 
37 X5          5.281e-03  4.042e-03   1.307  0.191366
38 X1:X4      -3.053e-04  1.855e-03  -0.165  0.869253
39 X1:X8      -8.321e-03  2.742e-03  -3.034  0.002411 ** 
40 X2:X7      -2.185e-02  9.985e-03  -2.188  0.028666 *
41 X2:X8      -6.206e-04  2.769e-04  -2.241  0.025004 *
42 X4:X8      7.845e-04  5.589e-04   1.404  0.160436
43 X6:X5      -1.634e-04  1.089e-04  -1.500  0.133561
44 X7:X5      -9.322e-04  1.662e-03  -0.561  0.574982
45 ---
46 Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1
47
48 (Dispersion parameter for binomial family taken to be 1)
49
50 Null deviance: 993.48  on 767  degrees of freedom
51 Residual deviance: 703.32  on 753  degrees of freedom
52 AIC: 733.32
53
54 Number of Fisher Scoring iterations: 5
55
56 >

```

此模型的 Residual Deviance 較 3.3.2.2 稍大，且 AIC 稍高，顯示取 Two-way ANOVA 中 p-value 大於 0.1 的交乘項會使模型的擬合程度與複雜度均變差。

3.3.2.4 加入 two-way ANOVA 中，p-value 小於 0.05 的交互作用項

此模型的 AIC 較佳，雖然 Residual Deviance 稍高，但 AIC 比單純將 4 個重要解釋變數相乘來的低。此模型中，我們的主要變數為懷孕次數、血糖濃度、BMI、胰島素、糖尿病家族史、與年齡，交互作用項為懷孕次數與年齡、血糖濃度與家族病史、血糖與年齡、胰島素與 BMI、胰島素濃度與家族病史。

雖然所有加入的交互作用項在單獨以 Two-way ANOVA 測試時皆達 5% 顯著水準，但因為 Two-way ANOVA 是以單獨測試兩個變數，在無其他更顯著、更具解釋力的變數下兩者為顯著變數，並不代表有其他更具解釋力的變數存在時，該變數或交互作用項仍具強解釋力。

以胰島素與糖尿病家族病史為例，兩者在單獨以 Two-way ANOVA 測試時，兩者皆為顯著變數且交乘項對於判定是否罹患糖尿病具有影響力，但因為此模型有更具解釋力的變數如血糖、懷孕次數等，故胰島素在此模型反而變成不顯著，其與糖尿病使得交乘項意義亦同。

```
> raw_data <- read.csv("diabetes.csv")
```

```

2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 +
19 >
20 > model1 <- glm(response ~ X1+X2+X6+X7+X1*X8+X2*X7+X2*X8+X5*X6+X5*X7 , data=diabetes ,
21   family=binomial)
22 > summary(model1)
23
24 Call:
25 glm(formula = response ~ X1 + X2 + X6 + X7 + X1 * X8 + X2 * X7 +
26   X2 * X8 + X5 * X6 + X5 * X7, family = binomial, data = diabetes)
27
28 Coefficients:
29             Estimate Std. Error z value Pr(>|z|)
30 (Intercept) -1.494e+01  1.803e+00 -8.284 < 2e-16 ***
31 X1          4.413e-01  1.124e-01  3.927 8.60e-05 ***
32 X2          6.789e-02  1.261e-02  5.382 7.37e-08 ***
33 X6          1.000e-01  1.777e-02  5.629 1.81e-08 ***
34 X7          3.845e+00  1.260e+00  3.052  0.00227 **
35 X8          1.226e-01  3.803e-02  3.223  0.00127 **
36 X5          6.038e-03  3.950e-03  1.528  0.12639
37 X1:X8      -8.313e-03  2.716e-03 -3.061  0.00221 **
38 X2:X7      -2.149e-02  1.011e-02 -2.127  0.03346 *
39 X2:X8      -6.029e-04  2.705e-04 -2.229  0.02584 *
40 X6:X5      -1.832e-04  1.077e-04 -1.701  0.08891 .
41 X7:X5      -1.051e-03  1.675e-03 -0.627  0.53047
42 ---
43 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
44
45
46 (Dispersion parameter for binomial family taken to be 1)
47
48 Null deviance: 993.48 on 767 degrees of freedom
Residual deviance: 705.67 on 756 degrees of freedom
AIC: 729.67

```

```

49
50 Number of Fisher Scoring iterations: 5
51
52 >

```

3.3.2.5 考慮 Age(X8) 與 4 個重要解釋變數的所有交互作用項

由以上統計報表可知 Age(X8) 與其他解釋變數有重要的交互作用存在，因此我們嘗試將 4 個重要解釋變數與 Age(X8) 相乘，發現 residual deviance 與 AIC 均大幅降低，代表年齡與其他的交互作用能很好的解釋是否罹患糖尿病的資料變異、在不同的年齡下不同的血糖濃度、血壓、肥胖程度會有不同表現。

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 + )
19 >
20 > model1 <- glm(response ~ X1*X2*X6*X7*X8, data=diabetes, family=binomial)
21 警告訊息:
22 glm.fit: 擬合機率算出來是數值零或一
23 > summary(model1)

24

25 Call:
26 glm(formula = response ~ X1 * X2 * X6 * X7 * X8, family = binomial,
27      data = diabetes)

28

29 Coefficients:
30                               Estimate Std. Error z value Pr(>|z|)
31 (Intercept) -7.098e+01  2.830e+01 -2.508   0.0121 *
32 X1           1.124e+01  6.674e+00  1.683   0.0923 .
33 X2           5.036e-01  2.144e-01  2.349   0.0188 *
```

```

34 X6          1.799e+00 8.108e-01 2.219  0.0265 *
35 X7          9.333e+01 5.251e+01 1.778  0.0755 .
36 X8          1.590e+00 8.200e-01 1.939  0.0525 .
37 X1:X2      -9.444e-02 5.362e-02 -1.761  0.0782 .
38 X1:X6      -3.175e-01 2.009e-01 -1.581  0.1139
39 X2:X6      -1.379e-02 6.162e-03 -2.237  0.0253 *
40 X1:X7      -2.361e+01 1.430e+01 -1.652  0.0986 .
41 X2:X7      -7.882e-01 4.176e-01 -1.887  0.0591 .
42 X6:X7      -2.802e+00 1.577e+00 -1.777  0.0755 .
43 X1:X8      -2.262e-01 1.665e-01 -1.359  0.1743
44 X2:X8      -1.240e-02 6.059e-03 -2.047  0.0407 *
45 X6:X8      -4.227e-02 2.361e-02 -1.790  0.0734 .
46 X7:X8      -2.417e+00 1.536e+00 -1.574  0.1156
47 X1:X2:X6   2.917e-03 1.604e-03 1.819  0.0689 .
48 X1:X2:X7   2.374e-01 1.145e-01 2.073  0.0382 *
49 X1:X6:X7   7.415e-01 4.375e-01 1.695  0.0901 .
50 X2:X6:X7   2.479e-02 1.244e-02 1.992  0.0464 *
51 X1:X2:X8   2.066e-03 1.292e-03 1.599  0.1097
52 X1:X6:X8   5.849e-03 5.061e-03 1.156  0.2478
53 X2:X6:X8   3.575e-04 1.762e-04 2.029  0.0425 *
54 X1:X7:X8   4.483e-01 3.525e-01 1.272  0.2035
55 X2:X7:X8   2.132e-02 1.201e-02 1.776  0.0758 .
56 X6:X7:X8   6.960e-02 4.612e-02 1.509  0.1313
57 X1:X2:X6:X7  -7.623e-03 3.468e-03 -2.198  0.0279 *
58 X1:X2:X6:X8  -6.066e-05 3.929e-05 -1.544  0.1226
59 X1:X2:X7:X8  -4.984e-03 2.767e-03 -1.801  0.0717 .
60 X1:X6:X7:X8  -1.290e-02 1.096e-02 -1.177  0.2390
61 X2:X6:X7:X8  -6.443e-04 3.583e-04 -1.798  0.0722 .
62 X1:X2:X6:X7:X8 1.529e-04 8.526e-05 1.793  0.0729 .
63 ---
64 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
65
66 (Dispersion parameter for binomial family taken to be 1)
67
68 Null deviance: 993.48 on 767 degrees of freedom
69 Residual deviance: 649.47 on 736 degrees of freedom
70 AIC: 713.47
71
72 Number of Fisher Scoring iterations: 7
73
74 >

```

經過缺失值處理後，模型的 Residual Deviance 下降了 18.2，AIC 下降了 8.2，代表缺失值對於模型解釋具有強烈解釋意義，因為如果某些重要解釋變數的資料缺失可以由增加解釋變數代價。

3.3.2.6 缺失值處理-加入 5 個額外解釋變數

以下為將上一個模型加入 5 個額外解釋變數的結果

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 +
19 >
20 > model1 <- glm(response ~ X1*X2*X6*X7*X8+X9+X10+X11+X12+X13, data=diabetes, family=
21   binomial)
22 警告訊息:
23 glm.fit: 擬合機率算出來是數值零或一
24 > summary(model1)
25
26 Call:
27 glm(formula = response ~ X1 * X2 * X6 * X7 * X8 + X9 + X10 +
28   X11 + X12 + X13, family = binomial, data = diabetes)
29
30 Coefficients:
31                               Estimate Std. Error z value Pr(>|z|)
32 (Intercept) -7.374e+01  2.889e+01 -2.553  0.01069 *
33 X1            1.112e+01  6.897e+00  1.612  0.10702
34 X2            5.316e-01  2.193e-01  2.424  0.01533 *
35 X6            1.875e+00  8.386e-01  2.236  0.02534 *
36 X7            9.252e+01  5.251e+01  1.762  0.07809 .
37 X8            1.702e+00  8.655e-01  1.966  0.04926 *
38 X9TRUE        6.266e+00  2.005e+00  3.126  0.00177 **
39 X10TRUE       9.746e-01  5.460e-01  1.785  0.07429 .
40 X11TRUE      -6.082e-02  3.183e-01 -0.191  0.84845
41 X12TRUE       2.656e-01  2.959e-01  0.898  0.36937
42 X13TRUE       2.551e+00  1.701e+00  1.500  0.13368
43 X1:X2        -9.988e-02  5.556e-02 -1.798  0.07222 .
44 X1:X6        -3.156e-01  2.074e-01 -1.522  0.12808
45 X2:X6        -1.458e-02  6.380e-03 -2.285  0.02232 *
```

```

45 X1:X7      -2.280e+01  1.440e+01  -1.583  0.11346
46 X2:X7      -7.872e-01  4.197e-01  -1.876  0.06070 .
47 X6:X7      -2.736e+00  1.575e+00  -1.737  0.08240 .
48 X1:X8      -2.332e-01  1.721e-01  -1.355  0.17550
49 X2:X8      -1.336e-02  6.400e-03  -2.087  0.03685 *
50 X6:X8      -4.607e-02  2.533e-02  -1.819  0.06891 .
51 X7:X8      -2.446e+00  1.553e+00  -1.575  0.11532
52 X1:X2:X6   3.080e-03  1.660e-03  1.855  0.06361 .
53 X1:X2:X7   2.387e-01  1.164e-01  2.050  0.04036 *
54 X1:X6:X7   7.084e-01  4.382e-01  1.616  0.10599
55 X2:X6:X7   2.449e-02  1.250e-02  1.959  0.05009 .
56 X1:X2:X8   2.214e-03  1.342e-03  1.650  0.09892 .
57 X1:X6:X8   6.132e-03  5.244e-03  1.169  0.24233
58 X2:X6:X8   3.889e-04  1.890e-04  2.058  0.03959 *
59 X1:X7:X8   4.466e-01  3.524e-01  1.267  0.20501
60 X2:X7:X8   2.146e-02  1.214e-02  1.767  0.07723 .
61 X6:X7:X8   6.974e-02  4.662e-02  1.496  0.13464
62 X1:X2:X6:X7 -7.586e-03  3.508e-03  -2.163  0.03057 *
63 X1:X2:X6:X8 -6.536e-05  4.089e-05  -1.598  0.10998
64 X1:X2:X7:X8 -5.063e-03  2.790e-03  -1.815  0.06957 .
65 X1:X6:X7:X8 -1.271e-02  1.090e-02  -1.166  0.24367
66 X2:X6:X7:X8 -6.443e-04  3.623e-04  -1.778  0.07537 .
67 X1:X2:X6:X7:X8 1.541e-04  8.563e-05  1.800  0.07193 .
68 ---
69 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
70
71 (Dispersion parameter for binomial family taken to be 1)
72
73 Null deviance: 993.48 on 767 degrees of freedom
74 Residual deviance: 631.27 on 731 degrees of freedom
75 AIC: 705.27
76
77 Number of Fisher Scoring iterations: 7
78
79 >

```

3.3.2.7 缺失值處理-中位數填充

我們對於 5 個有缺失值的解釋變數，計算它們各自的正常值的中位數，並將結果填入缺失值中。我們發現這個做法雖然也能讓 Residual deviance 與 AIC 下降，但效果並沒有加入 5 個額外解釋變數那麼好，因此最終選擇加入 5 個額外解釋變數來處理缺失值。

```

1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(

```

```

4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 +
19 >
20 > Glucose_fill <- replace(diabetes$X2, diabetes$X2==0, median(diabetes[diabetes$X2!=0,
21     "X2"]))
21 > BloodPressure_fill <- replace(diabetes$X3, diabetes$X3==0, median(diabetes[diabetes$X3!=0,
22     "X3"]))
22 > SkinThickness_fill <- replace(diabetes$X4, diabetes$X4==0, median(diabetes[diabetes$X4!=0,
23     "X4"]))
23 > Insulin_fill <- replace(diabetes$X5, diabetes$X5==0, median(diabetes[diabetes$X5!=0,
24     "X5"]))
24 > BMI_fill <- replace(diabetes$X6, diabetes$X6==0, median(diabetes[diabetes$X6!=0, "X6"
25     ]))
25 >
26 > diabetes[, "X2"] = Glucose_fill
27 > diabetes[, "X3"] = BloodPressure_fill
28 > diabetes[, "X4"] = SkinThickness_fill
29 > diabetes[, "X5"] = Insulin_fill
30 > diabetes[, "X6"] = BMI_fill
31 >
32 > model1 <- glm(response ~ X1*X2*X6*X7*X8, data=diabetes, family=binomial)
33 警告訊息:
34 glm.fit:擬合機率算出來是數值零或一
35 > summary(model1)

36

37 Call:
38 glm(formula = response ~ X1 * X2 * X6 * X7 * X8, family = binomial,
39     data = diabetes)
40

41 Coefficients:
42             Estimate Std. Error z value Pr(>|z|)
43 (Intercept) -7.515e+01 3.040e+01 -2.472  0.0134 *
44 X1           1.433e+01 7.362e+00  1.946  0.0517 .
45 X2           5.429e-01 2.268e-01  2.394  0.0167 *
46 X6           1.965e+00 8.822e-01  2.227  0.0259 *
```

```

47 X7          9.624e+01  5.325e+01   1.807  0.0707 .
48 X8          1.761e+00  9.155e-01   1.924  0.0544 .
49 X1:X2      -1.199e-01  5.734e-02  -2.090  0.0366 *
50 X1:X6      -4.238e-01  2.223e-01  -1.906  0.0566 .
51 X2:X6      -1.519e-02  6.590e-03  -2.306  0.0211 *
52 X1:X7      -2.691e+01  1.473e+01  -1.827  0.0678 .
53 X2:X7      -8.174e-01  4.153e-01  -1.968  0.0490 *
54 X6:X7      -2.951e+00  1.600e+00  -1.844  0.0651 .
55 X1:X8      -3.144e-01  1.881e-01  -1.672  0.0946 .
56 X2:X8      -1.381e-02  6.681e-03  -2.067  0.0388 *
57 X6:X8      -4.894e-02  2.686e-02  -1.822  0.0684 .
58 X7:X8      -2.531e+00  1.556e+00  -1.626  0.1039
59 X1:X2:X6   3.749e-03  1.720e-03   2.179  0.0293 *
60 X1:X2:X7   2.640e-01  1.162e-01   2.271  0.0232 *
61 X1:X6:X7   8.595e-01  4.493e-01   1.913  0.0558 .
62 X2:X6:X7   2.597e-02  1.238e-02   2.098  0.0359 *
63 X1:X2:X8   2.713e-03  1.416e-03   1.916  0.0553 .
64 X1:X6:X8   8.881e-03  5.719e-03   1.553  0.1205
65 X2:X6:X8   4.094e-04  1.975e-04   2.073  0.0382 *
66 X1:X7:X8   5.521e-01  3.622e-01   1.524  0.1274
67 X2:X7:X8   2.218e-02  1.190e-02   1.864  0.0624 .
68 X6:X7:X8   7.506e-02  4.691e-02   1.600  0.1096
69 X1:X2:X6:X7  -8.499e-03 3.510e-03  -2.422  0.0155 *
70 X1:X2:X6:X8  -8.218e-05 4.312e-05  -1.906  0.0567 .
71 X1:X2:X7:X8  -5.709e-03 2.795e-03  -2.042  0.0411 *
72 X1:X6:X7:X8  -1.661e-02 1.120e-02  -1.483  0.1379
73 X2:X6:X7:X8  -6.821e-04 3.566e-04  -1.913  0.0558 .
74 X1:X2:X6:X7:X8 1.775e-04 8.582e-05   2.069  0.0386 *
75 ---
76 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
77
78 (Dispersion parameter for binomial family taken to be 1)
79
80 Null deviance: 993.48 on 767 degrees of freedom
81 Residual deviance: 644.28 on 736 degrees of freedom
82 AIC: 708.28
83
84 Number of Fisher Scoring iterations: 7
85
86 >

```

捕捉數據缺失的模式，可能提供更多的解釋力，特別是在數據缺失本身與結果變量有關的情況下，中位數補值雖然可以維持數據一致性，但忽略了數據缺失的信息。

3.3.2.8 最終模型

在反覆嘗試並實驗後，發現對於 5 個額外解釋變數，僅保留具統計顯著性的項目可以得到最小的 AIC，因此決定選擇其作為最終模型，即 $X1*X2*X6*X7*X8 + X9 + X10$ 。

```
1 > raw_data <- read.csv("diabetes.csv")
2 >
3 > diabetes <- data.frame(
4 +   response = raw_data$Outcome,
5 +   X1 = raw_data$Pregnancies,
6 +   X2 = raw_data$Glucose,
7 +   X3 = raw_data$BloodPressure,
8 +   X4 = raw_data$SkinThickness,
9 +   X5 = raw_data$Insulin,
10 +  X6 = raw_data$BMI,
11 +  X7 = raw_data$DiabetesPedigreeFunction,
12 +  X8 = raw_data$Age,
13 +  X9 = raw_data$Glucose == 0,
14 +  X10 = raw_data$BloodPressure == 0,
15 +  X11 = raw_data$SkinThickness == 0,
16 +  X12 = raw_data$Insulin == 0,
17 +  X13 = raw_data$BMI == 0
18 + )
19 >
20 > model1 <- glm(response ~ X1*X2*X6*X7*X8+X9+X10, data=diabetes, family=binomial)
21 警告訊息：
22 glm.fit: 擬合機率算出來是數值零或一
23 > summary(model1)
24
25 Call:
26 glm(formula = response ~ X1 * X2 * X6 * X7 * X8 + X9 + X10, family = binomial,
27      data = diabetes)
28
29 Coefficients:
30                               Estimate Std. Error z value Pr(>|z|)
31 (Intercept) -7.686e+01  2.849e+01 -2.698 0.006980 ***
32 X1            1.251e+01  6.701e+00  1.867 0.061940 .
33 X2            5.465e-01  2.148e-01  2.544 0.010960 *
34 X6            1.956e+00  8.264e-01  2.367 0.017934 *
35 X7            9.735e+01  5.169e+01  1.883 0.059677 .
36 X8            1.776e+00  8.498e-01  2.089 0.036663 *
37 X9TRUE        6.600e+00  1.926e+00  3.426 0.000612 ***
38 X10TRUE       1.157e+00  5.162e-01  2.241 0.025053 *
39 X1:X2         -1.073e-01  5.390e-02 -1.991 0.046501 *
40 X1:X6         -3.582e-01  2.024e-01 -1.770 0.076716 .
41 X2:X6         -1.495e-02  6.241e-03 -2.395 0.016599 *
42 X1:X7         -2.433e+01  1.421e+01 -1.712 0.086830 .
43 X2:X7         -8.089e-01  4.083e-01 -1.981 0.047564 *
44 X6:X7         -2.869e+00  1.550e+00 -1.850 0.064268 .
```

```

45 X1:X8      -2.544e-01  1.687e-01  -1.508  0.131472
46 X2:X8      -1.366e-02  6.246e-03  -2.187  0.028730 *
47 X6:X8      -4.761e-02  2.483e-02  -1.918  0.055118 .
48 X7:X8      -2.547e+00  1.521e+00  -1.674  0.094042 .
49 X1:X2:X6   3.315e-03  1.616e-03  2.051  0.040307 *
50 X1:X2:X7   2.458e-01  1.137e-01  2.162  0.030651 *
51 X1:X6:X7   7.572e-01  4.338e-01  1.745  0.080899 .
52 X2:X6:X7   2.506e-02  1.213e-02  2.066  0.038848 *
53 X1:X2:X8   2.326e-03  1.307e-03  1.780  0.075143 .
54 X1:X6:X8   6.748e-03  5.157e-03  1.309  0.190678
55 X2:X6:X8   3.942e-04  1.839e-04  2.144  0.032069 *
56 X1:X7:X8   4.693e-01  3.484e-01  1.347  0.177950
57 X2:X7:X8   2.179e-02  1.175e-02  1.854  0.063761 .
58 X6:X7:X8   7.206e-02  4.554e-02  1.582  0.113589
59 X1:X2:X6:X7 -7.822e-03  3.432e-03 -2.279  0.022680 *
60 X1:X2:X6:X8 -6.873e-05  3.993e-05 -1.721  0.085200 .
61 X1:X2:X7:X8 -5.151e-03  2.724e-03 -1.891  0.058608 .
62 X1:X6:X7:X8 -1.339e-02  1.080e-02 -1.240  0.215135
63 X2:X6:X7:X8 -6.499e-04  3.492e-04 -1.861  0.062727 .
64 X1:X2:X6:X7:X8  1.568e-04  8.364e-05  1.875  0.060844 .
65 ---
66 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
67
68 (Dispersion parameter for binomial family taken to be 1)
69
70 Null deviance: 993.48  on 767  degrees of freedom
71 Residual deviance: 634.65  on 734  degrees of freedom
72 AIC: 702.65
73
74 Number of Fisher Scoring iterations: 7
75
76 >

```

原資料集中重要的解釋變數為懷孕次數、血糖、BMI、家族病史與年齡，此結果與 3.1.3 吻合，但此模型因為考慮交互作用使 Residual Deviance 由 715.70 降至 634.65。顯著影響的交互作用為此五個變數的各項交互作用、與血糖、血壓是否缺失的指標變數。

血糖濃度與懷孕次數的交互作用效應(圖12a)在中低血糖的人較不明顯，唯相同血糖範圍的個體罹患糖尿病的機率隨懷孕次數增長，對於高血糖範圍者，其罹患糖尿病的機率斜率平緩但平均值高，代表在高血糖的情況下，懷孕次數對是否罹患糖尿病的影響力較小，但每個懷孕次數的範圍中，血糖高低對罹患糖尿病的機率有絕對性的影響。

血糖與身體質量指數的交互作用如圖12b所示，在高 BMI 群體中，血糖範圍由中到高的斜率較陡，代表罹患糖尿病的機率在肥胖族群中對血糖濃度極為敏感，而對於 BMI

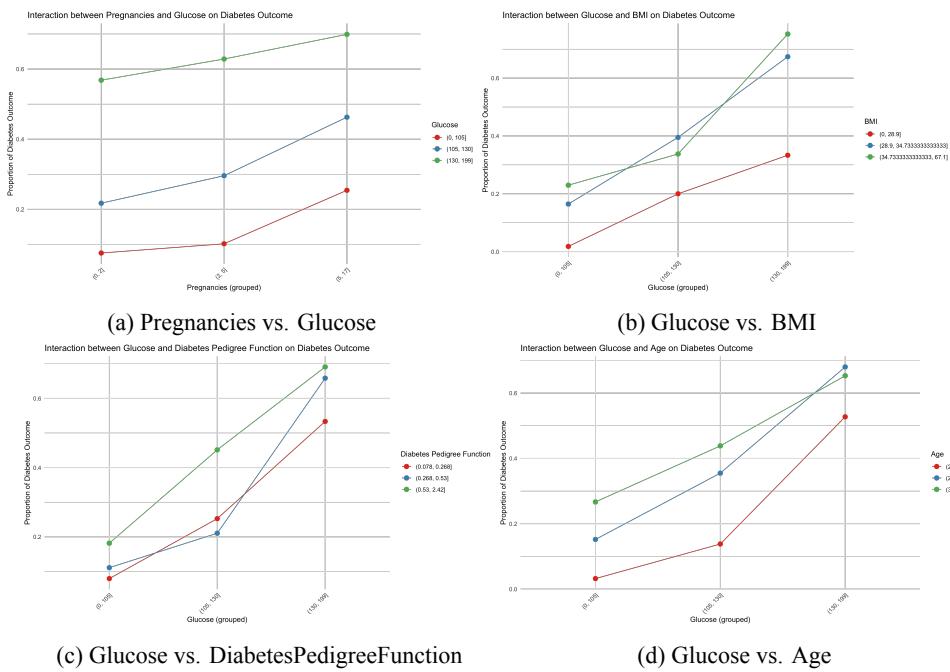


Figure 12: 顯著的交互作用影響 Outcome 視覺畫結果

較低的群體，血糖濃度對於罹患糖尿病的機率較不敏感，代表控制體重有助於減少血糖對罹患糖尿病機率的敏感度。

血糖與糖尿病家族史的交互作用如12c所示，在所有血糖範圍中，擁有較多糖尿病家族史比例的個體罹患糖尿病的比例均最高，而在正常飯後血糖(<140)範圍的個體，較低血糖者罹患糖尿病的機率與家族史分數順序一致，但家族史所計算出中等風險者得到糖尿病的機率反而比 DiabetesPedigreeFunction 最小者低，可能解釋為血糖偏高者且有家族史的個體因為注意到血糖接近臨界值而開始控制飲食進而避免罹患糖尿病。

血糖與年齡的交互作用如圖12d所示，青年組罹患糖尿病的機率均最低，但是青年的血糖由中區段跨到高範圍時風險急劇上升。中年群體在血糖範圍為中間值時風險偏高但並非最高，但在高血糖群體中年人的罹患率居首。對於高齡人口，血糖濃度增加帶來的風險高，須謹慎預防、及早控制。

3.4 模型精確度

使用全部資料擬合的模型準確度為 0.796875

```

1 > raw_data <- read.csv("Diabetes.csv")
2 > diabetes <- data.frame(
3 +   response = raw_data$Outcome,
4 +   X1 = raw_data$Pregnancies,
5 +   X2 = raw_data$Glucose,
6 +   X3 = raw_data$BloodPressure,
```

```

7 +   X4 = raw_data$SkinThickness,
8 +   X5 = raw_data$Insulin,
9 +   X6 = raw_data$BMI,
10 +  X7 = raw_data$DiabetesPedigreeFunction,
11 +  X8 = raw_data$Age,
12 +  X9 = raw_data$Glucose == 0,
13 +  X10 = raw_data$BloodPressure == 0,
14 +  X11 = raw_data$SkinThickness == 0,
15 +  X12 = raw_data$Insulin == 0,
16 +  X13 = raw_data$BMI == 0
17 +
18 > model1 <- glm(response ~ X1*X2*X6*X7*X8+X9+X10, data=diabetes, family=binomial)
19 Warning message:
20 glm.fit: fitted probabilities numerically 0 or 1 occurred
21 >
22 > predicted_probs <- predict(model1, diabetes, type = "response")
23 >
24 > threshold <- 0.5
25 > predicted_classes <- ifelse(predicted_probs > threshold, 1, 0)
26 >
27 > accuracy <- mean(predicted_classes == diabetes$response)
28 > print(paste("Model Accuracy: ", accuracy))
29 [1] "Model Accuracy: 0.796875"

```

使用訓練資料集擬合，使用測試資料集測試模型準確度得到精準度為 0.727272，僅略低與使用全部資料擬合，代表模型的泛化能力佳。另外，得到的 F1-score 和 AUROC 分別為 0.7907 和 0.8011。

```

1 > library(caret)
2 > library(pROC)
3 >
4 > raw_data <- read.csv("train_data.csv")
5 >
6 > diabetes <- data.frame(
7 +   response = raw_data$Outcome,
8 +   X1 = raw_data$Pregnancies,
9 +   X2 = raw_data$Glucose,
10 +  X3 = raw_data$BloodPressure,
11 +  X4 = raw_data$SkinThickness,
12 +  X5 = raw_data$Insulin,
13 +  X6 = raw_data$BMI,
14 +  X7 = raw_data$DiabetesPedigreeFunction,
15 +  X8 = raw_data$Age,
16 +  X9 = raw_data$Glucose == 0,
17 +  X10 = raw_data$BloodPressure == 0,
18 +  X11 = raw_data$SkinThickness == 0,

```

```

19 +     X12 = raw_data$Insulin == 0,
20 +     X13 = raw_data$BMI == 0
21 +
22 >
23 > model <- glm(response ~ X1*X2*X6*X7*X8+X9+X10, data=diabetes, family=binomial)
24 警告訊息:
25 glm.fit: 擬合機率算出來是數值零或一
26 > summary(model)
27
28 Call:
29 glm(formula = response ~ X1 * X2 * X6 * X7 * X8 + X9 + X10, family = binomial,
30      data = diabetes)
31
32 Coefficients:
33                               Estimate Std. Error z value Pr(>|z|)
34 (Intercept)           -8.835e+01  4.190e+01 -2.108   0.0350 *
35 X1                    1.261e+01  1.029e+01  1.225   0.2205
36 X2                   6.857e-01  3.279e-01  2.091   0.0365 *
37 X6                   2.222e+00  1.202e+00  1.848   0.0646 .
38 X7                   1.621e+02  8.828e+01  1.836   0.0663 .
39 X8                   2.127e+00  1.246e+00  1.708   0.0877 .
40 X9TRUE                6.675e+00  2.956e+00  2.258   0.0239 *
41 X10TRUE               1.059e+00  7.567e-01  1.399   0.1618
42 X1:X2                -1.322e-01  8.270e-02 -1.598   0.1100
43 X1:X6                -3.187e-01  3.061e-01 -1.041   0.2978
44 X2:X6                -1.856e-02  9.396e-03 -1.975   0.0483 *
45 X1:X7                -3.562e+01  2.310e+01 -1.542   0.1230
46 X2:X7                -1.425e+00  7.075e-01 -2.014   0.0440 *
47 X6:X7                -4.831e+00  2.655e+00 -1.820   0.0688 .
48 X1:X8                -2.349e-01  2.596e-01 -0.905   0.3654
49 X2:X8                -1.788e-02  9.807e-03 -1.823   0.0682 .
50 X6:X8                -5.722e-02  3.568e-02 -1.604   0.1088
51 X7:X8                -4.550e+00  2.733e+00 -1.665   0.0959 .
52 X1:X2:X6               3.735e-03  2.451e-03  1.524   0.1275
53 X1:X2:X7               3.854e-01  1.841e-01  2.094   0.0363 *
54 X1:X6:X7               1.042e+00  6.918e-01  1.507   0.1319
55 X2:X6:X7               4.338e-02  2.100e-02  2.066   0.0388 *
56 X1:X2:X8               2.765e-03  2.058e-03  1.344   0.1791
57 X1:X6:X8               5.429e-03  7.792e-03  0.697   0.4859
58 X2:X6:X8               5.167e-04  2.817e-04  1.834   0.0666 .
59 X1:X7:X8               6.679e-01  5.956e-01  1.121   0.2622
60 X2:X7:X8               4.078e-02  2.198e-02  1.855   0.0635 .
61 X6:X7:X8               1.335e-01  8.131e-02  1.641   0.1008
62 X1:X2:X6:X7            -1.156e-02  5.482e-03 -2.109   0.0349 *
63 X1:X2:X6:X8            -7.641e-05  6.177e-05 -1.237   0.2161
64 X1:X2:X7:X8            -8.086e-03  4.664e-03 -1.734   0.0830 .
65 X1:X6:X7:X8            -1.851e-02  1.804e-02 -1.026   0.3048
66 X2:X6:X7:X8            -1.218e-03  6.436e-04 -1.893   0.0584 .

```

```

67 X1:X2:X6:X7:X8  2.370e-04  1.405e-04   1.687   0.0917 .
68 ---
69 Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
70
71 (Dispersion parameter for binomial family taken to be 1)
72
73 Null deviance: 695.42  on 536  degrees of freedom
74 Residual deviance: 410.35  on 503  degrees of freedom
75 AIC: 478.35
76
77 Number of Fisher Scoring iterations: 7
78
79 >
80 >
81 > test_data <- read.csv("test_data.csv")
82 >
83 > diabetes_test <- data.frame(
84 +   response = test_data$Outcome,
85 +   X1 = test_data$Pregnancies,
86 +   X2 = test_data$Glucose,
87 +   X3 = test_data$BloodPressure,
88 +   X4 = test_data$SkinThickness,
89 +   X5 = test_data$Insulin,
90 +   X6 = test_data$BMI,
91 +   X7 = test_data$DiabetesPedigreeFunction,
92 +   X8 = test_data$Age,
93 +   X9 = test_data$Glucose == 0,
94 +   X10 = test_data$BloodPressure == 0,
95 +   X11 = test_data$SkinThickness == 0,
96 +   X12 = test_data$Insulin == 0,
97 +   X13 = test_data$BMI == 0
98 + )
99 >
100 > # 預測機率
101 > threshold <- 0.5
102 > predicted_probs <- predict(model, newdata=diabetes_test, type="response")
103 > predicted_classes <- ifelse(predicted_probs > threshold, 1, 0)
104 >
105 > # 計算準確率
106 > accuracy <- mean(predicted_classes == diabetes_test$response)
107 > print(paste("Model Accuracy: ", accuracy))
108 [1] "Model Accuracy:  0.727272727272727"
109 >
110 > # 計算F1-score
111 > conf_matrix <- confusionMatrix(factor(predicted_classes), factor(diabetes_test$`response`))
112 > f1_score <- conf_matrix$byClass["F1"]
113 > print(paste("F1 Score: ", f1_score))

```

```
114 [1] "F1 Score:  0.790697674418605"
115 >
116 > # 計算AUROC
117 > roc_curve <- roc(diabetes_test$response, predicted_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
118
119
120 > auroc <- auc(roc_curve)
121 > print(paste("AUROC: ", auroc))
122 [1] "AUROC:  0.801076158940397"
```

4 機器學習模型比較

4.1 方法介紹

這裡比較了五種不同的機器學習模型預測糖尿病的結果，模型包含：邏輯回歸（Logistic Regression）、決策樹（Decision Tree）、隨機森林（Random Forest）、支持向量機（Support Vector Machine, SVM）和 XGBoost（XGBoost Classifier）。每種方法都有不同的優勢和適用情境，因此通過比較這些方法的性能，可以更全面地了解模型在不同場景下的表現。

4.2 模型訓練與評估

為了評估模型的性能，我們使用了多種評估指標，包括準確率（Accuracy）、精確率（Precision）、召回率（Recall）、F1 分數（F1 Score）以及受試者工作特徵曲線下面積（Area Under the ROC Curve, AUC）。這些指標能夠全面反映模型的分類能力。

Models	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7359	0.74	0.74	0.74	0.7978
Decision Tree	0.7056	0.73	0.71	0.71	0.7043
Random Forest	0.7576	0.76	0.76	0.76	0.8046
SVM	0.7446	0.74	0.74	0.74	0.7974
XGBoost	0.7273	0.73	0.73	0.73	0.7833

Table 3: 使用 Raw Data 預測結果

4.3 模型準確率和分類報告

4.3.1 使用 Raw Data 預測結果

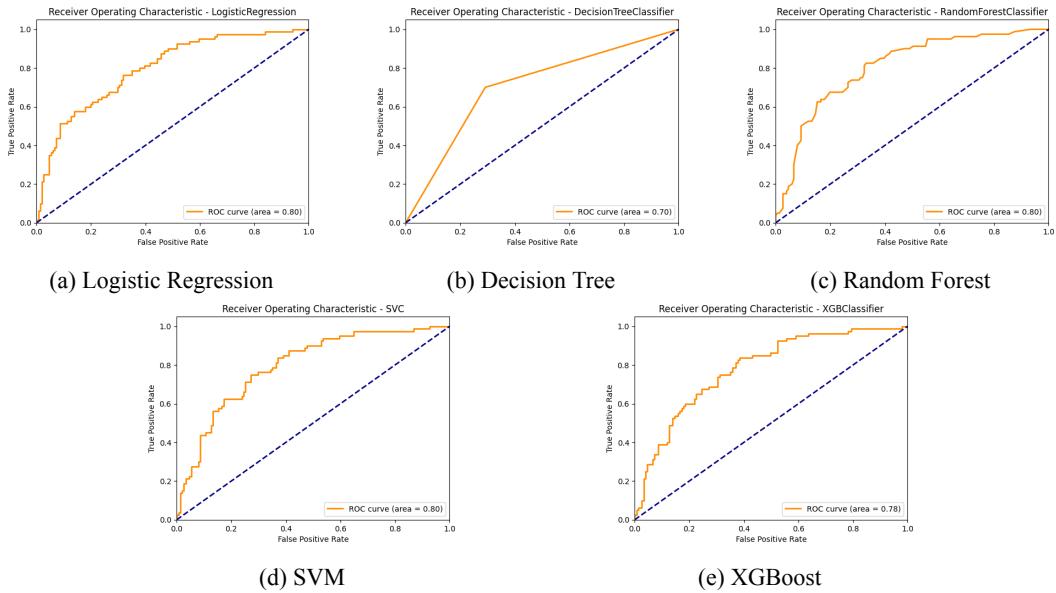


Figure 13: 各模型的 ROC 曲線（使用 Raw Data）

從表 3 可以看出，Random Forest 在所有評估指標上都表現最佳，特別是在 AUC 指標上達到 0.8046。Logistic Regression 和 SVM 的表現相似，準確率均約為 0.74，AUC 接近 0.797，表現穩定。相較之下，Decision Tree 在所有指標上表現最差，特別是在 AUC 上僅為 0.7043。然而，從整體上來看，不同模型在此資料集的表現上並沒有顯著差異。

Models	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7712	0.77	0.77	0.77	0.8035
Decision Tree	0.7119	0.72	0.71	0.71	0.6907
Random Forest	0.7542	0.75	0.75	0.75	0.7772
SVM	0.7542	0.75	0.75	0.75	0.7968
XGBoost	0.7542	0.75	0.75	0.75	0.7824

Table 4: 使用 Filtered Data 預測結果

4.3.2 使用 Filtered Data 預測結果

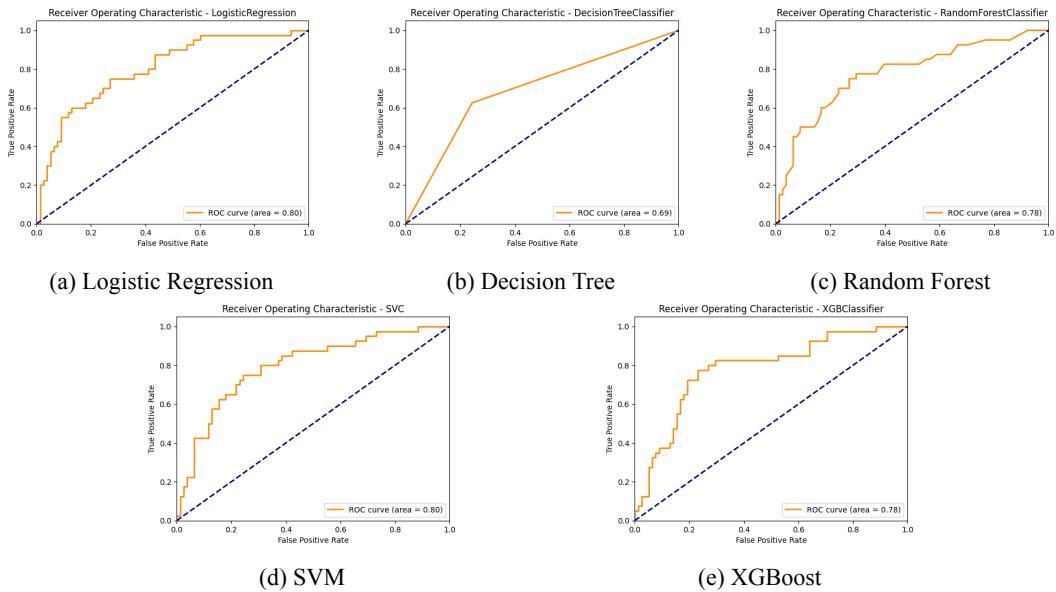


Figure 14: 各模型的 ROC 曲線（使用 Filtered Data）

在表 4 中可以看到，除了 Random Forest 和 Decision Tree 外，其他模型在所有指標上都有些微提升，其中 Logistic Regression 在所有評估指標上均表現最佳，特別是在準確率 (0.7712) 和 F1-Score (0.77) 上。這表明在過濾數據後，Logistic Regression 能夠更好地捕捉數據特徵。而 Random Forest 的表現則在所有指標上略有下降，這可能是由於資料量減少，使得較為複雜的模型表現下降。

Models	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7403	0.74	0.74	0.74	0.7976
Decision Tree	0.6926	0.70	0.69	0.69	0.6679
Random Forest	0.7576	0.76	0.76	0.76	0.7975
SVM	0.7446	0.74	0.74	0.74	0.7935
XGBoost	0.7359	0.74	0.74	0.74	0.7792

Table 5: 使用 Imputed Data 預測結果

4.3.3 使用 Imputed Data 預測結果

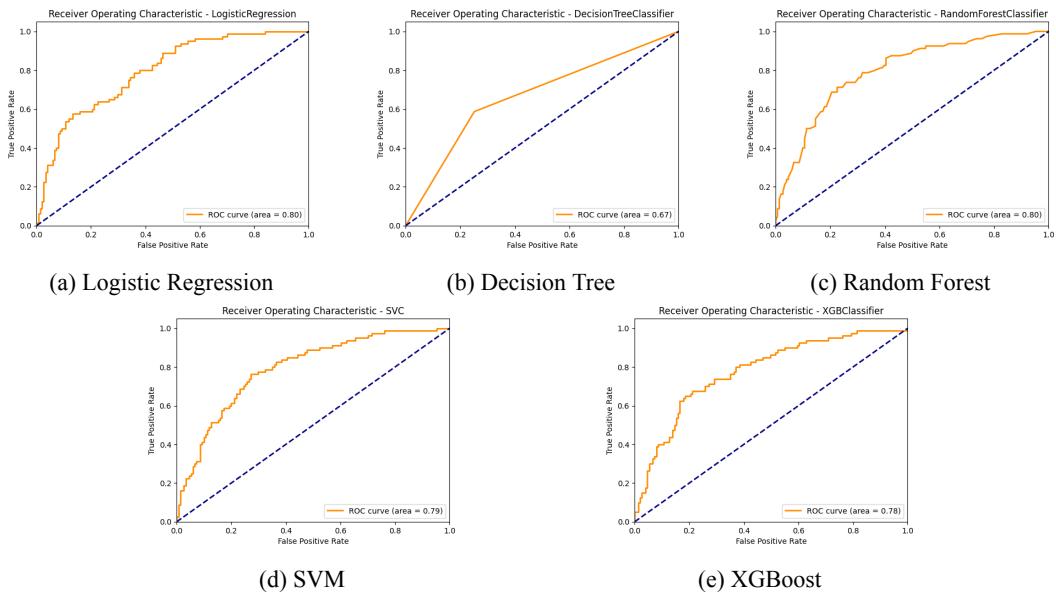


Figure 15: 各模型的 ROC 曲線（使用 Imputed Data）

表 5 顯示使用中位數填充後的資料訓練的模型預測結果與原始數據的結果相差不大。Random Forest 依然表現最佳，Logistic Regression 和 SVM 表現相似。XGBoost 在除 AUC 外的指標上略有提升。值得注意的是，Decision Tree 的指標反而略有下降，AUC 更是從 0.7043 跌至 0.6679。

4.4 總結

綜合以上結果可以看出，不同數據處理方法對於模型性能的影響並不顯著，模型的準確率基本上都在 69% 到 75% 之間。這表明在此資料集上，各種機器學習模型的表現差距不大。因此，選擇較為簡單且解釋性強的模型，如邏輯回歸，可能更有利於實際應用。邏輯回歸在數據過濾後性能最佳，適合於數據較為乾淨的情境，而隨機森林則在處

理複雜數據時具有優勢。支持向量機（SVM）和 XGBoost 在多數情況下表現良好，顯示其在不同數據處理策略下的適應性。決策樹（Decision Tree）的表現相對較弱，特別是在數據未經過濾或填補的情況下，泛化能力較差。

總體而言，考慮到性能、解釋性和實際應用的需求，邏輯回歸是一個較為理想的選擇，尤其是在處理此類中小型數據集時。

4.5 與自定義統計模型3.3.2.8比較

在直接使用 Raw data 的情況下，我們經過篩選變數所得到的模型準確率為 0.73，略低於使用機器學習方法 Random Forest 與 SVM 的準確度，但 AUC 與 Random Forest 相近且 F1 Score 高於 Random Forest，代表模型分類能力已經可以與 Random Forest 媲美，可推斷自定義的模型已準確捕捉重要解釋變數與交互作用，更重要的是，相較於機器學習模型，統計模型具有較高可解釋性，在分析致病因子的具體影響方面更具優勢，因為它們提供明確的係數和顯著性水平，更能理解每個變量及其交互作用對結果的影響。

5 結論

本研究中，我們根據解釋變數「Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age」以及應變數「Outcome」來分析並建立統計模型。分析的部分包含「單因素方差分析(One-way ANOVA), 雙因素方差分析(Two-way ANOVA), 繪製折線圖觀察關聯性」，而統計模型的建立是以 AIC (Akaike Information Criterion) 作為模型好壞、解釋力與複雜度平衡的指標，並與機器學習方法的模型之準確率做比較。

資料分析過程中，我們嘗試不同缺失值的處理方式：刪除缺失資料，中位數填充與增加是否為缺失值的額外解釋變數。並發現缺失值的處理方式對模型的準確性有顯著影響。新增據影響力的解釋變數是否為缺失的額外解釋變數有助於提高模型的解釋力。

由統計方法與分析得出血糖、懷孕次數、BMI、糖尿病家族史與年齡具有較大影響力，而年齡較大、懷孕較多次、具有較多家族史與較高血糖者均有較高罹患糖尿病的風險。我們將由統計分析得到的模型與機器學習方法比較，得到相似準確度，但我們的模型具有高可解釋性的優點，在探討罹患糖尿病因素上可以有更清楚的分析，為預防與提早治療提供科學依據。

此資料集僅針對印度裔 21-81 歲女性的身體各項指標，並未針對較多樣化的人口採樣，優點為樣本同質性高，同種族與性別可降低潛藏變數如基因、性別習慣等帶來的影響。但缺點為此研究僅能適用於特定族群或提供大致方向，但泛化能力或許並非如此理想。未來研究可針對較廣人口採樣，以驗證不同民族、生活方式、甚至氣候等各種因素

對於糖尿病的影響深入探討。

6 心得

吳秉澍:

建立一個完整且有效的統計模型需要考慮多個前提條件。像是變數的選擇應盡量精簡，以提高模型的準確性和解釋力；同時，資料量及資料品質也是分析的重要因素，資料量太少或未載明缺失原因的缺失值也可能會導致結果的不準確和偏頗。即便檢定結果顯示某項變因或交互作用具有統計顯著性，有時仍需要結合專業知識來進一步確認，因為可能會有一些在選用資料外的變因對結果造成影響。

溫柏萱:

期末專題讓我整合了每次作業所進行的分析、處理，並應用了上課所學到的觀念，使我對如何分析、預測資料有更系統與結構的認識。雖然我們直接採用 Logistic Regression，但對於 Residual 的判斷、模型是否過度擬合、如何解決過度擬合與資料缺失讓我對分析資料的方法有更深刻的理解。謝謝同組隊友高效率的合作與互相包容使我們都能發揮所長，也感謝老師即時回覆，悉心指點我們分析、判斷的方向與用心的教學！

黃昱翰:

這次的專題讓我深刻體驗到了統計在實際應用上不是簡單做個檢定、套個模型就結束了。要考慮的事情意外地多，除了數據處理和模型選擇外，還有很多細節需要注意，例如資料的質量、變數之間的關係、模型的解釋力等。即使現在的運算能力很強、複雜的深度學習模型可以做到精準預測，但未來的數據資料只會越來越多，因此，我們仍應遵循精簡的原則，用盡可能小的成本達到令人滿意的表現。感謝老師用心的指教和同學的合作！

References

- [1] International Diabetes Federation, “Idf diabetes atlas, 10th edition,” 2021, accessed: 2024-06-17. [Online]. Available: <https://diabetesatlas.org>
- [2] Health Promotion Administration, Ministry of Health and Welfare, Taiwan, “National health insurance annual report,” 2024, accessed: 2024-06-17. [Online]. Available: <https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=642&pid=1235>
- [3] Mayo Clinic, “Diabetes,” 2020. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444>
- [4] National Institute of Diabetes and Digestive and Kidney Diseases, “Diabetes overview,” 2024, accessed: 2024-06-17. [Online]. Available: <https://www.niddk.nih.gov/health-information/diabetes/overview>
- [5] C. Kim, K. M. Newton, and R. H. Knopp, “Gestational diabetes and the incidence of type 2 diabetes: a systematic review,” *Diabetes care*, vol. 25, no. 10, pp. 1862–1868, 2002.
- [6] J. Bigby, “Harrison’s principles of internal medicine,” *Archives of Dermatology*, vol. 124, no. 2, pp. 287–287, 1988.
- [7] D. P. Guh, W. Zhang, N. Bansback, Z. Amarsi, C. L. Birmingham, and A. H. Anis, “The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis,” *BMC public health*, vol. 9, pp. 1–20, 2009.
- [8] V. Lyssenko and M. Laakso, “Genetic screening for the risk of type 2 diabetes: worthless or valuable?” *Diabetes care*, vol. 36, no. Suppl 2, p. S120, 2013.
- [9] C. C. Cowie, K. F. Rust, E. S. Ford, M. S. Eberhardt, D. D. Byrd-Holt, C. Li, D. E. Williams, E. W. Gregg, K. E. Bainbridge, S. H. Saydah *et al.*, “Full accounting of diabetes and pre-diabetes in the us population in 1988–1994 and 2005–2006,” *Diabetes care*, vol. 32, no. 2, pp. 287–294, 2009.