資料探勘導論

期末專題

Introduction to Data Mining

Final Project

報告主題：糖尿病預測

系級：資工碩一

學生姓名：陳菀菁、李祐瑄

學號：11177008、11177011

中　華　民　國　一　百　一　十　二　年　一　月

目錄

第一章 數據探索

因糖尿病已日漸成為現代人的文明病，因此想透過資料探勘了解，何種身理特徵容易罹患糖尿病。本實驗將會依據病人身體的狀況以及各項指標等去做預測分析，來判斷病人是否有糖尿病。

1-1 數據解釋

本實驗之數據集使用來自 Kaggle 的糖尿病資料集[1]，數據筆數共有 520 筆，欄位共有 15 欄，下表為各欄位的說明：

表 1. 變數說明

變數名稱	對應中文解釋	值域
Age	年齡	16~90
Gender	性別	Male/Female
Polyuria	多尿(頻尿)	Yes/No
Polydipsia	容易口渴	Yes/No
Sudden weight loss	體重驟降	Yes/No
Weakness	虛弱	Yes/No
Polyphagia	多食症	Yes/No
Genital thrush	念珠菌感染	Yes/No
Visual blurring	視線模糊	Yes/No
Itching	搔癢	Yes/No
Irritability	應激性	Yes/No
Delayed healing	延遲癒合(傷口不易癒合)	Yes/No
Partial paresis	局部麻痺	Yes/No
Muscle stiffness	肌肉僵硬	Yes/No
Alopecia	脫髮	Yes/No
Obesity	肥胖	Yes/No
Class (Y 預測)	是否有糖尿病 (Y 預測)	Positive/Negative

1-2 數據敘述統計

根據視覺化處理，觀察各變數與是否患有糖尿病的欄位的關係。

1. 年齡與糖尿病：

根據圖 1，我們將年齡分成四個階段，分別是青少年(15~21)、青年(22~39)、中年(40~69)、老年(70~90)，由於青少年與老年的資料本身較少，但由青年與中年來看，還是可以看出隨著年齡的增長，有更高的比例罹患糖尿病。

圖 1. 年齡與糖尿病之統計圖

2. 性別與糖尿病：

根據圖 2 可以發現到，相對於女性較高比率患有糖尿病，男性只有不到 50%的比率患有糖尿病。

圖 2. 性別與糖尿病之統計圖

3. 其他特徵與糖尿病：

根據圖 3 可以發現，有 polyuria(頻尿)、polydipsia(口渴)、sudden weight loss(體重驟降)、weakness(虛弱)、polyphagia(多食症)、Genital thrush(念珠菌感染)、visual blurring(視線模糊)、irritability(應激性)、partial paresis(局部麻痺)、Obesity(肥胖)等以上情況者，患有糖尿病的比率明顯較高。

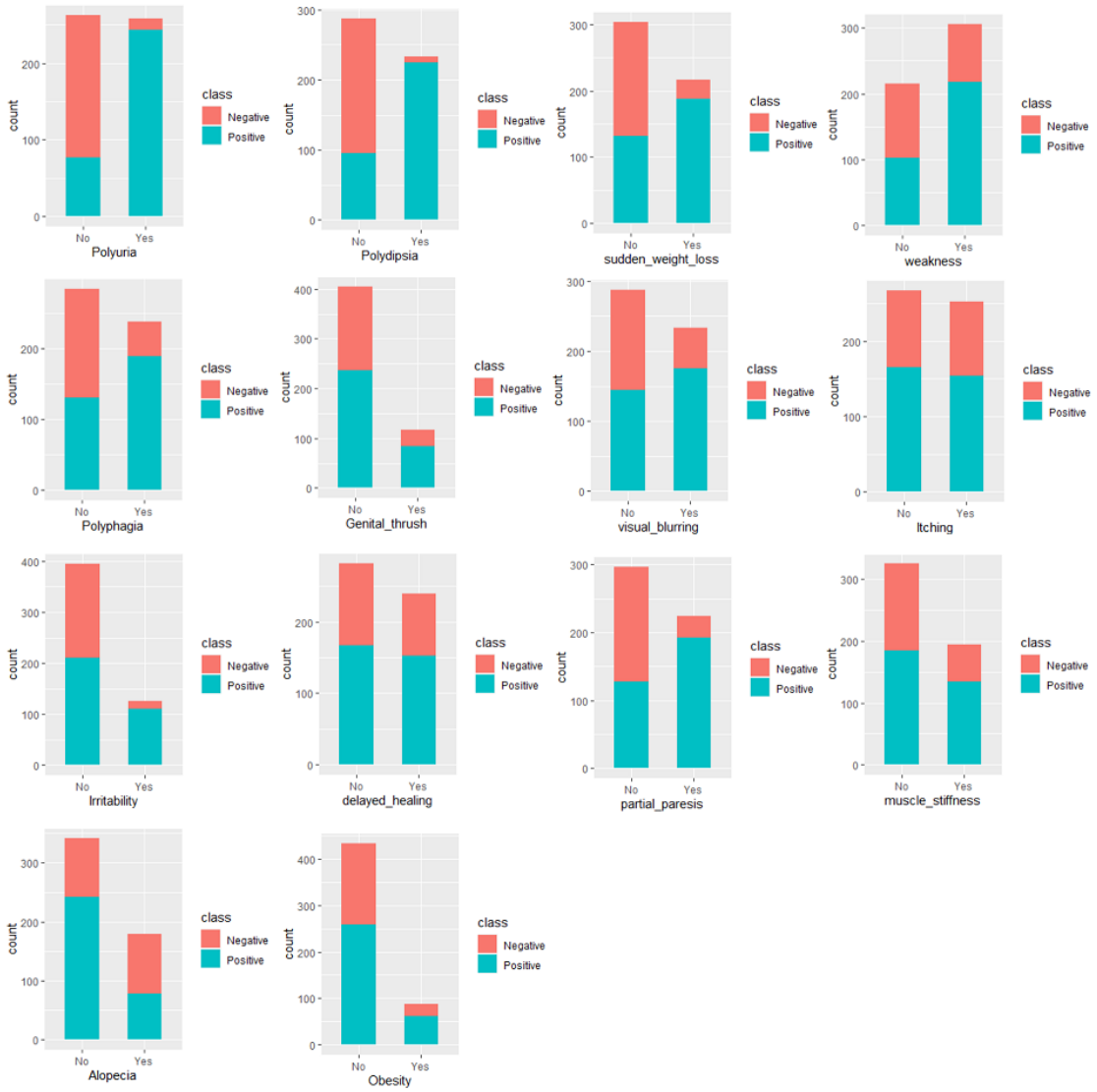


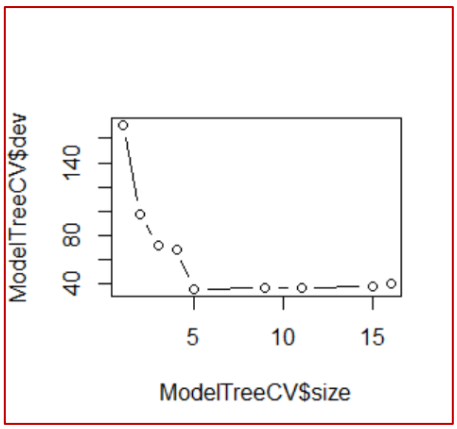
圖 3. 其他特徵與糖尿病之統計圖

第二章 建立模型

本實驗使用 R 語言[2]去建立 Classification tree、Bagging 和 Random Forest[3]模型。

2-1 Classification tree

在分類樹的建立，我們使用 10-Fold CV 去選擇 tree size，可從圖 4 中觀察到當 tree size = 5 時，有最低的 error 值，其 error 值為 36。




```
> ModelTreeCV$size;ModelTreeCV$dev
[1] 16 15 11  9  5  4  3  2  1
[1]  41  38  37  37  36  69  72  98 171
```

圖 4. 透過 10-Fold CV 產生的樹深與 error

再將剛剛透過交叉驗證所選擇的 tree size 繪出(圖 5)，可從分類樹中看到此分類樹透過 polyuria(頻尿)、polydipsia(口渴)、gender(性別)及 alopecia(脫髮)這四個特徵去做分類樹的分類。

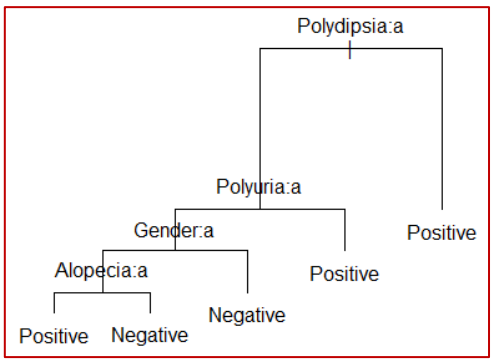


圖 5. 分類樹(tree size = 5)

最後產生分類樹的相關分析結果，根據圖 6 可看出準確度為 0.8462，而 AUC 曲線則為 0.856 (圖 7)。

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       26        2
  Positive        6       18

               Accuracy : 0.8462
                 95% CI : (0.7192, 0.9312)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 0.0002607

                  Kappa : 0.6867

 Mcnemar's Test P-Value : 0.2888444

            Sensitivity : 0.9000
            Specificity : 0.8125
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.9286
             Prevalence : 0.3846
         Detection Rate : 0.3462
   Detection Prevalence : 0.4615
      Balanced Accuracy : 0.8562

       'Positive' Class : Positive
```

圖 6. 分類樹分析結果

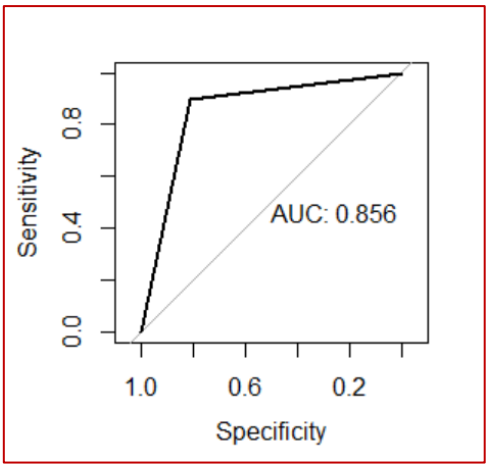


圖 7. 分類樹 AUC 曲線

2-2 Bagging & Random Forest

在 Bagging 模型的建立，從圖 8 中可觀察出各變數與預測是否患有糖尿病的重要程度，Mean Decrease Accuracy 為刪除此變數，會使得 accuracy 下降多少數值愈大代表此變數愈重。Mean Decrease Gini 為刪除此變數，會使得 Gini index 下降多少數值愈大代表此變數愈重要。而前三名重要程度依序為 polyuria(頻尿)、polydipsia(口渴)及 gender(性別)。

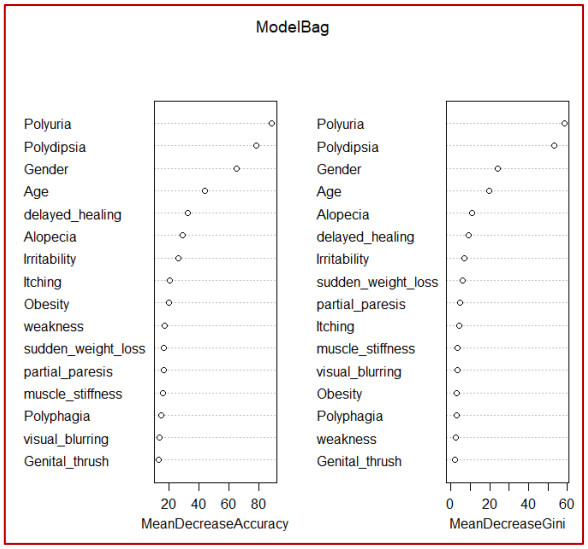


圖 8. 各變數 X 對於預測 Y(是否有病)的重要程度

最後產生的分析結果，根據圖 9 可看出準確度為 0.9424，比分類樹提升了 10%左右，而 AUC 曲線則為 0.934 (圖 10)。

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       31        2
  Positive        1       18

               Accuracy : 0.9423
                 95% CI : (0.8405, 0.9879)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 6.455e-08

                  Kappa : 0.877

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9000
            Specificity : 0.9688
         Pos Pred Value : 0.9474
         Neg Pred Value : 0.9394
             Prevalence : 0.3846
         Detection Rate : 0.3462
   Detection Prevalence : 0.3654
      Balanced Accuracy : 0.9344

       'Positive' Class : Positive
```

圖 9. Bagging 分析結果

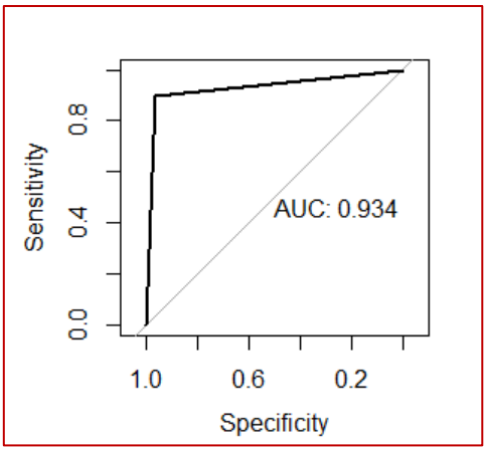


圖 10. Bagging AUC 曲線

接著我們使用 Random Forest 去做改進，雖然各變數對於是否有病的重要程度(圖 11)沒有變化，但在準確率方面達到了 0.9615(圖 12)，相較於 Bagging 提升了 2%左右。

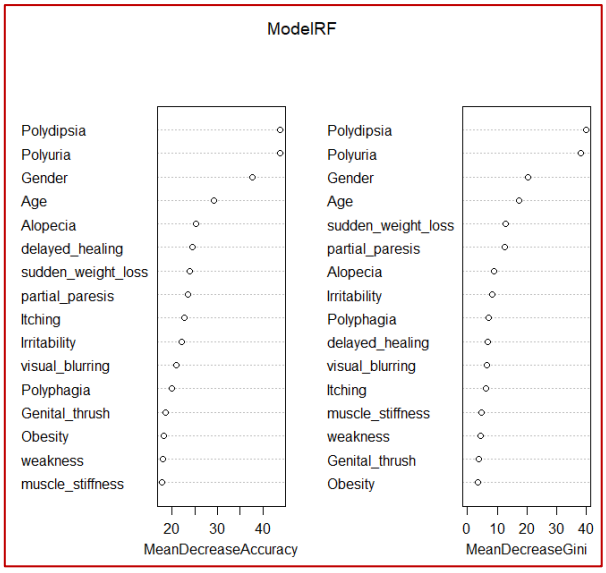


圖 11. 各變數 X 對於預測 Y(是否有病)的重要程度

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       31        1
  Positive        1       19

               Accuracy : 0.9615
                 95% CI : (0.8679, 0.9953)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 5.986e-09

                  Kappa : 0.9188

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9500
            Specificity : 0.9688
         Pos Pred Value : 0.9500
         Neg Pred Value : 0.9687
             Prevalence : 0.3846
         Detection Rate : 0.3654
   Detection Prevalence : 0.3846
      Balanced Accuracy : 0.9594

       'Positive' Class : Positive
```

圖 12. 透過 Random Forest 改進後的分析結果

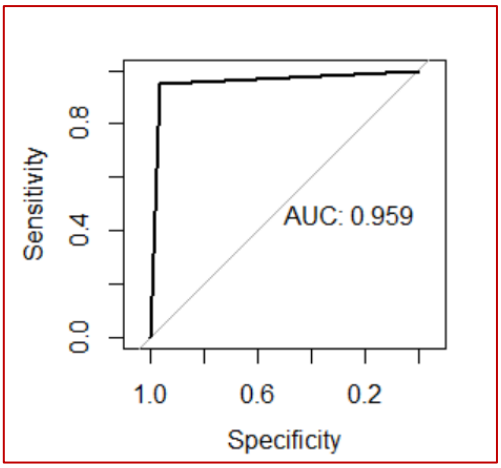


圖 13. 透過 Random Forest 改進後的 AUC 曲線

2-3 選擇模型

在這三種模型的建立後，發現以 Random Forest 改進過的 Bagging 在 Accuracy 與 AUC 都有最好的表現(表 2)，因此在第三章會做更深入的分析。

表 2. 模型比較

method	Accuracy	AUC
Classification tree	**0.8462**	**0.856**
Bagging & Random Forest	**0.9423**	**0.934**
	0.9615	**0.959**

第三章 Random Forest

根據上章節所看到透過 Random Forest 改進後的分析結果(圖 14)，在此章節會做更詳細的介紹[4]。

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       31        1
  Positive        1       19

               Accuracy : 0.9615
                 95% CI : (0.8679, 0.9953)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 5.986e-09

                  Kappa : 0.9188

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9500
            Specificity : 0.9688
         Pos Pred Value : 0.9500
         Neg Pred Value : 0.9687
             Prevalence : 0.3846
         Detection Rate : 0.3654
   Detection Prevalence : 0.3846
      Balanced Accuracy : 0.9594

       'Positive' Class : Positive
```

圖 14. 透過 Random Forest 改進後的分析結果

3-1 Accuracy

根據圖 14 中的混淆矩陣[5]，可從下列式子計算出 Accuracy(式 1)

$$Accuracy = \frac{TP+TN}{n} = \frac{19+31}{31+1+1+19} = \frac{50}{52} \approx 0.9615 \qquad (式\ 1)$$

3-2 H_0

設立 H_0：Accuracy ≦ No Information rate ，所有資料被分到樣本最多的那群實際為 Positive (陰性、沒有糖尿病)，根據式 2 計算可算出 NIR(No Information Rate)。從圖 14 可觀察出因 P -value < 0.05，所以拒絕 H_0，因此 Acc > NIR。

$$NIR = (31+1)/52 \approx 0.6154 \qquad (式\ 2)$$

3-3 H_0 vs. H_1

比較錯誤分類的比例，Mcnemar's Test P-Value：1 > 0.05，所以不會拒絕 H_0，而總樣本數是 52 , FP = 1 且 FN = 1，因此 1/52= 1/52 確實符合 H_0。

$$H_0 : \frac{FP}{n} = \frac{FN}{n} \quad vs. \quad H_1 : \frac{FP}{n} \neq \frac{FN}{n} \qquad \text{(式 3)}$$

3-4 Kappa

根據表 3，Random Forest 實驗中的 Kappa 值 0.9188 介於 0.81~1，表示模型結果與實際結果具有很好的一致性

表 3. Kappa 數值區分

Cohen's Kappa Statistic	Strengh of agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

3-5 其他統計值的計算

以下的式子為 Sensitivity(式 4)、Specificity(式 5)、Pos Pred Value(式 6)、Neg Pred Value(式 7)、Prevalence(式 8)、Detection Rate(式 9)、Detection Prevalence(式 10) 及 Balanced Accuracy(式 11)的計算方式，其中 Prevalence 為真實正樣本占總樣本的比例，Detection Rate 為實際為真且預測對占總樣本的比例，Detection Prevalence 預測為真的樣本占總樣本的比例。[6]

$$Sensitivity = \frac{TP}{TP+FN} = \frac{19}{19+1} = 0.9500 \qquad \text{(式 4)}$$

$$Specificity = \frac{TN}{FP+TN} = \frac{31}{1+31} \approx 0.9688 \qquad \text{(式 5)}$$

$$Pos\ Pred\ Value = \frac{TP}{TP+FP} = \frac{19}{19+1} = 0.9500 \qquad \text{(式 6)}$$

$$Neg\ Pred\ Value = \frac{TN}{FN+TN} = \frac{31}{1+31} \approx 0.9687 \quad \text{(式 7)}$$

$$Prevalence = \frac{TP+FN}{n} = \frac{19+1}{52} \approx 0.3846 \quad \text{(式 8)}$$

$$Detection\ Rate = \frac{TP}{n} = \frac{19}{52} \approx 0.3654 \quad \text{(式 9)}$$

$$Detection\ Prevalence = \frac{TP+FP}{n} = \frac{19+1}{52} \approx 0.3846 \quad \text{(式 10)}$$

$$Balanced\ Accruacy = \frac{Sensitivity + Specificity}{2} = \frac{0.9500+0.9688}{2} = 0.9594$$

(式 11)

3-6 Out-of-Bag Error

在隨機森林在抽取資料的過程中，有一部分資料會一直沒有被抽取到而此稱作 Out-of-Bag Error(以下簡稱 OOB)，而本實驗的 OOB 值為 1.28%，可從圖 15 中看到更詳細的數值變化，紅色線為預測陰性 Negative 的錯誤率；綠色線為預測陽性 Positive 的錯誤率；黑線為整體錯誤率。

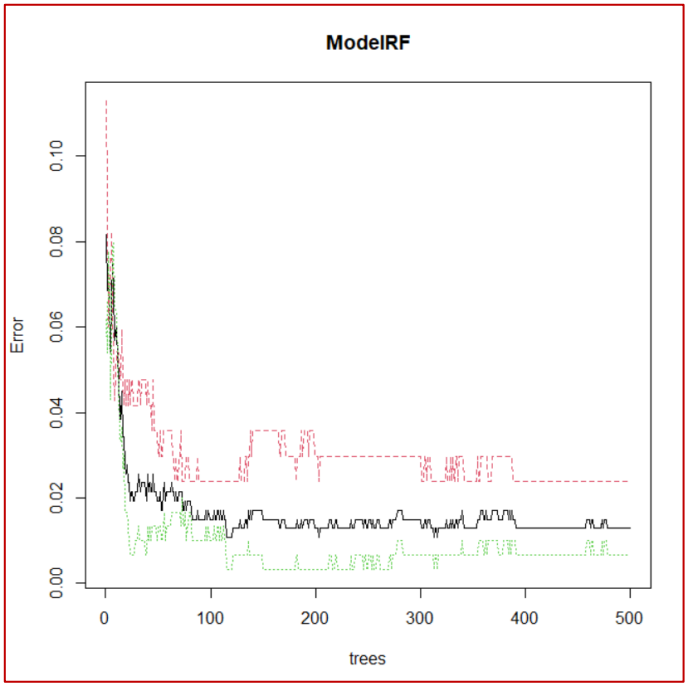


圖 15. Out-of-Bag Error

第四章 結論

從第二章的實驗中，Bagging 透過 Random Forest 改進後為最優，準確率為 0.9615，而從實驗中發現 polyuria(頻尿)、polydipsia(口渴)及 gender(性別)這三項特徵對於是否罹患糖尿病的影響最大。

對於醫學中的應用，隨機森林可用於識別醫學中分組的正確組合，並通過分析患者的醫療記錄來識別疾病，如依據患者症狀(是否容易口渴、是否頻尿)，來去預測此人是否有糖尿病，藉此幫助增加診斷時的效率，來去決定是否要做進一步的檢查及確認。

參考文獻

[1]Kaggle, Diabetes UCI Dataset(2020), 取自:
https://www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset

[2] Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.

[3] Report abuse , Random Forest Regression in R (2016), 取自：
https://gist.github.com/geandersonlenz/aa7ec3b46e797029f2ad6f0a09176de9

[4] Ryan Lu，Learning Model : Random Forest (2019), 取自：
https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/learning-model-random-forest-ca4e3f8a63d3

[5] chengdehe，混淆矩阵（confusionMatrix）——基于R语言的输出结果理解(2020), 取自：
https://blog.csdn.net/chengdehe/article/details/105008115

[6] Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. Springer.