



巨量資料分析工具與應用期末報告

# 以美國少數民族之女性 糖尿病之分析

研究同學：張家豪072214112、朱家佑072214113  
黃聖岳072214120、林妤柔072214121  
三子傑072214124

# CONTENT 目錄

## S



PART 1

資料來源

PART 2

描述性統計

PART 3

決策樹

PART 4

K-means

PART 5

結論

PART 6

資料來源

# 小組分工

小組報告-工作分配	
張家豪072214112	尋找可用資料集、所有資料分析
朱家佑072214113	文書製作與整理
黃聖岳072214120	文書製作與整理
林妤柔072214121	尋找可用資料集、k-mean、文書微調
三子傑072214124	簡報製作與整理

# PART1 資料來源

1. 使用 KAGGLE 資料科學競賽平台上所提供的開放資料集
2. 資料集為美國21歲以上印地安族女性之糖尿病數據集
3. 於 2015年 10月7日公布，由美國國家糖尿病、消化和腎臟疾病研究所提供

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	0	110	80	25	0	30	0.262	20	1

圖1 印地安女性糖尿病資料集

# PART2 描述性統計

## 2.1 資料匯入與清洗

```
> #input_data
> Diabetes <- read.csv("d:/Big Data Analysis/Final_Presentation/Pima_Indians_Diabetes .csv", header = TRUE)
>
> #data_clean
> library(magrittr)
Warning message:
套件 'magrittr' 是用 R 版本 4.1.2 來建造的
> Diabetes_Clean <- Diabetes[, c(1,2,3,4,5,6,7,8,9)] %>% na.omit
> |
```

圖2-1 資料匯入與清洗結果

## 2.2 分析維度與欄位

輸出表示共有768筆資料、9個欄位，分別為：Pregnancies、Glucose、BloodPressure、SkinThickness、Insulin、BMI、DiabetesPedigreeFunction、Age、Outcome。

```
> #check the dimensionality
> dim(Diabetes_Clean)
[1] 768 9
> #column names
> names(Diabetes_Clean)
[1] "Pregnancies"      "Glucose"           "BloodPressure"
[4] "SkinThickness"    "Insulin"           "BMI"
[7] "DiabetesPedigreeFunction" "Age"               "Outcome"
```

圖2-2 分析維度與欄位結果

# PART2 描述性統計

## 2. 3分析結構

```
> #structure
> str(Diabetes_Clean)
'data.frame': 768 obs. of 9 variables:
 $ Pregnancies      : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose           : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure     : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness     : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin           : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI               : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age              : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome           : int  1 0 1 0 1 0 1 0 1 1 ...
```

圖2-3 分析結構結果

## 2. 4分析屬性

```
> #attributes
> attributes(Diabetes_Clean)
$names
[1] "Pregnancies"      "Glucose"          "BloodPressure"
[4] "SkinThickness"    "Insulin"          "BMI"
[7] "DiabetesPedigreeFunction" "Age"              "Outcome"

$row.names
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
[76] 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

===== 中間省略 =====

[701] 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725
[726] 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750
[751] 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768

$class
[1] "data.frame"
```

圖2-4 分析屬性結果

# PART2 描述性統計

## 2. 5分析前五筆資料

```
> #get the first 5 rows
> Diabetes_Clean[1:5,]
  Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
1           6     148             72           35         0   33.6              0.627    50         1
2           1      85             66           29         0   26.6              0.351    31         0
3           8     183             64           0          0   23.3              0.672    32         1
4           1      89             66           23        94   28.1              0.167    21         0
5           0     137             40           35       168   43.1              2.288    33         1
> head(Diabetes_Clean)
  Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
1           6     148             72           35         0   33.6              0.627    50         1
2           1      85             66           29         0   26.6              0.351    31         0
3           8     183             64           0          0   23.3              0.672    32         1
4           1      89             66           23        94   28.1              0.167    21         0
5           0     137             40           35       168   43.1              2.288    33         1
6           5     116             74           0          0   25.6              0.201    30         0
```

圖2-5 分析前五筆資料結果

## 2. 6取得資料摘要

顯示各個欄位的統計相關值：最小值、最大值、第一四分位數、中位數、平均數、第三四分位數。

```
> #summary
> summary(Diabetes_Clean)
  Pregnancies      Glucose  BloodPressure  SkinThickness      Insulin      BMI
Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00   Min.   : 0.0   Min.   : 0.00
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00  1st Qu.: 0.00   1st Qu.: 0.0   1st Qu.:27.30
Median : 3.000   Median :117.0   Median : 72.00  Median :23.00   Median : 30.5   Median :32.00
Mean   : 3.845   Mean   :120.9   Mean   : 69.11  Mean   :20.54   Mean   : 79.8   Mean   :31.99
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00  3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60
Max.   :17.000   Max.   :199.0   Max.   :122.00  Max.   :99.00   Max.   :846.0   Max.   :67.10
DiabetesPedigreeFunction  Age  Outcome
Min.   :0.0780           Min.   :21.00   Min.   :0.000
1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
Median :0.3725           Median :29.00   Median :0.000
Mean   :0.4719           Mean   :33.24   Mean   :0.349
3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
Max.   :2.4200           Max.   :81.00   Max.   :1.000
>
```

圖2-6 資料摘要結果

## PART2 描述性統計

### 2. 7確診頻率分析

由2-7圖可知，沒有確診的人數為500人，有確診的人數為268人。

```
> #frequency  
> table(Diabetes_Clean$Outcome)  
  
 0    1  
500 268  
> #pie chart  
> pie(table(Diabetes_Clean$Outcome))  
> |
```

圖2-7 分析前五筆資料結果

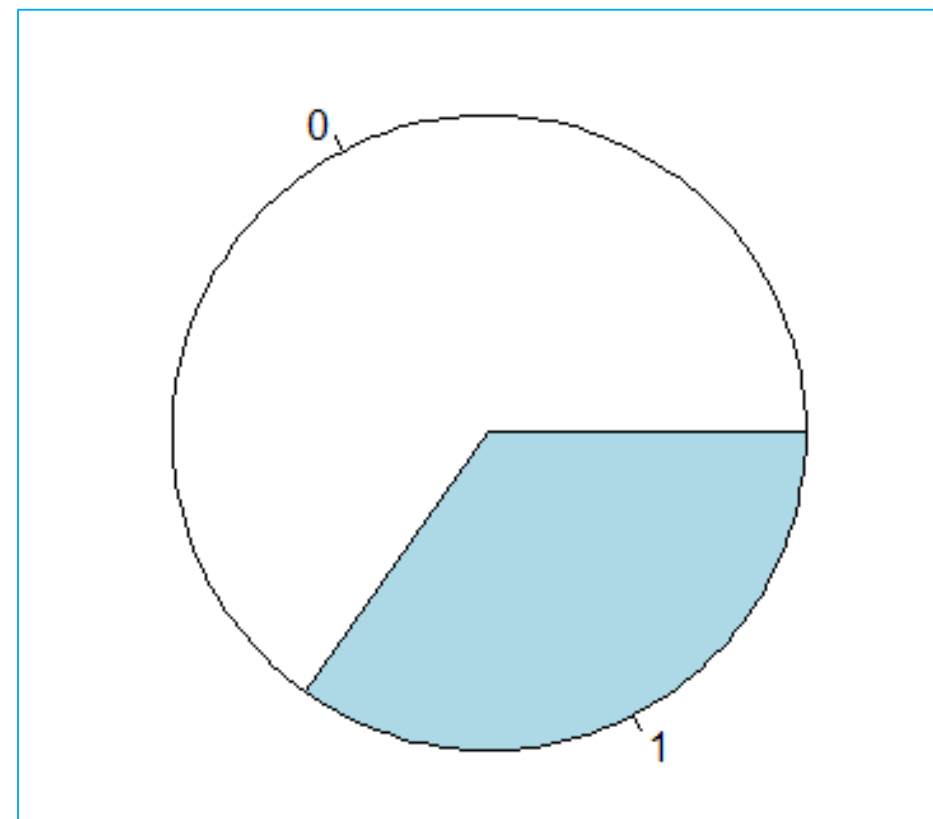


圖2-8 分析屬性結果



# PART3 決策樹

## 3.1 以rpart為例

### 3.1.1 資料預處理

```
> #Decision Trees
> #rpart
> # 隨機排列資料集
> n<-nrow(Diabetes_Clean)
> set.seed(123)
> shuffled_Diabetes <- Diabetes_Clean[sample(n), ]    #sample: random select
>
> # 將資料集分為訓練與測試
> train_Indices <- 1:round(0.7 * n)
> train <- shuffled_Diabetes[train_Indices, ]
> test_Indices <- (round(0.7 * n) + 1):n
> test <- shuffled_Diabetes[test_Indices, ]
> |
```




▶ shuffled_Dia...	768 obs. of 9 variables	
▶ test	230 obs. of 9 variables	
▶ train	538 obs. of 9 variables	
Values		
n	768L	
test_Indices	int [1:230] 539 540 541 542 543 544 ...	
train_Indices	int [1:538] 1 2 3 4 5 6 7 8 9 10 ...	

圖3-1-1 資料處理結果

# PART3 決策樹

## 3.1 以rpart為例

### 3.1.2 繪出決策樹

```
> # 建立一個決策樹模型  
> Diabetes_rpart_tree <- rpart(formula = Outcome ~ ., data = train, method = "class")  
> prediction <- predict(Diabetes_rpart_tree, newdata = test, type="class")  
> rpart.plot(Diabetes_rpart_tree ,extra=106)  
> |
```

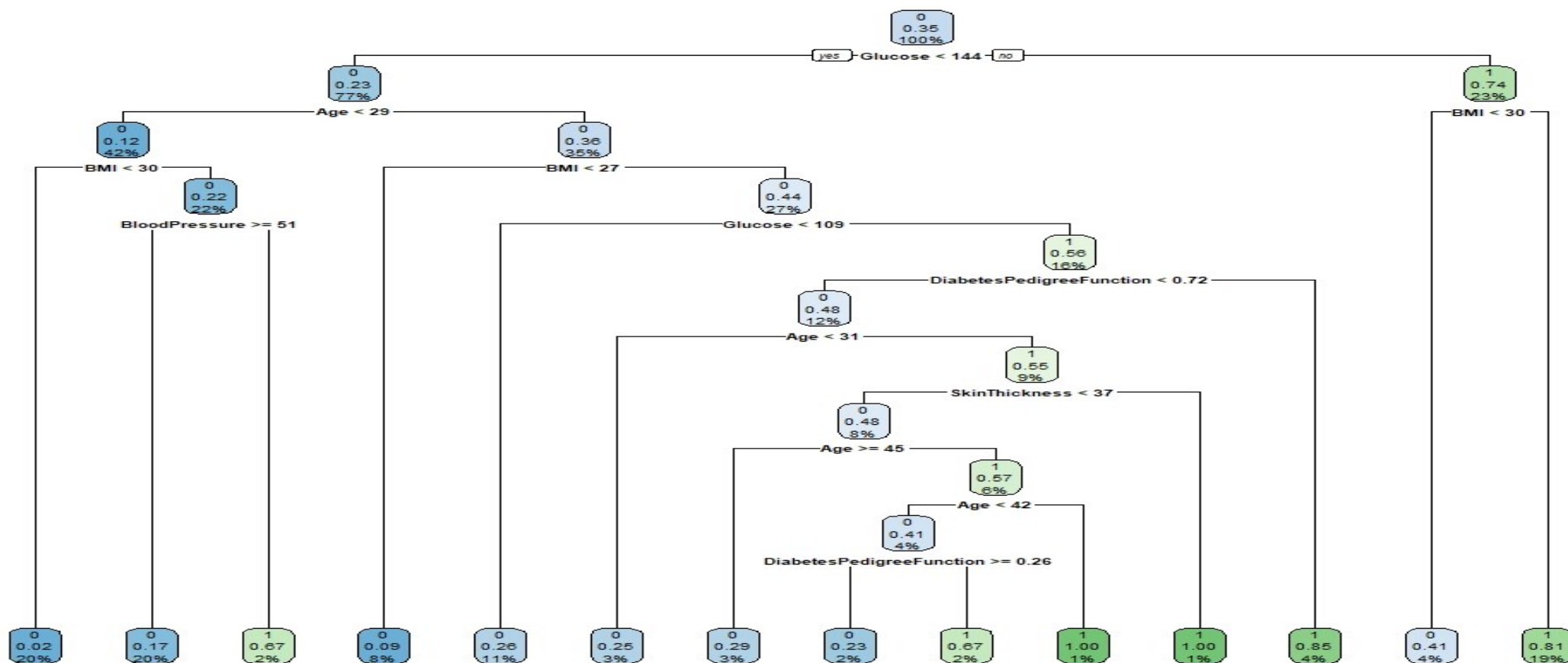


圖3-1-2 rpart決策樹結果

# PART3 決策樹

## 3.1 以rpart為例

### 3.1.2(續) 決策樹結果(由左至右)

- 血糖值 $<144$ ，年齡 $<29$ ，BMI $<30$ ，未患有糖尿病，佔有20%。
- 血糖值 $<144$ ，年齡 $<29$ ，BMI $>30$ ，血壓 $\geq 51\text{mmHg}$ ，未患有糖尿病，佔有20%。
- 血糖值 $<144$ ，年齡 $<29$ ，BMI $>30$ ，血壓 $<51\text{mmHg}$ ，患有糖尿病，佔有2%。
- 血糖值 $<144$ ，年齡 $>29$ ，BMI $<27$ ，未患有糖尿病，佔有8%。
- 年齡 $>29$ ，BMI $>27$ ，血糖值 $<109$ ，未患有糖尿病，佔有11%。
- 血糖值 $109\sim144$ ，年齡 $29\sim31$ ，BMI $>27$ ，糖尿病函數 $<0.72$ ，未患有糖尿病，佔有3%。
- 血糖值 $109\sim144$ ，年齡 $\geq 45$ ，BMI $>27$ ，糖尿病函數 $<0.72$ ，皮膚厚度 $<37$ ，未患有糖尿病，佔有3%。
- 血糖值 $109\sim144$ ，年齡 $42\sim31$ ，BMI $>27$ ，糖尿病函數 $0.26\sim0.72$ ，皮膚厚度 $<37$ ，未患有糖尿病，佔有2%。
- 血糖值 $109\sim144$ ，年齡 $42\sim31$ ，BMI $>27$ ，糖尿病函數 $\leq 0.26$ ，皮膚厚度 $<37$ ，患有糖尿病，佔有2%。
- 血糖值 $109\sim144$ ，年齡 $45\sim42$ ，BMI $>27$ ，糖尿病函數 $<0.72$ ，皮膚厚度 $<37$ ，未患有糖尿病，佔有1%。
- 血糖值 $109\sim144$ ，年齡 $>31$ ，BMI $>27$ ，糖尿病函數 $<0.72$ ，皮膚厚度 $>37$ ，患有糖尿病，佔有1%。
- 血糖值 $109\sim144$ ，年齡 $>29$ ，BMI $>27$ ，糖尿病函數 $>0.72$ ，患有糖尿病，佔有4%。
- 血糖值 $>144$ ，BMI $<30$ ，未患有糖尿病，佔有4%。
- 血糖值 $>144$ ，BMI $>30$ ，患有糖尿病，佔有19%。

# PART3 決策樹

## 3.1 以rpart為例

### 3.1.3 混淆矩陣

```
> confusionMatrix <- table(x = test$outcome, y = prediction, dnn=c("Actual", "Prediction"))
> confusionMatrix
      Prediction
Actual    0    1
    0 128   21
    1   46   35

>
> # 獲得TP, TN, FP, FN
> TP <- confusionMatrix[1, 1]
> TN <- confusionMatrix[2, 2]
> FP <- confusionMatrix[2, 1]
> FN <- confusionMatrix[1, 2]
>
> # 計算accuracy, TPR, FPR
> #敏感度 Sensitivity
> TPR <- TP/(TP + FN)
> TPR
[1] 0.8590604
> #特異度 Specificity
> TNR <- TN/(FP + TN)
> TNR
[1] 0.4320988
> #False Postive Rate
> FPR <- FP/(FP + TN)
> FPR
[1] 0.5679012
> #False Negative Rate
> FNR <- FN/(TP + FN)
> FNR
[1] 0.1409396
>
> accuracy1 <- (TP + TN)/(TP + TN + FP + FN)
> accuracy1
[1] 0.7086957
> accuracy2 <- sum(diag(confusionMatrix))/sum(confusionMatrix)# 試試這樣算
> accuracy2
[1] 0.7086957
```



混淆矩陣顯示測試集有230筆資料。

敏感度為 $128 / (128 + 21) = 0.859$

特異度為 $35 / (35 + 46) = 0.432$

偽陽性率為 $46 / (35 + 46) = 0.568$

偽陰性率為 $21 / (128 + 21) = 0.141$

正確率為 $(128 + 35) / (128 + 21 + 46 + 35) = 0.709$

圖3-1-3 混淆矩陣結果

# PART3 決策樹

## 3.2 以ctree為例

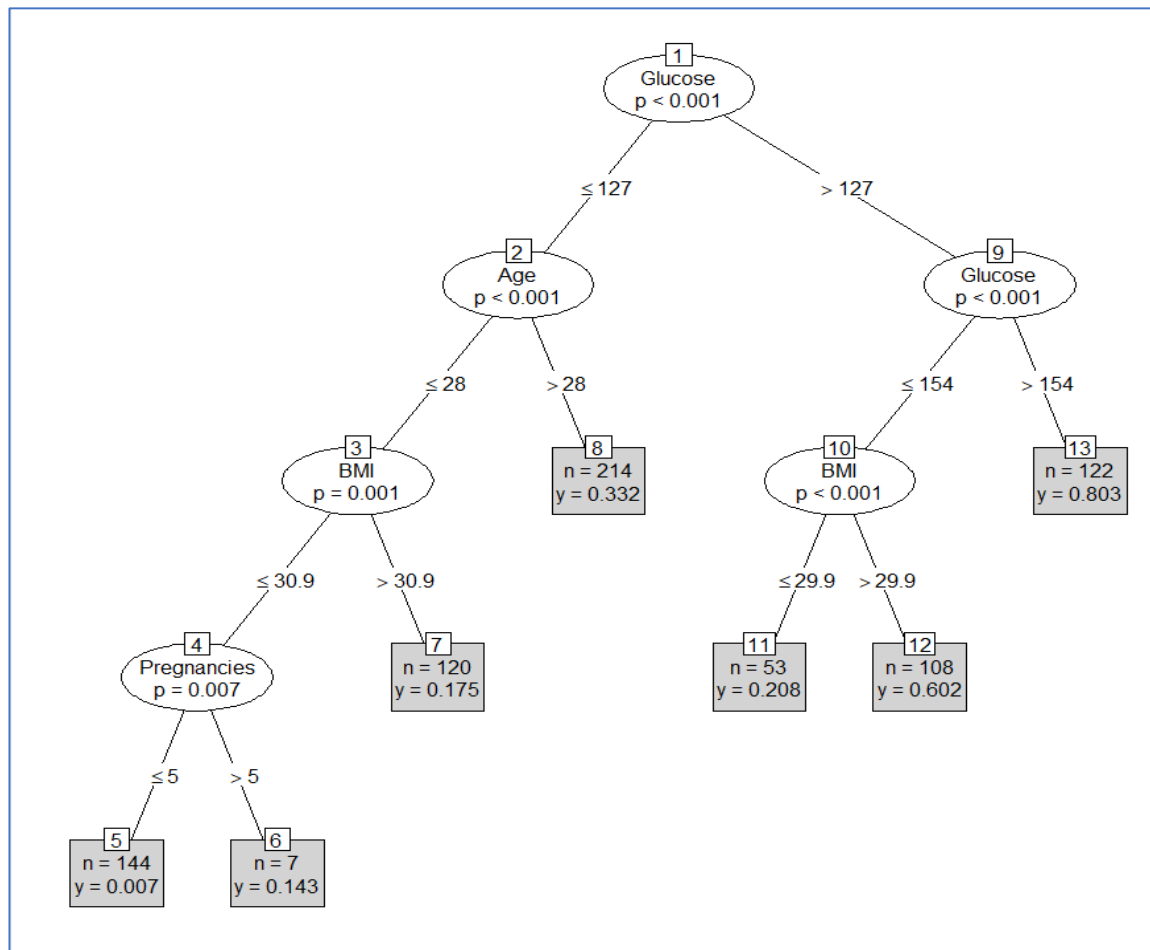


圖3-2-1 ctree決策樹結果

```
> #Decision Trees
> #ctree
> Diabetes_ctree <- ctree(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThicknes
s + Insulin + BMI + DiabetesPedigreeFunction + Age, data=Diabetes_Clean)
> print(Diabetes_ctree)

Conditional inference tree with 7 terminal nodes

Response: Outcome
Inputs: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedig
reeFunction, Age
Number of observations: 768

1) Glucose ≤ 127; criterion = 1, statistic = 166.975
2) Age ≤ 28; criterion = 1, statistic = 28.915
3) BMI ≤ 30.9; criterion = 0.999, statistic = 14.419
4) Pregnancies ≤ 5; criterion = 0.993, statistic = 11.077
5)* weights = 144
4) Pregnancies > 5
6)* weights = 7
3) BMI > 30.9
7)* weights = 120
2) Age > 28
8)* weights = 214
1) Glucose > 127
9) Glucose ≤ 154; criterion = 1, statistic = 26.787
10) BMI ≤ 29.9; criterion = 1, statistic = 20.552
11)* weights = 53
10) BMI > 29.9
12)* weights = 108
9) Glucose > 154
13)* weights = 122

>
> #ctree chart
> plot(Diabetes_ctree)
> plot(Diabetes_ctree, type="simple")
>
```

圖3-2-2 資料與處理結果

# PART3 決策樹

## 3.2 以ctree為例(續)

- 血糖值 $\leq 127$ ，年齡 $\leq 28$ ，BMI $\leq 30.9$ ，懷孕週期 $\leq 5$ 個月(包含未懷孕)，人數有144位，得糖尿病機率是0.7%。
- 血糖值 $\leq 127$ ，年齡 $\leq 28$ ，BMI $\leq 30.9$ ，懷孕週期 $> 5$ 個月，人數有7位，得糖尿病機率是14.3%。
- 血糖值 $\leq 127$ ，年齡 $\leq 28$ ，BMI $> 30.9$ ，人數有120位，得糖尿病機率是17.5%。
- 血糖值 $\leq 127$ ，年齡 $> 28$ ，人數有214位，得糖尿病機率是33.2%。
- 血糖值127~154，BMI $\leq 29.9$ ，人數有53位，得糖尿病機率是20.8%。
- 血糖值127~154，BMI $> 29.9$ ，人數有108位，得糖尿病機率是60.2%。
- 血糖值 $> 154$ ，人數有122位，得糖尿病機率是80.3%。

# PART4 K-means

## 4.1 選擇K值

經過K值訓練結果，我們將資料分為3群，  
分析結果如下。

```
> #k 要怎麼選擇 - Hands on
> ratio_ss <- rep(NA, times = 11)
> for (k in 1:length(ratio_ss)) {
+   fit_km <- kmeans(Diabetes_Clean, centers=k, nstart=20)
+   ratio_ss[k] <- fit_km$tot.withinss/fit_km$totss
+ }
> plot(ratio_ss, type="b", xlab="k", main = "screepplot") # "b" as in both
>
> ratio_ss
[1] 1.00000000 0.44271266 0.25081748 0.18521192 0.14953759 0.12626788 0.11219911
[8] 0.09992566 0.09379335 0.08768471 0.08279027
>
```

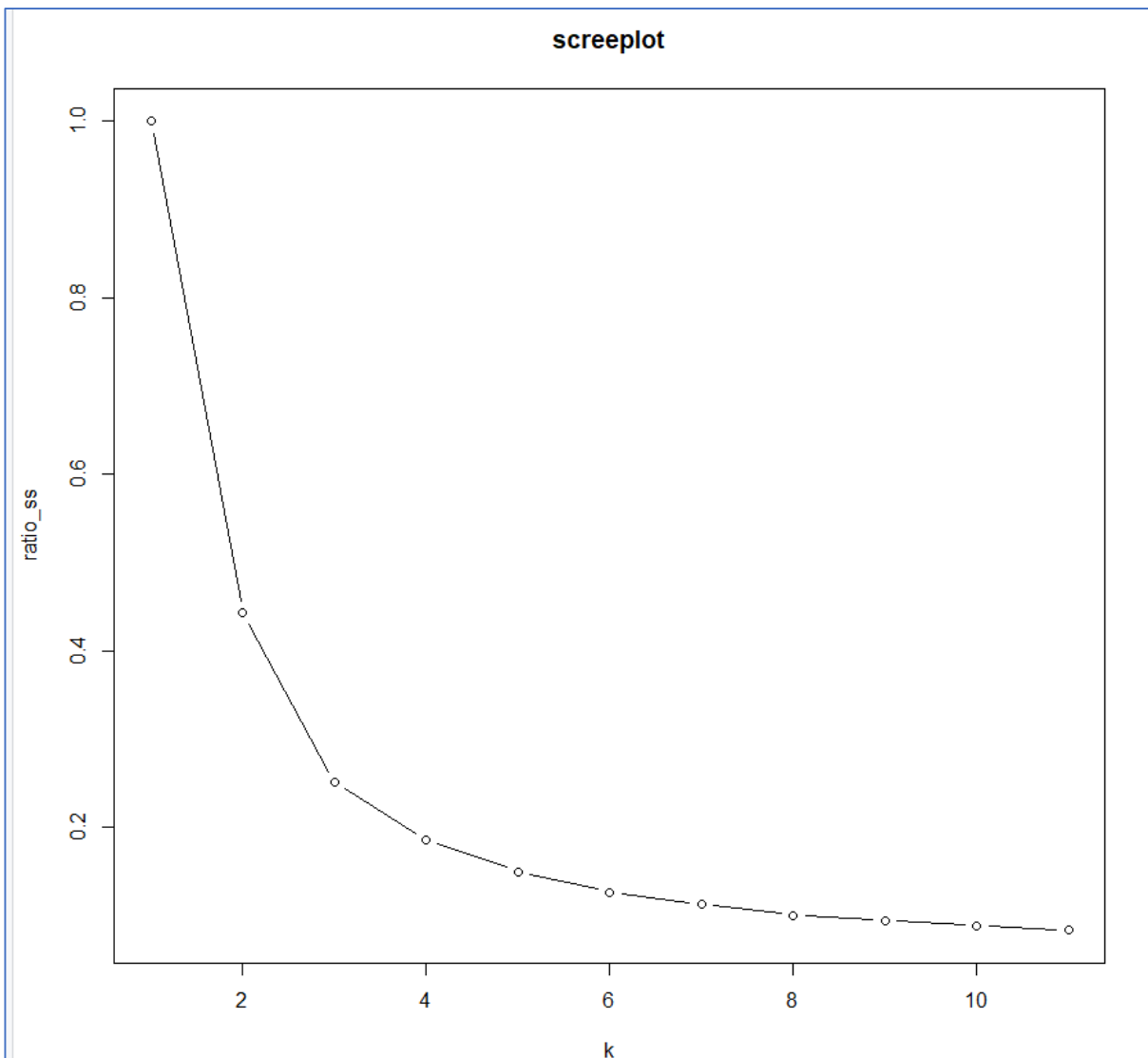


圖4-1 K值訓練圖

# PART4 K-means

## 4.2 K-means

```
736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756
  1   3   1   3   1   3   3   3   1   3   3   1   1   3   1   1   1   1   2   1   3
757 758 759 760 761 762 763 764 765 766 767 768
  1   1   1   1   1   1   1   3   1   3   1   1

within cluster sum of squares by cluster:
[1] 1342892.3 653441.4 917159.0
(between_SS / total_SS = 74.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"       "
```

===== 中間省略 =====

```
> kc <- kmeans(Diabetes_Clean, 3)
> kc
K-means clustering with 3 clusters of sizes 495, 38, 235

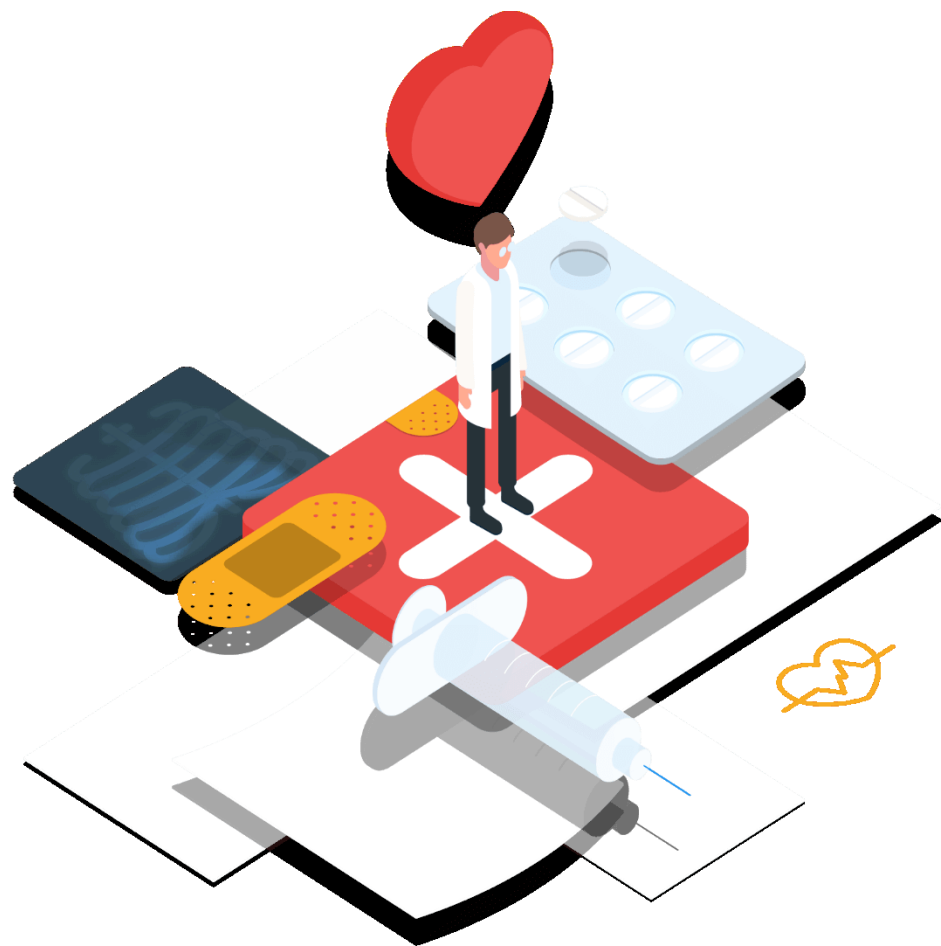
Cluster means:
  Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI
1    3.981818 114.0081     67.77172     14.99798  14.4000 30.80545
2    4.026316 158.4474     72.00000     32.26316 441.2895 35.10789
3    3.527660 129.3277     71.44681     30.30638 159.1021 33.98936
  DiabetesPedigreeFunction  Age  Outcome
1         0.4319313 33.75960 0.2989899
2         0.5692105 34.76316 0.5789474
3         0.5402766 31.90213 0.4170213

Clustering vector:
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
  1  1  1  3  3  1  1  1  2  1  1  1  1  2  3  1  3  1  1  3  3
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
  1  1  1  3  3  1  3  3  1  1  3  1  1  1  3  1  1  1  3  1  1
43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
  1  3  1  1  1  1  1  1  1  1  1  2  2  1  2  3  1  3  1  1  1
64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84
```

圖4-2 K-means結果



## PART5 結論



以rpart、ctree等決策樹分析，發現患有糖尿病的機率與血糖值、年齡、BMI有正相關性，K-means分群準確率為74.9%。

## PART6 資料來源

---

Diges.& Kidney Dis(2020, August 6).Diabetes  
Dataset.kaggle.

<https://www.kaggle.com/mathchi/diabetes-data-set>