1.利用 Weka 對 CardiologyCategorical.csv 進行 Unsupervised Clustering。使用 Simple K-Means 演算法做分群，其產生的群不能大於六群。調整群的樹木或刪除較不重要的屬性或是調整群的初始中心，再進行一次分群動作，直到找出最好的分群模型為止，再根據產生的結果進行分析：

(a)最好的分群模型判斷條件為何？列出使用了那些屬性、初始中心為多少等等。(10%)

The smaller value of sum of squared errors it is, the better the cluster model is.

NumCluster = 2 Within cluster sum of squared errors: 903.5806368731734
Seed = 10
DistanceFunction = EuclideanDistance

NumCluster = 3 Within cluster sum of squared errors: 632.1992389086358
Seed = 10
DistanceFunction = EuclideanDistance

NumCluster = 4 Within cluster sum of squared errors: 589.8257599289915
Seed = 10
DistanceFunction = EuclideanDistance

NumCluster = 5 Within cluster sum of squared errors: 553.4896470752894
Seed = 10
DistanceFunction = EuclideanDistance

This is the smallest value.

NumCluster = 6 Within cluster sum of squared errors: 531.1182943548731
Seed = 10
DistanceFunction = EuclideanDistance

NumCluster = 2 Within cluster sum of squared errors: 598.0254236996501
Seed = 100
DistanceFunction = EuclideanDistance
Ignore: Sex

NumCluster = Sum of within cluster distances: 868.6928788183081
2
Seed = 100
DistanceFunction = ManhattanDistance

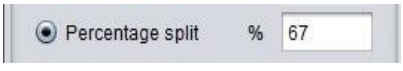
Final cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(303.0)	(59.0)	(85.0)	(22.0)	(65.0)	(43.0)	(29.0)
age	54.3663	49.9492	52.9412	58.6818	55.4462	60.8837	52.1724
sex	Male	Male	Female	Male	Male	Male	Male
chest pain type	Asymptomatic	Abnormal Angina	NoTang	Angina	Asymptomatic	Asymptomatic	Asymptomatic
blood pressure	131.6238	131.6102	128.7529	136.6364	133.6308	137.5814	122.931
cholesterol	246.264	242.6441	244.9059	260.5909	246.6	258.093	228.4483
Fasting blood sugar <120	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
resting ecg	Normal	Hyp	Normal	Hyp	Normal	Hyp	Normal
maximum heart rate	149.6469	164.5424	156.4471	157.5	127.6462	139.1628	158.3103
angina	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
peak	1.0396	0.3593	0.5459	1.2	1.7938	2.1814	0.3655
slope	Up	Up	Up	Flat	Flat	Flat	Up
#colored vessels	0.6667	0.3051	0.2706	0.2273	0.6769	2.0233	0.8621
thal	Normal	Normal	Normal	Rev	Rev	Rev	Rev
class	Healthy	Healthy	Healthy	Healthy	Sick	Sick	Sick

(b)承上題設定,此結果分了幾群 ?試著說明各群所代表之意義、特色為何 ?(10%)

Total 6 clusters.
Which respectively are:
<Class = Healthy>
Cluster0 which includes 59 instances.
Cluster1 which includes 85 instances.
Cluster2 which includes 22 instances.
<Class = Sick>
Cluster3 which includes 65 instances.
Cluster4 which includes 43 instances.
Cluster5 which includes 29 instances.

2.將 CardiologyCategorical.csv 分割成 training data: 203 筆，test data: 100 筆。使用上一題的 input attributes，並將 Attribute: class 設定為 output，利用 Naïve Bayes 進行 supervised learning：
(a)請說明如何將 CardiologyCategorical.csv 分割成訓練集和測試集。(5%)
Take advantage of Percentage Split, in order to split all instances into training data and test data. The former includes 203 instances and stands 67% of the whole data set, the latter includes 100 instances which stands for 33% of the whole data set.



(b)觀察訓練出來的模型，屬於 class = sick 及 class = healthy 的 instance 各有何特色？與第一題(b)比較有何相同或相異？(15%)

Attribute	Class Sick (0.46)	Healthy (0.54)
age		
mean	56.6	52.4655
std. dev.	7.9344	9.5284
weight sum	138	165
precision	1.2	1.2
sex		
Male	115.0	94.0
Female	25.0	73.0
[total]	140.0	167.0
chest pain type		
Asymptomatic	105.0	40.0
Abnormal Angina	10.0	42.0
Angina	8.0	17.0
NoTang	19.0	70.0
[total]	142.0	169.0
blood pressure		
mean	134.3723	129.1273
std. dev.	18.6815	16.0421
weight sum	138	165
precision	2.2083	2.2083
cholesterol		
mean	251.1385	242.1789
std. dev.	49.3271	53.384
weight sum	138	165
precision	2.9007	2.9007
Fasting blood sugar <120		
FALSE	117.0	143.0
TRUE	23.0	24.0
[total]	140.0	167.0
resting ecg		
Hyp	80.0	69.0
Normal	57.0	97.0
Abnormal	4.0	2.0
[total]	141.0	168.0
maximum heart rate		
mean	139.1005	158.4174
std. dev.	22.5146	19.0916
weight sum	138	165
precision	1.4556	1.4556
angina		
TRUE	77.0	24.0
FALSE	63.0	143.0
[total]	140.0	167.0
peak		
mean	1.5932	0.5868
std. dev.	1.2957	0.7744
weight sum	138	165
precision	0.159	0.159
slope		
Flat	92.0	50.0
Up	36.0	108.0
Down	13.0	10.0
[total]	141.0	168.0
#colored vessels		
mean	1.1449	0.2667
std. dev.	1.0112	0.6241
weight sum	138	165
precision	1	1
thal		
Rev	91.0	29.0
Normal	37.0	132.0
Fix	13.0	7.0
[total]	141.0	168.0

Take <maximum heart rate> for example.

Cluster	results	Cluster v.s Naïve Bayes	
Cluster0: 131.6102	Healthy	129.1273	same
Cluster1: 128.7529	Healthy	129.1273	same
Cluster2: 136.6364	Healthy	134.3723	Not same
Cluster3: 133.6308	Sick	134.3723	same

Cluster4: 137.5814	Sick	134.3723	same
Cluster5: 122.931	Sick	129.1273	Not same

Take <Blood pressure> for example.

Cluster	results	Cluster v.s Naïve Bayes	
Cluster0: 164.5424	Healthy	158.4714	same
Cluster1: 156.4471	Healthy	158.4714	Same
Cluster2: 157.5	Healthy	158.4714	Same
Cluster3: 127.6462	Sick	139.1105	Same
Cluster4: 139.1628	Sick	139.1105	ame
Cluster5: 158.3103	Sick	158.4714	Not same

(c)Test data 正確率為多少 ?預測兩個 class 的 F-Measure 各為多少 ?(截圖即可，不用計算過程) (10%)

Correctly Classified Instances	89	89 %
Incorrectly Classified Instances	11	11 %

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.809	0.038	0.950	0.809	0.874	0.785	0.954	0.958	Sick
	0.962	0.191	0.850	0.962	0.903	0.785	0.954	0.957	Healthy
Weighted Avg.	0.890	0.119	0.897	0.890	0.889	0.785	0.954	0.957	

(d)第 100 筆 test data 透過此 model 的預測 class 為 Sick 還是 Healthy ?它的 Probability distribution 各為多少 ?(請列出算式) (20%)

No. 100 instance 1:Sick 1:Sick 0.896 Predict as Sick.

56	Male	Asymptomatic	125	249	TRUE	Hyp	144	TRUE	1.2	Flat	1	Normal	Sick

Age: $\frac{1}{\sqrt{2\pi} * 9.082} e^{-\frac{(56-54.366)^2}{2*9.082^2}} = 0.043391$

Sex:

Male: Output=Sick: 144 instances. $P = \frac{114}{138}$

Chest pain type:

Asymptomatic: Output=Sick: 104 instances. $P = \frac{104}{138}$

Blood pressure: $\frac{1}{\sqrt{2\pi} * 17.538} e^{-\frac{(125-131.624)^2}{2*17.538^2}} = 0.0215$

Cholesterol: $\frac{1}{\sqrt{2\pi} * 51.831} e^{-\frac{(249-246.264)^2}{2*51.831^2}} = 0.007691$

Fasting blood sugar<120:

TRUE: Output=Sick: 22 instances. $P = \frac{22}{138}$

Resting ecg:

Hyper: Output=Sick: 79 instances. $P = \frac{79}{138}$

Maximum:

Angina: $\frac{1}{\sqrt{2\pi} * 22.905} e^{-\frac{(144-149.647)^2}{2*22.905^2}} = 0.01702$

True: Output=Sick: 76 instances. $P = \frac{76}{138}$

Peak: $\frac{1}{\sqrt{2\pi} * 1.161} e^{-\frac{(1.2-1.04)^2}{2*1.161^2}} = 0.341195$

Slope:

Flat: Output=Sick: 91 instances. $P = \frac{91}{138}$

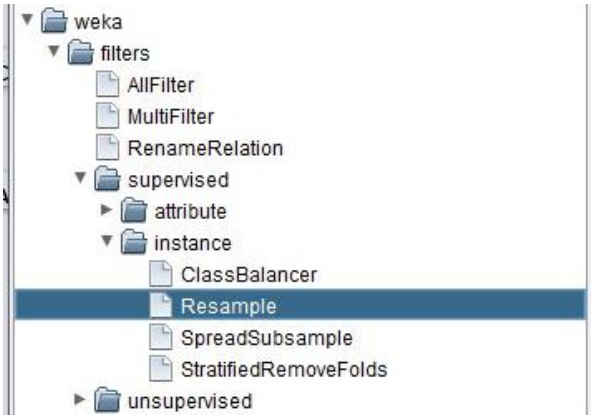
colored $\frac{1}{\sqrt{2\pi} * 0.934} e^{-\frac{(0.667-1)^2}{2*0.934^2}} = 0.40671$ vessels:

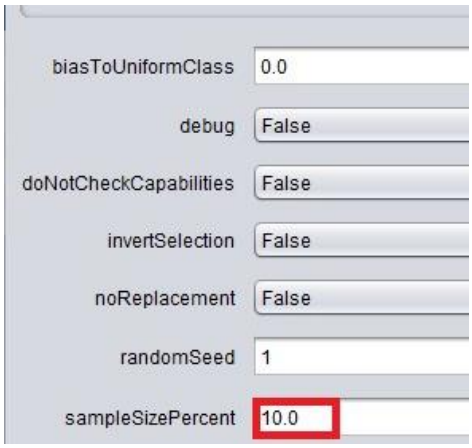
Thal:

Normal: Output=Sick: 36 instances. $P = \frac{36}{138}$

3.利用 Weka 對 diabetes.arff 進行分類。請比較該資料集利用 J48 與 Naïve Bayes 的分類運算下，有無明顯的優劣差別。

(a)載入 diabetes.arff 並在前處理隨機抽樣(Resample)10%的資料集來做分類 (sampleSizePercent = 10)





Choose>filters>supervised>instance>resample>set SampleSizePercent =10

(b)採用同樣 10 取樣的資料集，Test options 選擇 Cross-validation，並錄 J48 和 Naïve Bayes 分類後的正確率 (請將 10 次結果都截圖)

J48:

1	1:tested_negative	1:tested_negative		0.946	1	1:tested_negative	1:tested_negative	1
2	1:tested_negative	1:tested_negative		1	2	1:tested_negative	1:tested_negative	1
3	1:tested_negative	1:tested_negative		0.946	3	1:tested_negative	1:tested_negative	1
4	1:tested_negative	1:tested_negative		0.946	4	1:tested_negative	1:tested_negative	1
5	1:tested_negative	1:tested_negative		0.946	5	1:tested_negative	2:tested_positive	+ 0.75
6	2:tested_positive	2:tested_positive		0.929	6	2:tested_positive	2:tested_positive	0.75
7	2:tested_positive	1:tested_negative	+	1	7	2:tested_positive	1:tested_negative	+ 1
8	2:tested_positive	1:tested_negative	+	0.946	8	2:tested_positive	1:tested_negative	+ 1

1	1:tested_negative	1:tested_negative		0.946	1	1:tested_negative	1:tested_negative	1
2	1:tested_negative	1:tested_negative		0.946	2	1:tested_negative	2:tested_positive	+ 0.875
3	1:tested_negative	1:tested_negative		0.946	3	1:tested_negative	2:tested_positive	+ 0.875
4	1:tested_negative	1:tested_negative		1	4	1:tested_negative	1:tested_negative	0.974
5	1:tested_negative	1:tested_negative		0.946	5	1:tested_negative	1:tested_negative	0.974
6	2:tested_positive	1:tested_negative	+	1	6	2:tested_positive	2:tested_positive	0.875
7	2:tested_positive	1:tested_negative	+	0.946	7	2:tested_positive	2:tested_positive	1
8	2:tested_positive	2:tested_positive		0.929	8	2:tested_positive	2:tested_positive	0.875

1	1:tested_negative	1:tested_negative		0.972	1	1:tested_negative	1:tested_negative	1
2	1:tested_negative	1:tested_negative		0.972	2	1:tested_negative	1:tested_negative	1
3	1:tested_negative	1:tested_negative		0.972	3	1:tested_negative	1:tested_negative	1
4	1:tested_negative	1:tested_negative		0.972	4	1:tested_negative	1:tested_negative	1
5	1:tested_negative	1:tested_negative		1	5	1:tested_negative	1:tested_negative	0.667
6	2:tested_positive	2:tested_positive		1	6	2:tested_positive	2:tested_positive	0.929
7	2:tested_positive	2:tested_positive		1	7	2:tested_positive	1:tested_negative	+ 1
8	2:tested_positive	1:tested_negative	+	1	8	2:tested_positive	1:tested_negative	+ 1

1	1:tested_negative	1:tested_negative		1	1	1:tested_negative	1:tested_negative	0.947
2	1:tested_negative	1:tested_negative		0.973	2	1:tested_negative	1:tested_negative	0.947
3	1:tested_negative	1:tested_negative		0.973	3	1:tested_negative	1:tested_negative	0.947
4	1:tested_negative	1:tested_negative		0.973	4	1:tested_negative	2:tested_positive	+ 1
5	1:tested_negative	1:tested_negative		0.973	5	1:tested_negative	2:tested_positive	+ 1
6	2:tested_positive	1:tested_negative	+	0.973	6	2:tested_positive	1:tested_negative	+ 0.947
7	2:tested_positive	2:tested_positive		1	7	2:tested_positive	2:tested_positive	1

1	1:tested_negative	2:tested_positive	+	1	1	1:tested_negative	1:tested_negative		1
2	1:tested_negative	1:tested_negative		0.967	2	1:tested_negative	1:tested_negative		0.889
3	1:tested_negative	1:tested_negative		1	3	1:tested_negative	2:tested_positive	+	0.923
4	1:tested_negative	1:tested_negative		0.967	4	1:tested_negative	1:tested_negative		0.889
5	1:tested_negative	1:tested_negative		1	5	1:tested_negative	2:tested_positive	+	1
6	2:tested_positive	1:tested_negative	+	1	6	2:tested_positive	2:tested_positive		0.923
7	2:tested_positive	1:tested_negative	+	0.967	7	2:tested_positive	2:tested_positive		0.923

Correctly Classified Instances 55 72.3684 %

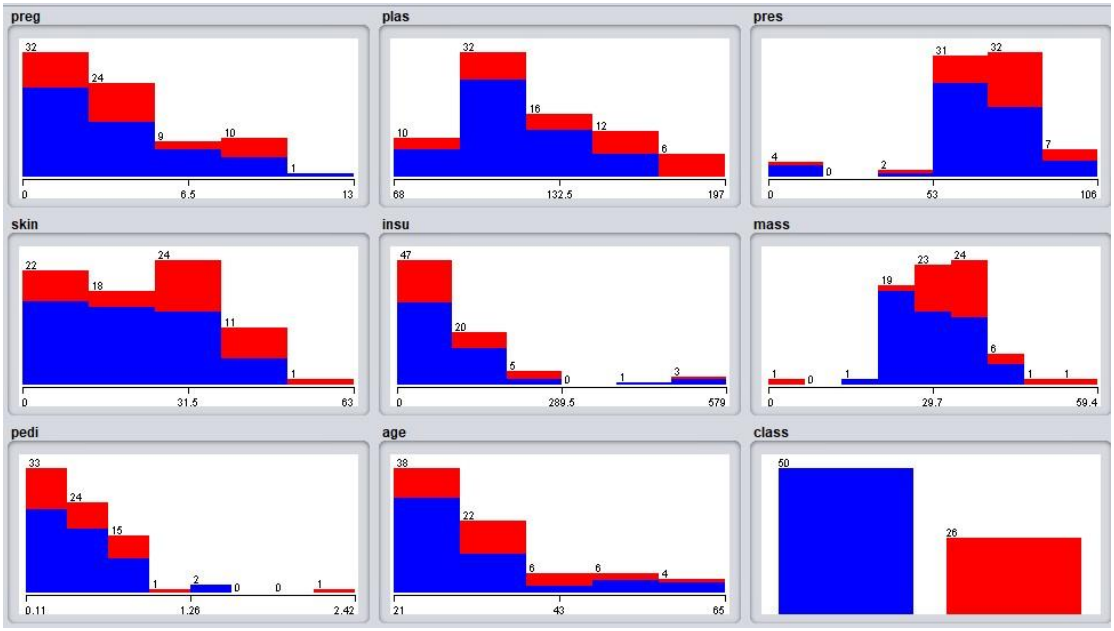
Naïve Bayes:

1	1:tested_negative	1:tested_negative		1	1	1:tested_negative	1:tested_negative		0.995
2	1:tested_negative	1:tested_negative		0.502	2	1:tested_negative	1:tested_negative		0.856
3	1:tested_negative	1:tested_negative		0.988	3	1:tested_negative	1:tested_negative		0.867
4	1:tested_negative	1:tested_negative		0.994	4	1:tested_negative	1:tested_negative		1
5	1:tested_negative	1:tested_negative		1	5	1:tested_negative	1:tested_negative		0.765
6	2:tested_positive	2:tested_positive		0.951	6	2:tested_positive	2:tested_positive		0.972
7	2:tested_positive	1:tested_negative	+	0.587	7	2:tested_positive	1:tested_negative	+	0.862
8	2:tested_positive	1:tested_negative	+	0.994	8	2:tested_positive	1:tested_negative	+	0.816
1	1:tested_negative	1:tested_negative		0.963	1	1:tested_negative	1:tested_negative		0.977
2	1:tested_negative	1:tested_negative		0.997	2	1:tested_negative	2:tested_positive	+	0.999
3	1:tested_negative	1:tested_negative		0.554	3	1:tested_negative	2:tested_positive	+	0.999
4	1:tested_negative	1:tested_negative		0.692	4	1:tested_negative	1:tested_negative		0.984
5	1:tested_negative	1:tested_negative		0.956	5	1:tested_negative	1:tested_negative		0.989
6	2:tested_positive	1:tested_negative	+	0.969	6	2:tested_positive	2:tested_positive		0.999
7	2:tested_positive	2:tested_positive		0.871	7	2:tested_positive	2:tested_positive		0.869
8	2:tested_positive	2:tested_positive		0.902	8	2:tested_positive	2:tested_positive		0.998
1	1:tested_negative	2:tested_positive	+	0.51	1	1:tested_negative	1:tested_negative		0.986
2	1:tested_negative	1:tested_negative		0.972	2	1:tested_negative	1:tested_negative		0.89
3	1:tested_negative	2:tested_positive	+	0.714	3	1:tested_negative	1:tested_negative		0.925
4	1:tested_negative	1:tested_negative		0.98	4	1:tested_negative	1:tested_negative		0.777
5	1:tested_negative	1:tested_negative		0.933	5	1:tested_negative	1:tested_negative		0.95
6	2:tested_positive	2:tested_positive		0.969	6	2:tested_positive	2:tested_positive		0.66
7	2:tested_positive	2:tested_positive		0.517	7	2:tested_positive	2:tested_positive		0.995
8	2:tested_positive	1:tested_negative	+	1	8	2:tested_positive	2:tested_positive		0.805
1	1:tested_negative	1:tested_negative		0.739	1	1:tested_negative	1:tested_negative		0.991
2	1:tested_negative	1:tested_negative		0.92	2	1:tested_negative	1:tested_negative		0.971
3	1:tested_negative	1:tested_negative		0.974	3	1:tested_negative	1:tested_negative		0.993
4	1:tested_negative	1:tested_negative		0.971	4	1:tested_negative	2:tested_positive	+	0.877
5	1:tested_negative	1:tested_negative		0.987	5	1:tested_negative	2:tested_positive	+	0.76
6	2:tested_positive	1:tested_negative	+	0.565	6	2:tested_positive	1:tested_negative	+	0.602
7	2:tested_positive	2:tested_positive		0.974	7	2:tested_positive	2:tested_positive		0.59
1	1:tested_negative	1:tested_negative		0.924	1	1:tested_negative	1:tested_negative		0.993
2	1:tested_negative	2:tested_positive	+	0.889	2	1:tested_negative	1:tested_negative		0.92
3	1:tested_negative	1:tested_negative		0.604	3	1:tested_negative	2:tested_positive	+	0.505
4	1:tested_negative	1:tested_negative		0.929	4	1:tested_negative	1:tested_negative		1
5	1:tested_negative	1:tested_negative		0.83	5	1:tested_negative	1:tested_negative		0.945
6	2:tested_positive	2:tested_positive		0.992	6	2:tested_positive	2:tested_positive		1
7	2:tested_positive	2:tested_positive		0.677	7	2:tested_positive	2:tested_positive		0.964

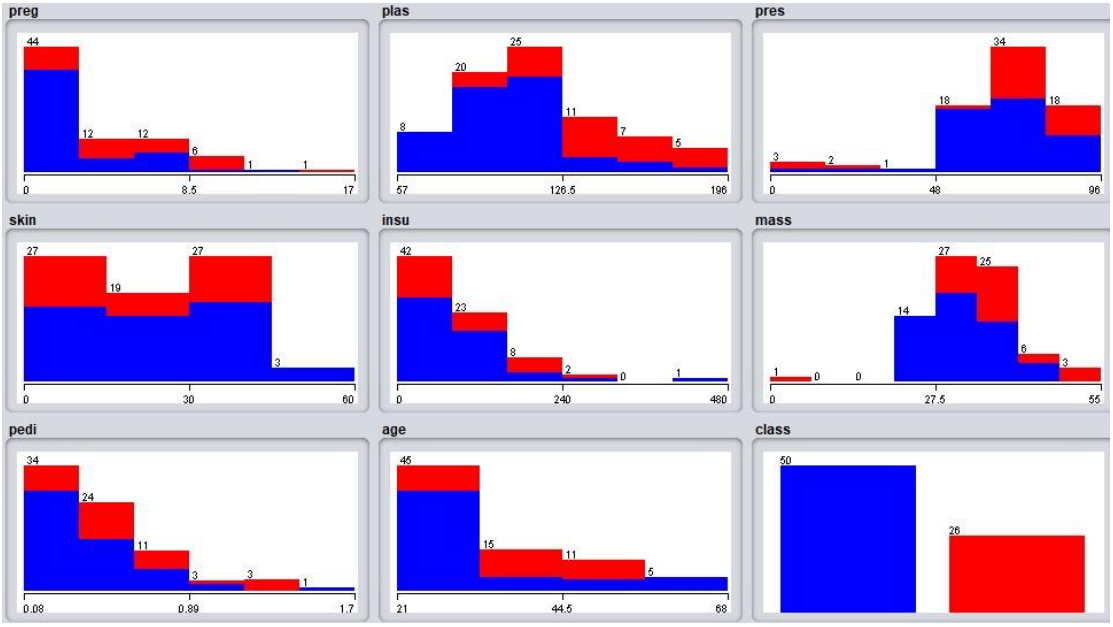
Correctly Classified Instances 60 78.9474 %

(c)重複 10 次不同的取樣，每一次取樣都要改變 randomSeed，使每次 10%的取樣得到不同的 instance (a)~(c) 配分 10%

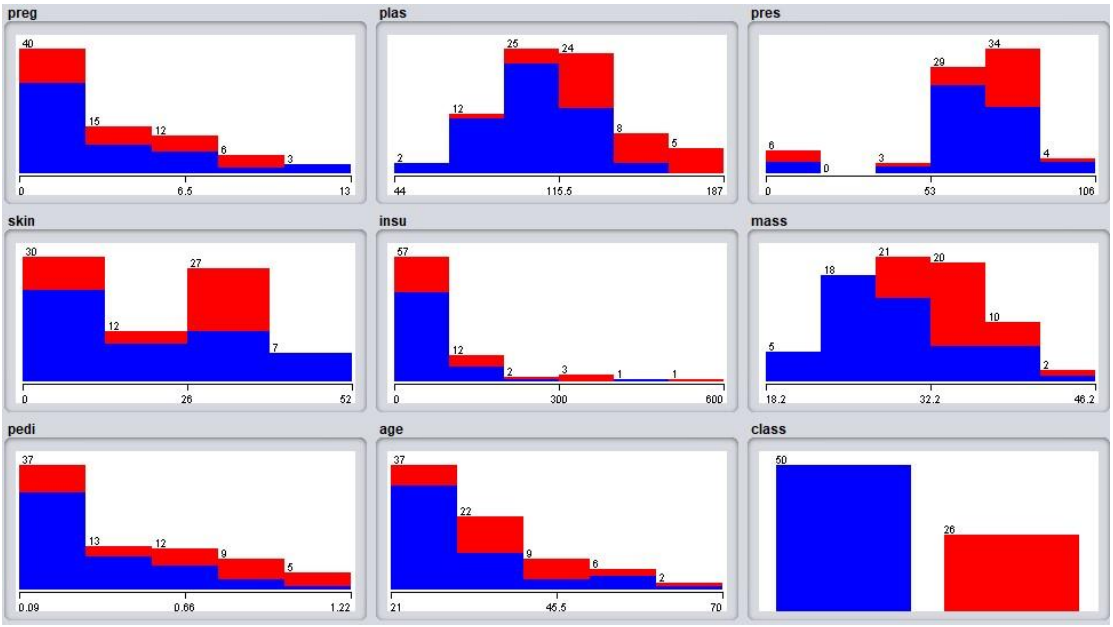
RandomSeed = 10



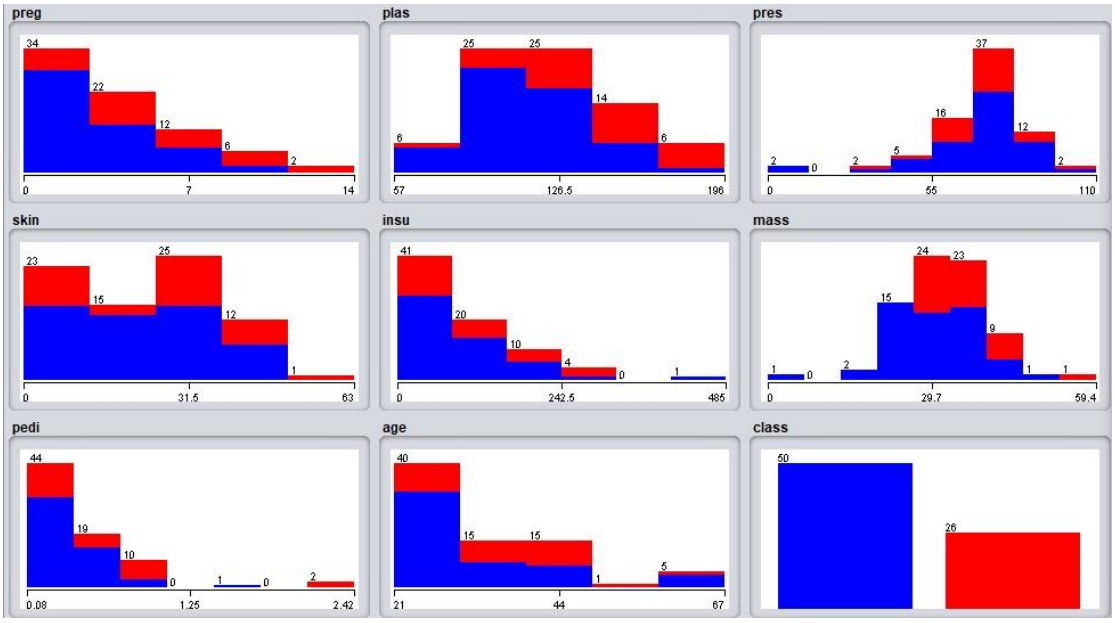
RandomSeed = 20



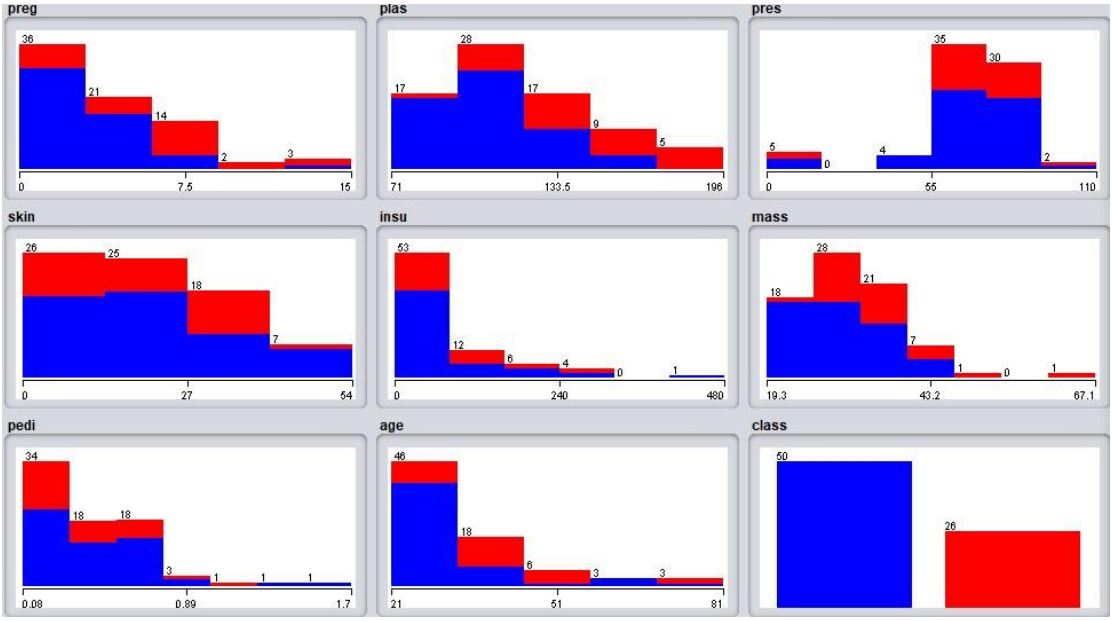
RandomSeed = 30



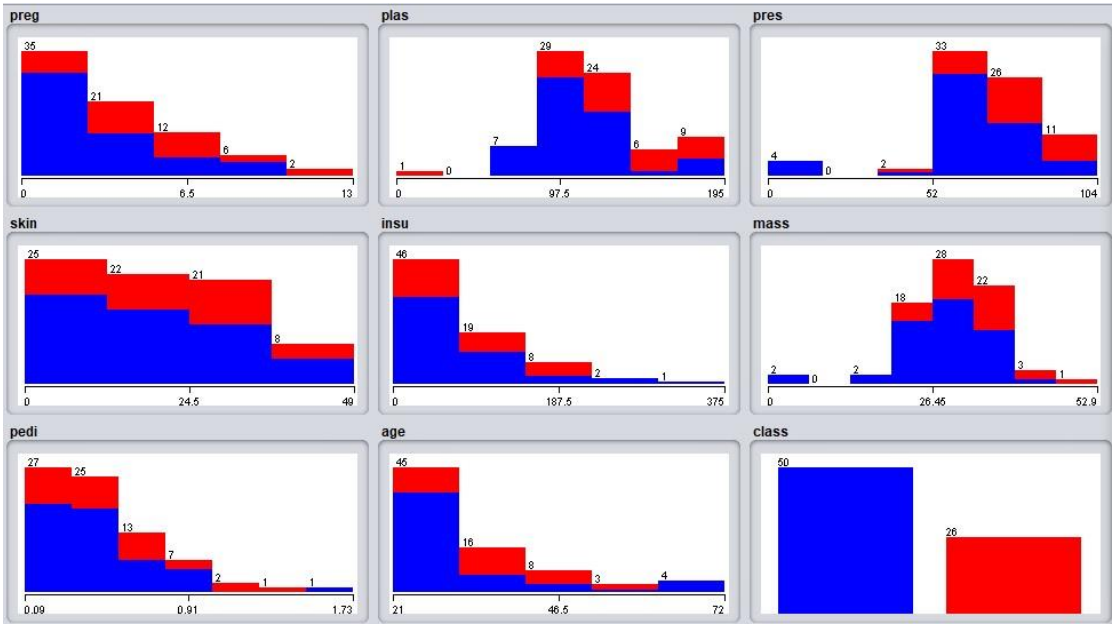
RandomSeed = 40



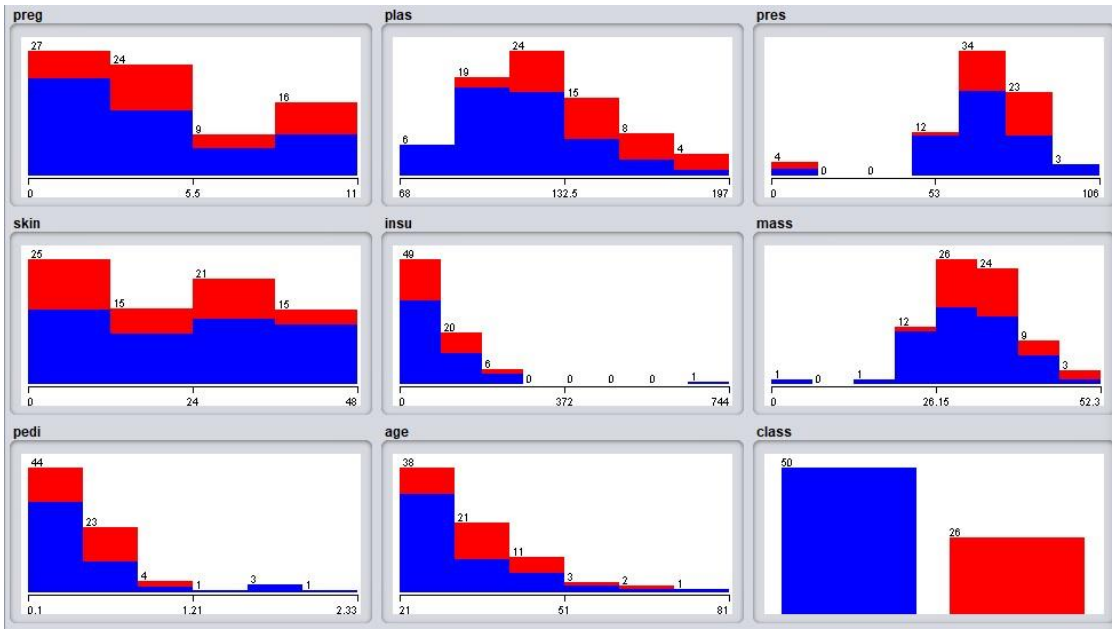
RandomSeed = 50



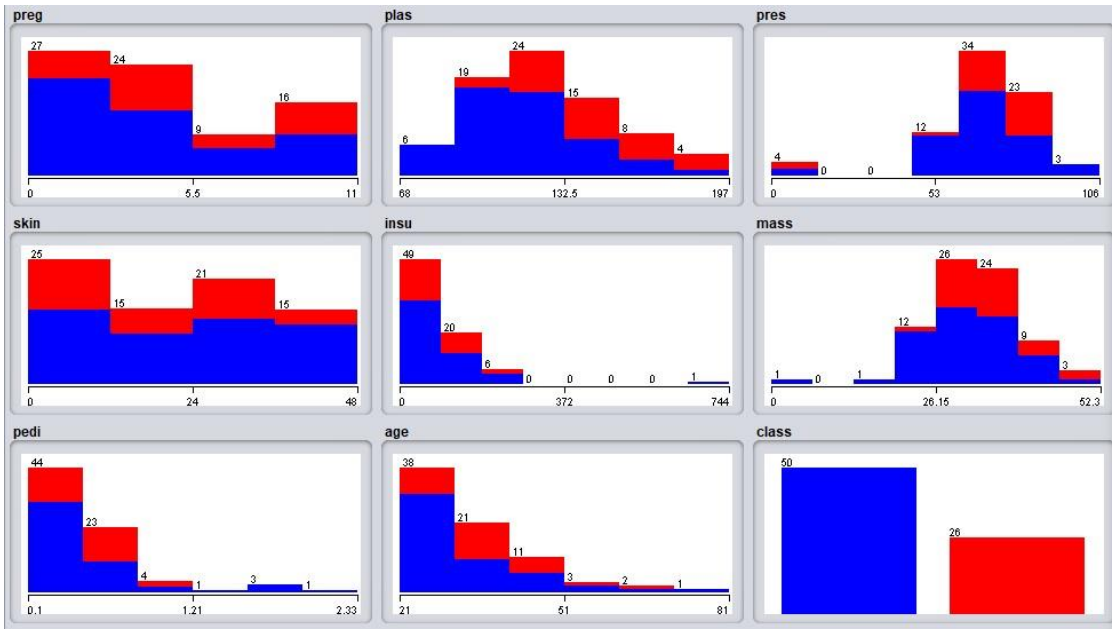
RandomSeed = 60



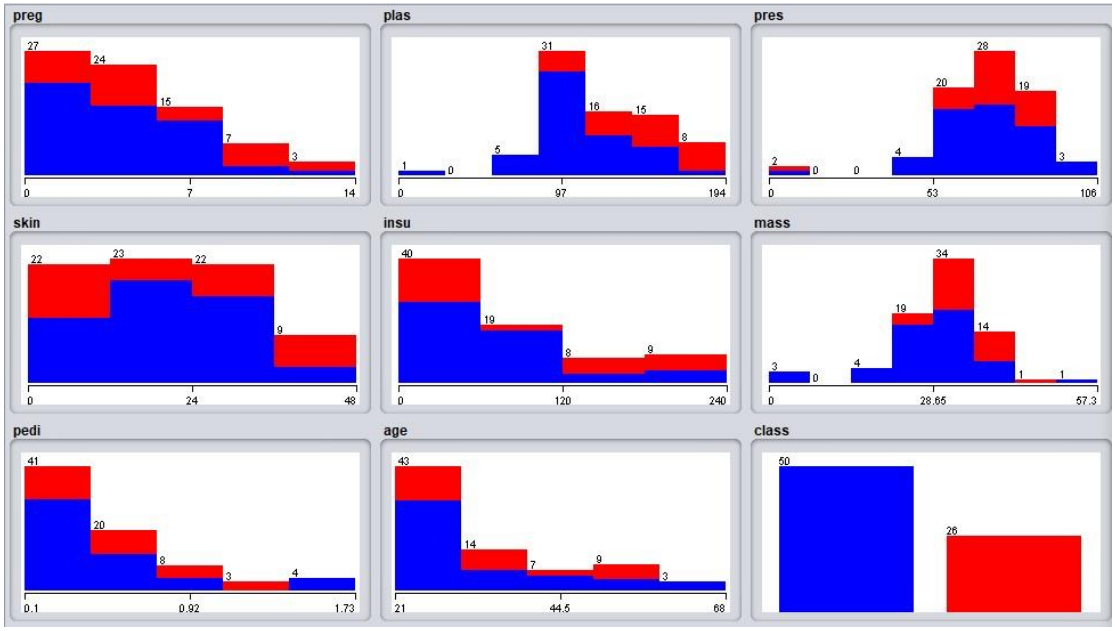
RandomSeed = 70



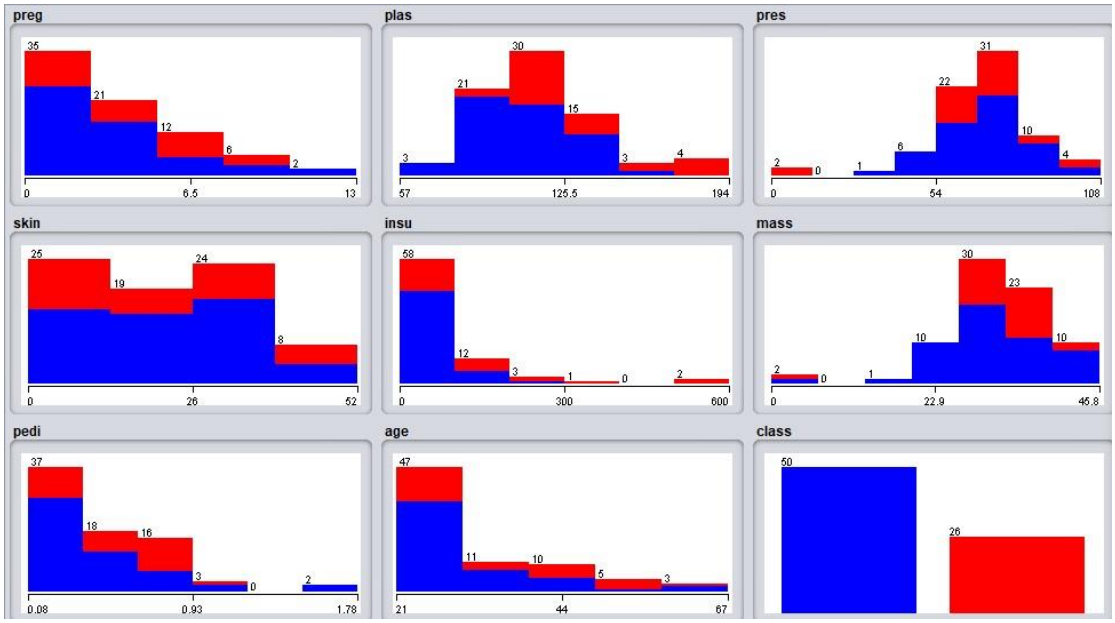
RandomSeed = 80



RandomSeed = 90



RandomSeed = 100



(d)請計算 10 次取樣中，兩分類器的平均正確率、變異數等，假設信心水準 99% 的情況下，兩分類器有無明顯的差別！（請用統計的觀點證明) (20%)

	naïve	J48
	0.710526	0.78947
	0.710526	0.69737
	0.842105	0.81579
	0.736842	0.68421
	0.75	0.75
	0.75	0.75
	0.710526	0.67105
	0.815789	0.76316
	0.657895	0.72368
	0.723684	0.64474
Mean	0.7407893	0.72895
Standard deviation	0.05097719	0.0517
Variation	0.002598674	0.00267

Make an hypothesis that:

H0: d = 0

H1: d != 0

When confidence level at 99%, degree = 10-1=9, z = 3.25

$\frac{d}{\sqrt{\frac{\sigma}{10}}}$<3.25

If-3.25<$\sqrt{\frac{\sigma}{10}}$, there is no difference between them.

σ^2=0.002887/10+0.00297/10=0.0005857

$\frac{740789-0.728947}{\sqrt{\frac{0.0005857}{10}}}$=1.5473 not in the interval (-3.25, 3.25)

Which indicates two classifier have vast difference.