

資料探勘

06160485

曾宏鈞

1. 資料

1-1 來源

[UCI 糖尿病性的視網膜病變資料集](#)

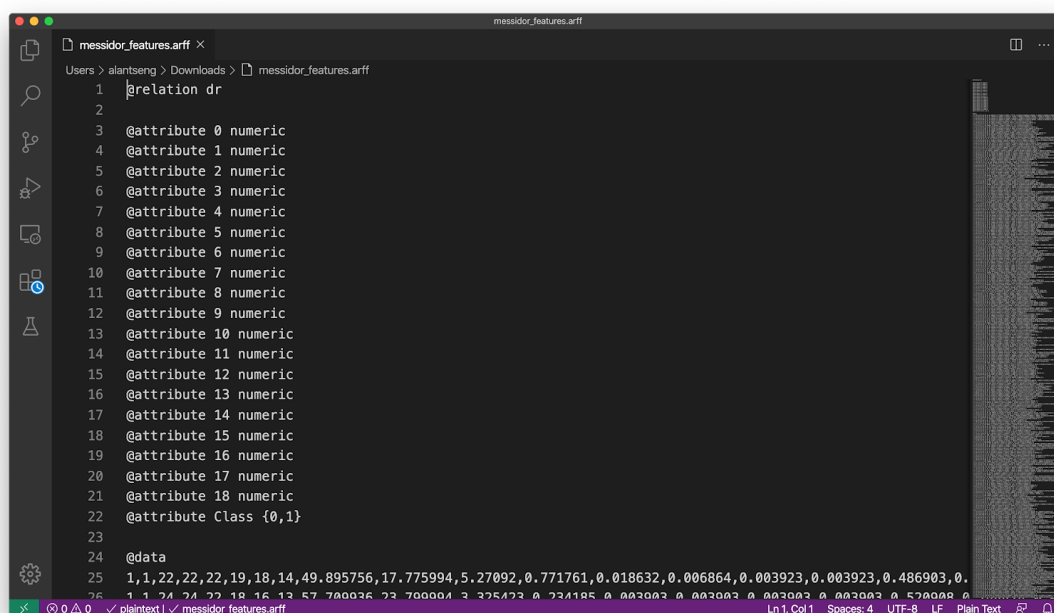
格式：.arff

打開可發現

@relation：代表有一個資料表叫做dr。

@attribute：共有20個屬性，有19個特徵及1個答案，型別為數字與{0,1}的兩個類別。

@data：資料的本身，很像CSV使用逗點隔開。



```
messidor_features.arff
Users > alantseng > Downloads > messidor_features.arff
1  @relation dr
2
3  @attribute 0 numeric
4  @attribute 1 numeric
5  @attribute 2 numeric
6  @attribute 3 numeric
7  @attribute 4 numeric
8  @attribute 5 numeric
9  @attribute 6 numeric
10 @attribute 7 numeric
11 @attribute 8 numeric
12 @attribute 9 numeric
13 @attribute 10 numeric
14 @attribute 11 numeric
15 @attribute 12 numeric
16 @attribute 13 numeric
17 @attribute 14 numeric
18 @attribute 15 numeric
19 @attribute 16 numeric
20 @attribute 17 numeric
21 @attribute 18 numeric
22 @attribute Class {0,1}
23
24 @data
25 1,1,22,22,22,19,18,14,49.895756,17.775994,5.27092,0.771761,0.018632,0.006864,0.003923,0.003923,0.486903,0.
26 1,1,24,24,22,18,16,13,57.700036,23.700004,3.325423,0.734185,0.003003,0.003003,0.003003,0.003003,0.570008,0.1
```

1-2 說明

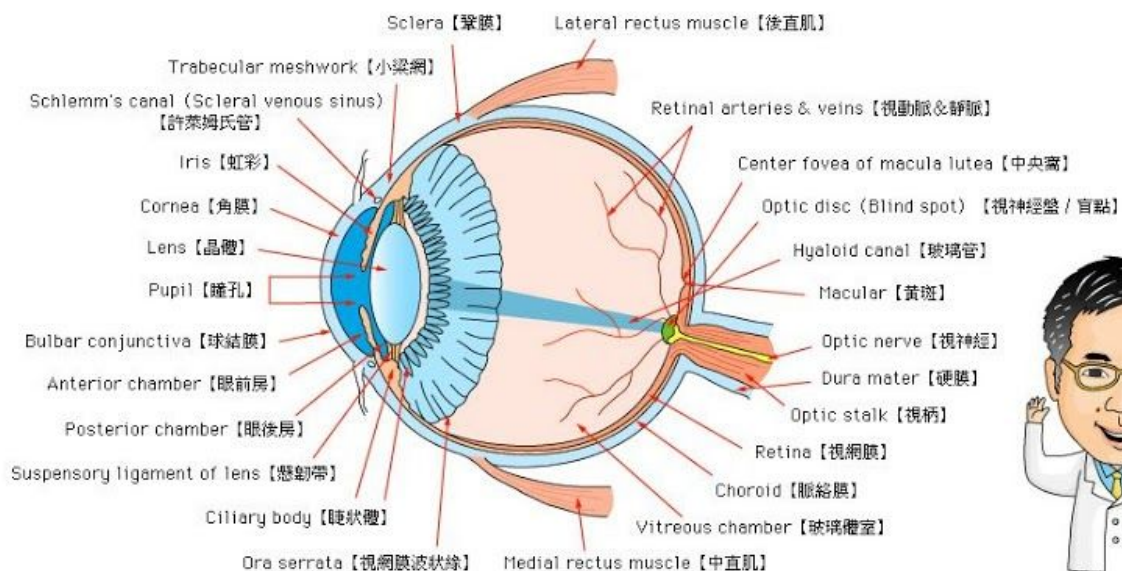
取自Messidor資料集，預測圖像是否包含糖尿病性視網膜病變的徵兆。使用在Balint Antal, Andras Hajdu：基於集合體的糖尿病性視網膜病變自動篩查系統，所有特徵都代表檢測到的病變資訊，且使用圖像分析和特徵提取以及分類。

適用問題	分類
資料筆數	1151
屬性數量	20
缺值	無

1-3 屬性介紹

特徵

0. 檢測質量	0:差, 1:好
1. 預篩選的結果(嚴重視網膜異常)	1:有, 0:沒有
2-7. MA檢測結果，每一項特徵代表在信心水準(alpha)下0.5... 1.0	數字
8-15. 包含2-7的滲出液訊息，由病變數/ROI直徑歸一化特徵，以縮減不同大小的差距	數字
16. 黃斑部到盲點的歐式距離，也使用ROI直徑標準化	數字
17. 盲點直徑	數字
18. 基於AM/FM的分類結果	0:沒有糖尿病視網膜病變(DR), 1,2,3:代表不同的DR標籤



1-4 屬性資訊

我們將原始資料讀入weka並視覺化結果

特徵

0

Name: 0

Missing: 0 (0%)

Distinct: 2

Type: Numeric

Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.997
StdDev	0.059

A bar chart for feature 0. The x-axis is labeled with 0, 0.5, and 1. There is a single bar at value 1, which is split into a blue bottom half and a red top half. The total height of the bar is approximately 1000 units.

1

Name: 1

Missing: 0 (0%)

Distinct: 2

Type: Numeric

Unique: 0 (0%)

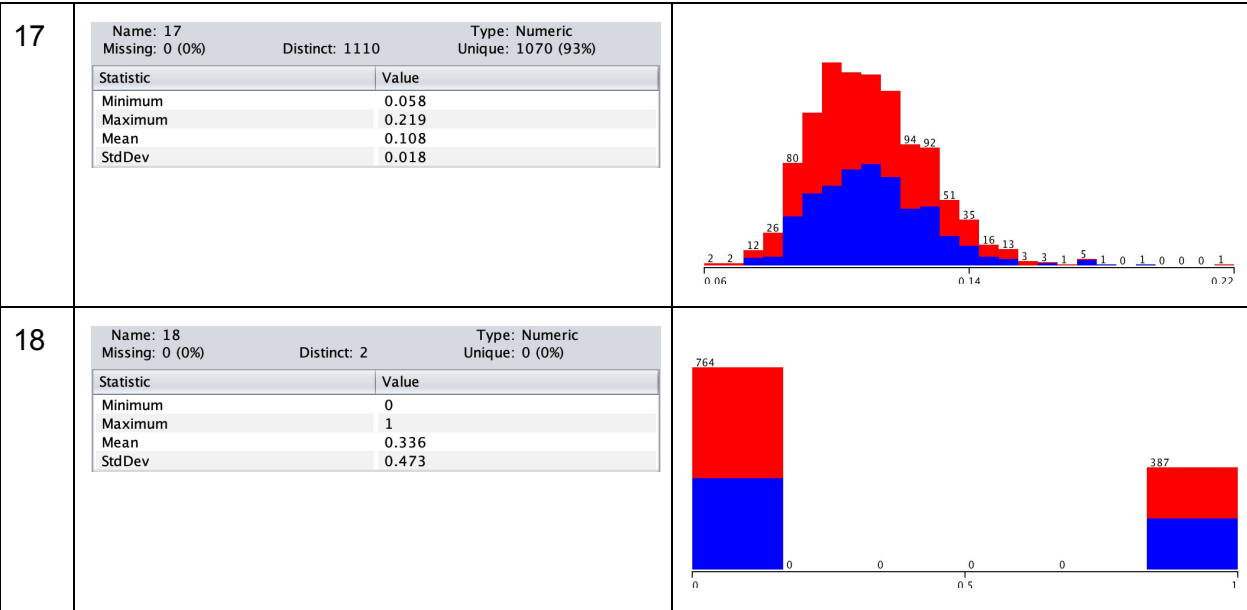
Statistic	Value
Minimum	0
Maximum	1
Mean	0.918
StdDev	0.274

A bar chart for feature 1. The x-axis is labeled with 0, 0.5, and 1. There are two bars: a small bar at value 0 with a height of 94, and a large bar at value 1 with a height of 1057. Both bars are split into a blue bottom half and a red top half.

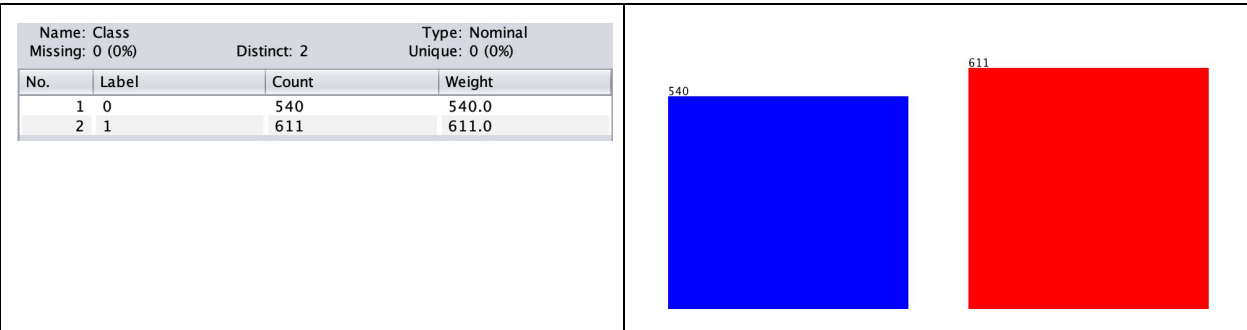
2	<div><div>Name: 2 Missing: 0 (0%)</div><div>Distinct: 110</div><div>Type: Numeric Unique: 11 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>151</td></tr><tr><td>Mean</td><td>38.428</td></tr><tr><td>StdDev</td><td>25.621</td></tr></table>	Statistic	Value	Minimum	1	Maximum	151	Mean	38.428	StdDev	25.621	<p>Frequency distribution for variable 2 (Name: 2, Distinct: 110, Type: Numeric, Unique: 11 (1%)). The distribution is right-skewed, with the highest frequency (199) at the minimum value (1). The x-axis ranges from 1 to 151, and the y-axis represents frequency.</p>
Statistic	Value											
Minimum	1											
Maximum	151											
Mean	38.428											
StdDev	25.621											
3	<div><div>Name: 3 Missing: 0 (0%)</div><div>Distinct: 104</div><div>Type: Numeric Unique: 8 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>132</td></tr><tr><td>Mean</td><td>36.91</td></tr><tr><td>StdDev</td><td>24.106</td></tr></table>	Statistic	Value	Minimum	1	Maximum	132	Mean	36.91	StdDev	24.106	<p>Frequency distribution for variable 3 (Name: 3, Distinct: 104, Type: Numeric, Unique: 8 (1%)). The distribution is right-skewed, with the highest frequency (207) at the minimum value (1). The x-axis ranges from 1 to 132, and the y-axis represents frequency.</p>
Statistic	Value											
Minimum	1											
Maximum	132											
Mean	36.91											
StdDev	24.106											
4	<div><div>Name: 4 Missing: 0 (0%)</div><div>Distinct: 99</div><div>Type: Numeric Unique: 13 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>120</td></tr><tr><td>Mean</td><td>35.141</td></tr><tr><td>StdDev</td><td>22.805</td></tr></table>	Statistic	Value	Minimum	1	Maximum	120	Mean	35.141	StdDev	22.805	<p>Frequency distribution for variable 4 (Name: 4, Distinct: 99, Type: Numeric, Unique: 13 (1%)). The distribution is right-skewed, with the highest frequency (192) at the minimum value (1). The x-axis ranges from 1 to 120, and the y-axis represents frequency.</p>
Statistic	Value											
Minimum	1											
Maximum	120											
Mean	35.141											
StdDev	22.805											
5	<div><div>Name: 5 Missing: 0 (0%)</div><div>Distinct: 91</div><div>Type: Numeric Unique: 9 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>105</td></tr><tr><td>Mean</td><td>32.297</td></tr><tr><td>StdDev</td><td>21.115</td></tr></table>	Statistic	Value	Minimum	1	Maximum	105	Mean	32.297	StdDev	21.115	<p>Frequency distribution for variable 5 (Name: 5, Distinct: 91, Type: Numeric, Unique: 9 (1%)). The distribution is right-skewed, with the highest frequency (204) at the minimum value (1). The x-axis ranges from 1 to 105, and the y-axis represents frequency.</p>
Statistic	Value											
Minimum	1											
Maximum	105											
Mean	32.297											
StdDev	21.115											
6	<div><div>Name: 6 Missing: 0 (0%)</div><div>Distinct: 84</div><div>Type: Numeric Unique: 9 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>97</td></tr><tr><td>Mean</td><td>28.747</td></tr><tr><td>StdDev</td><td>19.509</td></tr></table>	Statistic	Value	Minimum	1	Maximum	97	Mean	28.747	StdDev	19.509	<p>Frequency distribution for variable 6 (Name: 6, Distinct: 84, Type: Numeric, Unique: 9 (1%)). The distribution is right-skewed, with the highest frequency (206) at the minimum value (1). The x-axis ranges from 1 to 97, and the y-axis represents frequency.</p>
Statistic	Value											
Minimum	1											
Maximum	97											
Mean	28.747											
StdDev	19.509											

7	<div><div>Name: 7 Missing: 0 (0%)</div><div>Distinct: 69</div><div>Type: Numeric Unique: 8 (1%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>89</td></tr><tr><td>Mean</td><td>21.151</td></tr><tr><td>StdDev</td><td>15.102</td></tr></table>	Statistic	Value	Minimum	1	Maximum	89	Mean	21.151	StdDev	15.102	
Statistic	Value											
Minimum	1											
Maximum	89											
Mean	21.151											
StdDev	15.102											
8	<div><div>Name: 8 Missing: 0 (0%)</div><div>Distinct: 1141</div><div>Type: Numeric Unique: 1131 (98%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0.349</td></tr><tr><td>Maximum</td><td>403.939</td></tr><tr><td>Mean</td><td>64.097</td></tr><tr><td>StdDev</td><td>58.485</td></tr></table>	Statistic	Value	Minimum	0.349	Maximum	403.939	Mean	64.097	StdDev	58.485	
Statistic	Value											
Minimum	0.349											
Maximum	403.939											
Mean	64.097											
StdDev	58.485											
9	<div><div>Name: 9 Missing: 0 (0%)</div><div>Distinct: 1141</div><div>Type: Numeric Unique: 1131 (98%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>167.131</td></tr><tr><td>Mean</td><td>23.088</td></tr><tr><td>StdDev</td><td>21.603</td></tr></table>	Statistic	Value	Minimum	0	Maximum	167.131	Mean	23.088	StdDev	21.603	
Statistic	Value											
Minimum	0											
Maximum	167.131											
Mean	23.088											
StdDev	21.603											
10	<div><div>Name: 10 Missing: 0 (0%)</div><div>Distinct: 1130</div><div>Type: Numeric Unique: 1119 (97%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>106.07</td></tr><tr><td>Mean</td><td>8.705</td></tr><tr><td>StdDev</td><td>11.568</td></tr></table>	Statistic	Value	Minimum	0	Maximum	106.07	Mean	8.705	StdDev	11.568	
Statistic	Value											
Minimum	0											
Maximum	106.07											
Mean	8.705											
StdDev	11.568											
11	<div><div>Name: 11 Missing: 0 (0%)</div><div>Distinct: 1032</div><div>Type: Numeric Unique: 1022 (89%)</div></div> <table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>59.766</td></tr><tr><td>Mean</td><td>1.836</td></tr><tr><td>StdDev</td><td>3.923</td></tr></table>	Statistic	Value	Minimum	0	Maximum	59.766	Mean	1.836	StdDev	3.923	
Statistic	Value											
Minimum	0											
Maximum	59.766											
Mean	1.836											
StdDev	3.923											

12	<div><div><div>Name: 12 Missing: 0 (0%)</div><div>Distinct: 795</div><div>Type: Numeric Unique: 771 (67%)</div></div><table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>51.423</td></tr><tr><td>Mean</td><td>0.561</td></tr><tr><td>StdDev</td><td>2.484</td></tr></table></div> <div></div>	Statistic	Value	Minimum	0	Maximum	51.423	Mean	0.561	StdDev	2.484
Statistic	Value										
Minimum	0										
Maximum	51.423										
Mean	0.561										
StdDev	2.484										
13	<div><div><div>Name: 13 Missing: 0 (0%)</div><div>Distinct: 579</div><div>Type: Numeric Unique: 547 (48%)</div></div><table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>20.099</td></tr><tr><td>Mean</td><td>0.212</td></tr><tr><td>StdDev</td><td>1.057</td></tr></table></div> <div></div>	Statistic	Value	Minimum	0	Maximum	20.099	Mean	0.212	StdDev	1.057
Statistic	Value										
Minimum	0										
Maximum	20.099										
Mean	0.212										
StdDev	1.057										
14	<div><div><div>Name: 14 Missing: 0 (0%)</div><div>Distinct: 415</div><div>Type: Numeric Unique: 385 (33%)</div></div><table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>5.938</td></tr><tr><td>Mean</td><td>0.086</td></tr><tr><td>StdDev</td><td>0.399</td></tr></table></div> <div></div>	Statistic	Value	Minimum	0	Maximum	5.938	Mean	0.086	StdDev	0.399
Statistic	Value										
Minimum	0										
Maximum	5.938										
Mean	0.086										
StdDev	0.399										
15	<div><div><div>Name: 15 Missing: 0 (0%)</div><div>Distinct: 351</div><div>Type: Numeric Unique: 325 (28%)</div></div><table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>3.087</td></tr><tr><td>Mean</td><td>0.037</td></tr><tr><td>StdDev</td><td>0.179</td></tr></table></div> <div></div>	Statistic	Value	Minimum	0	Maximum	3.087	Mean	0.037	StdDev	0.179
Statistic	Value										
Minimum	0										
Maximum	3.087										
Mean	0.037										
StdDev	0.179										
16	<div><div><div>Name: 16 Missing: 0 (0%)</div><div>Distinct: 1132</div><div>Type: Numeric Unique: 1113 (97%)</div></div><table><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0.368</td></tr><tr><td>Maximum</td><td>0.592</td></tr><tr><td>Mean</td><td>0.523</td></tr><tr><td>StdDev</td><td>0.028</td></tr></table></div> <div></div>	Statistic	Value	Minimum	0.368	Maximum	0.592	Mean	0.523	StdDev	0.028
Statistic	Value										
Minimum	0.368										
Maximum	0.592										
Mean	0.523										
StdDev	0.028										



答案

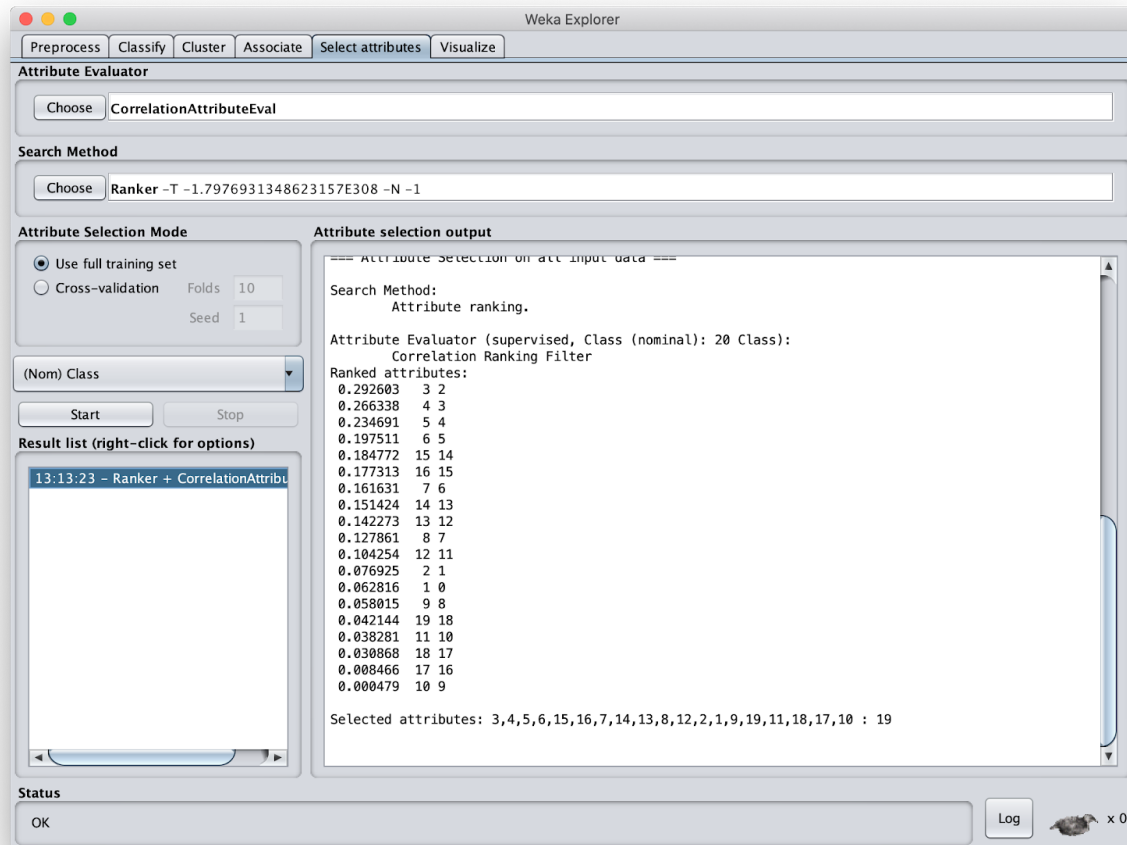


由於兩類別資料數目都差不多，因此不需針對答案去做抽樣

1-5 特徵工程

1-5-1 基於關聯性

- Weka操作步驟：
 - Select attributes
 - Attribute Evaluator > CorrelationAttributeEval
 - Search Method > Ranker



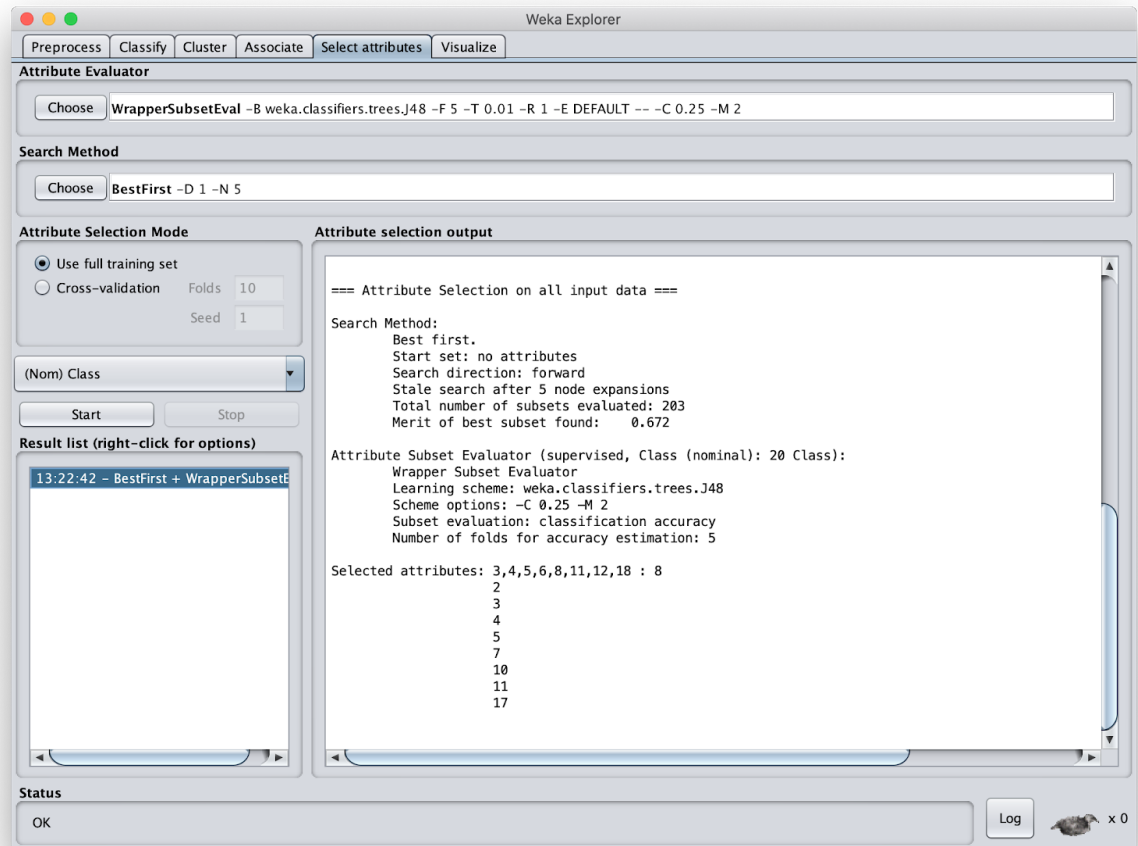
1-5-2 基於學習

- Weka操作步驟：

Select attributes

Attribute Evaluator > WrapperSubsetEval > classifier > tree > J48

Search Method > BestFirst



1-6 前處理原因

- 觀察資料值域

由1-4可知，特徵19個標籤最大值從0.219-403不等，因此資料前處理要做正規化的動作，避免訓練模型時，因特徵值域範圍不同造成某項特徵重要性權重的錯誤。

- 抽樣

我們將原始資料進行抽樣訓練，使得訓練時使用的時間及空間會較少。

3. 降低維度

一般資料都有大量的特徵，我們使用主成分分析法(PCA)而不是單純的特徵選擇來選擇對模型比較有幫助的特徵，把資料集的複雜度變小有以下的優點：

- 避免維度詛咒
- 較好視覺化
- 消除資料本身雜訊
- 更容易解釋

2. 資料前處理

2-1 資料正規化

- Weka操作步驟：
Preprocess>Filter
filters>unsupervised>attribute>Normalize

before	after																																
<table><tr><td colspan="2">Name: 8</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1141</td><td>Unique: 1131 (98%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0.349</td></tr><tr><td>Maximum</td><td>403.939</td></tr><tr><td>Mean</td><td>64.097</td></tr><tr><td>StdDev</td><td>58.485</td></tr></table>	Name: 8		Type: Numeric	Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)	Statistic	Value	Minimum	0.349	Maximum	403.939	Mean	64.097	StdDev	58.485	<table><tr><td colspan="2">Name: 8</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1141</td><td>Unique: 1131 (98%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>1</td></tr><tr><td>Mean</td><td>0.158</td></tr><tr><td>StdDev</td><td>0.145</td></tr></table>	Name: 8		Type: Numeric	Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)	Statistic	Value	Minimum	0	Maximum	1	Mean	0.158	StdDev	0.145
Name: 8		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)																															
Statistic	Value																																
Minimum	0.349																																
Maximum	403.939																																
Mean	64.097																																
StdDev	58.485																																
Name: 8		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)																															
Statistic	Value																																
Minimum	0																																
Maximum	1																																
Mean	0.158																																
StdDev	0.145																																
<table><tr><td colspan="2">Name: 9</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1141</td><td>Unique: 1131 (98%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>167.131</td></tr><tr><td>Mean</td><td>23.088</td></tr><tr><td>StdDev</td><td>21.603</td></tr></table>	Name: 9		Type: Numeric	Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)	Statistic	Value	Minimum	0	Maximum	167.131	Mean	23.088	StdDev	21.603	<table><tr><td colspan="2">Name: 9</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1141</td><td>Unique: 1131 (98%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>1</td></tr><tr><td>Mean</td><td>0.138</td></tr><tr><td>StdDev</td><td>0.129</td></tr></table>	Name: 9		Type: Numeric	Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)	Statistic	Value	Minimum	0	Maximum	1	Mean	0.138	StdDev	0.129
Name: 9		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)																															
Statistic	Value																																
Minimum	0																																
Maximum	167.131																																
Mean	23.088																																
StdDev	21.603																																
Name: 9		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1141	Unique: 1131 (98%)																															
Statistic	Value																																
Minimum	0																																
Maximum	1																																
Mean	0.138																																
StdDev	0.129																																
<table><tr><td colspan="2">Name: 17</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1110</td><td>Unique: 1070 (93%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0.058</td></tr><tr><td>Maximum</td><td>0.219</td></tr><tr><td>Mean</td><td>0.108</td></tr><tr><td>StdDev</td><td>0.018</td></tr></table>	Name: 17		Type: Numeric	Missing: 0 (0%)	Distinct: 1110	Unique: 1070 (93%)	Statistic	Value	Minimum	0.058	Maximum	0.219	Mean	0.108	StdDev	0.018	<table><tr><td colspan="2">Name: 17</td><td>Type: Numeric</td></tr><tr><td>Missing: 0 (0%)</td><td>Distinct: 1110</td><td>Unique: 1070 (93%)</td></tr><tr><td>Statistic</td><td>Value</td></tr><tr><td>Minimum</td><td>0</td></tr><tr><td>Maximum</td><td>1</td></tr><tr><td>Mean</td><td>0.313</td></tr><tr><td>StdDev</td><td>0.111</td></tr></table>	Name: 17		Type: Numeric	Missing: 0 (0%)	Distinct: 1110	Unique: 1070 (93%)	Statistic	Value	Minimum	0	Maximum	1	Mean	0.313	StdDev	0.111
Name: 17		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1110	Unique: 1070 (93%)																															
Statistic	Value																																
Minimum	0.058																																
Maximum	0.219																																
Mean	0.108																																
StdDev	0.018																																
Name: 17		Type: Numeric																															
Missing: 0 (0%)	Distinct: 1110	Unique: 1070 (93%)																															
Statistic	Value																																
Minimum	0																																
Maximum	1																																
Mean	0.313																																
StdDev	0.111																																

- 說明：
由上表可知，原始資料中特徵17的值域是 [0.058, 0.219]、特徵8的的值域是[0.349, 403.939]，標準差為0.018及58.485，平均為0.108、64.097，兩者相差640倍，有可能會因值域造成重要性的錯誤，因此我們將其正規化，使特徵的值域範圍在[0,1]之間，每一個特徵的重要性都相同。

2-2 資料抽樣

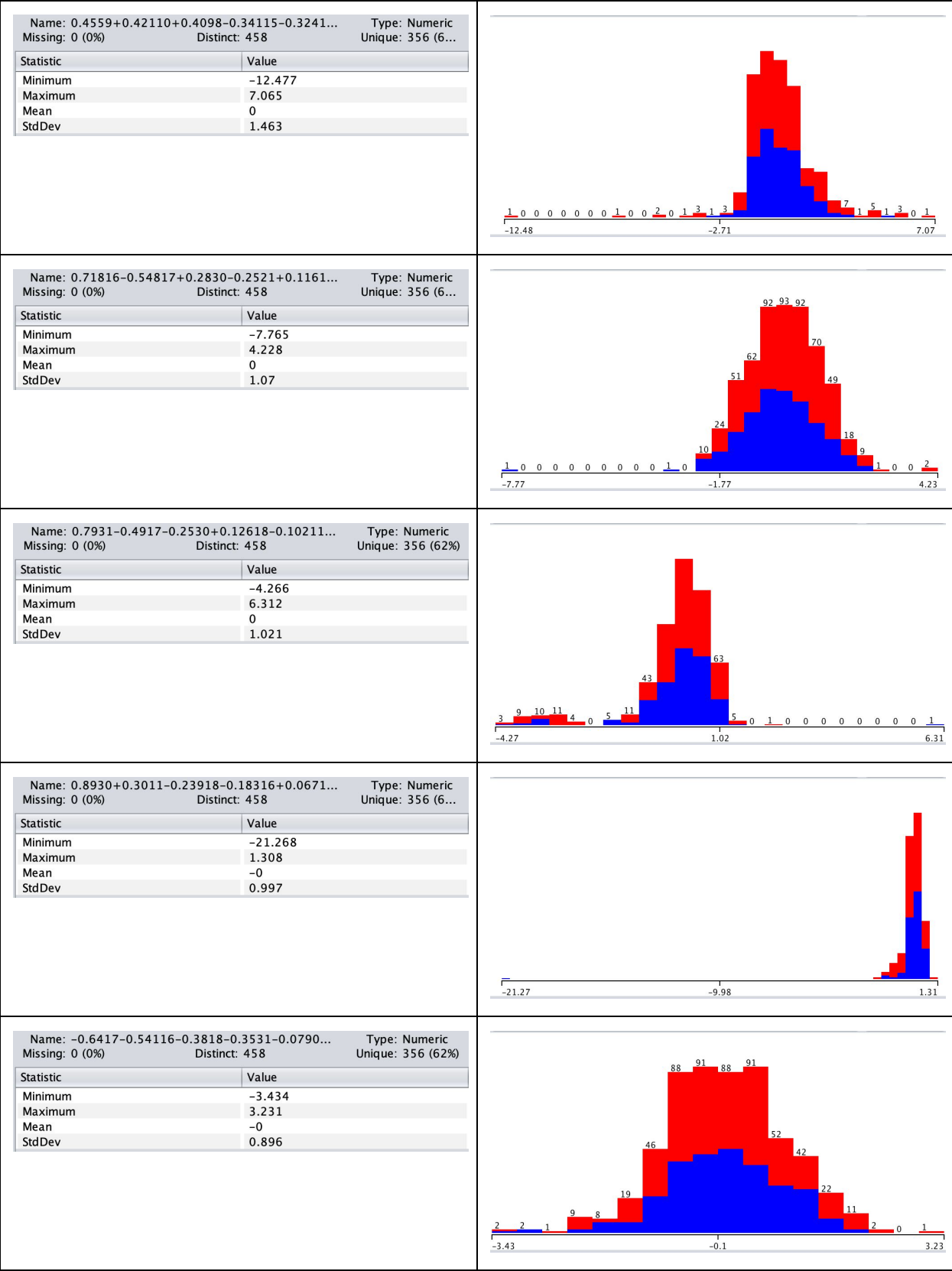
- Weka操作步驟：
Preprocess>Filter
filters>unsupervised>instance>resample
sampleSizePresentage:50

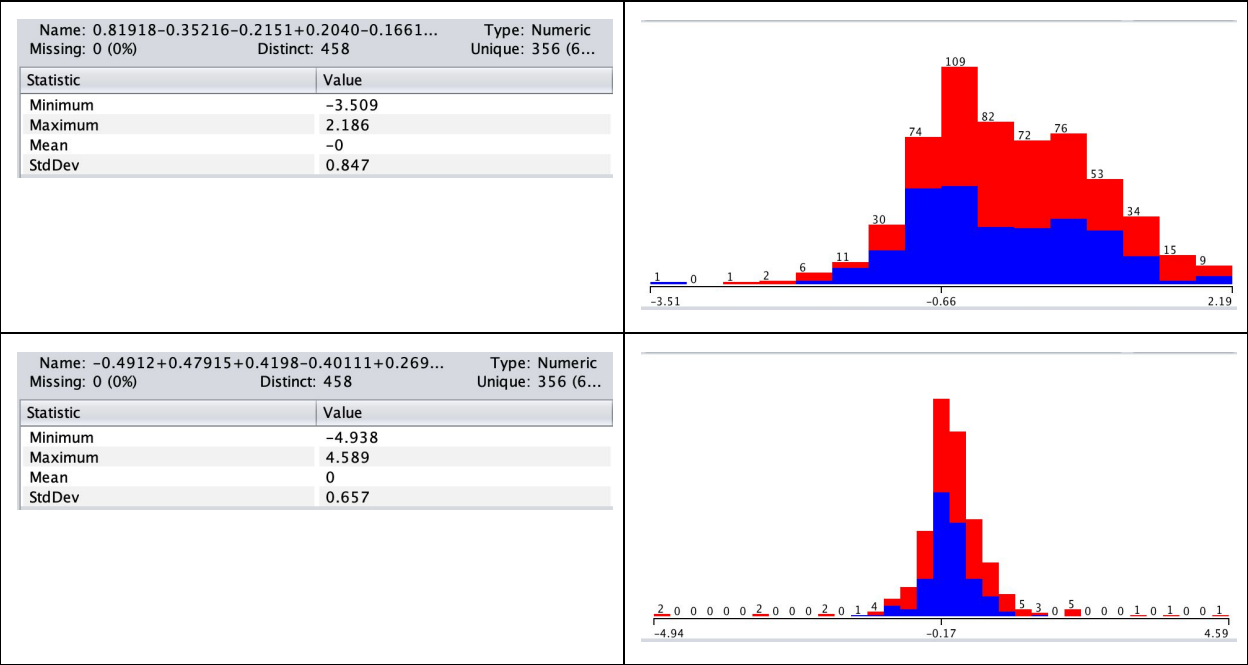
Current relation	Current relation
Relation: dr Instances: 1151 Attributes: 20 Sum of weights: 1151	Relation: dr-weka.filters.unsupervised.instance.Resa... Instances: 575 Attributes: 20 Sum of weights: 575

2-3 主成分分析法(PCA)

- Weka操作步驟：
Preprocess>Filter
filters>unsupervised>attribute>PrincipleComponent

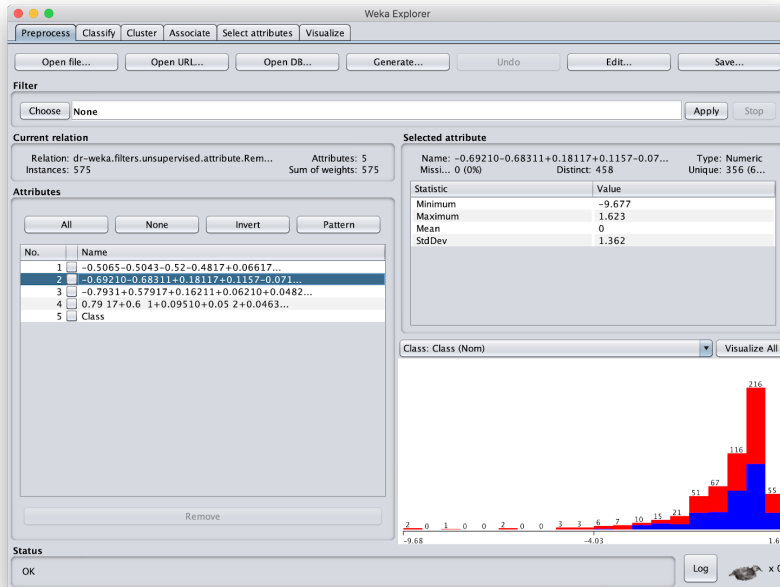
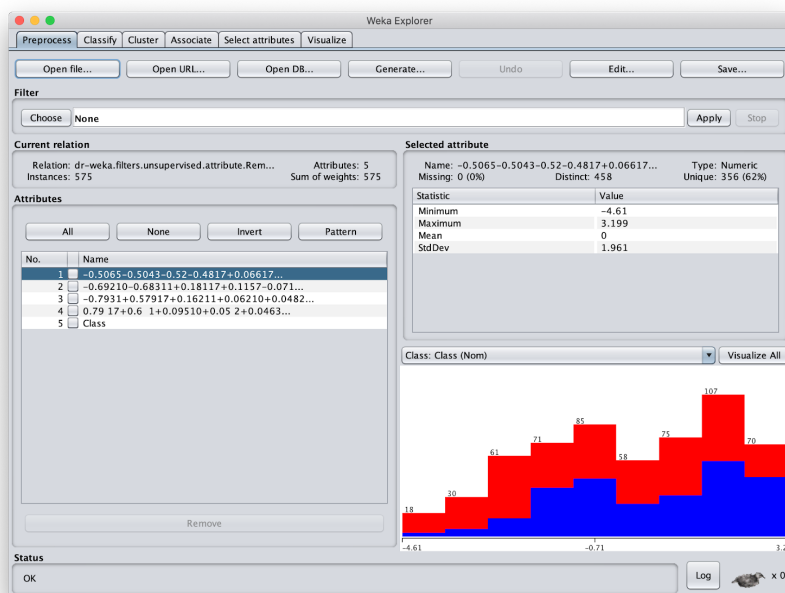
Name: -0.3944-0.3945-0.3913-0.3886-0.3882... Missing: 0 (0%) Distinct: 458 Type: Numeric Unique: 356 (62%)											
<table> <tr> <th>Statistic</th><th>Value</th></tr> <tr> <td>Minimum</td><td>-7.18</td></tr> <tr> <td>Maximum</td><td>4.474</td></tr> <tr> <td>Mean</td><td>-0</td></tr> <tr> <td>StdDev</td><td>2.493</td></tr> </table>	Statistic	Value	Minimum	-7.18	Maximum	4.474	Mean	-0	StdDev	2.493	
Statistic	Value										
Minimum	-7.18										
Maximum	4.474										
Mean	-0										
StdDev	2.493										
Name: -0.41111-0.3913-0.38312-0.36614-0.333... Missing: 0 (0%) Distinct: 458 Type: Numeric Unique: 356 (6...)											
<table> <tr> <th>Statistic</th><th>Value</th></tr> <tr> <td>Minimum</td><td>-20.651</td></tr> <tr> <td>Maximum</td><td>1.846</td></tr> <tr> <td>Mean</td><td>-0</td></tr> <tr> <td>StdDev</td><td>2.212</td></tr> </table>	Statistic	Value	Minimum	-20.651	Maximum	1.846	Mean	-0	StdDev	2.212	
Statistic	Value										
Minimum	-20.651										
Maximum	1.846										
Mean	-0										
StdDev	2.212										





以上可知， 我們使用PCA將19個特徵縮減到只有9個特徵(9個維度)

2-4 特徵選擇且資料處理過

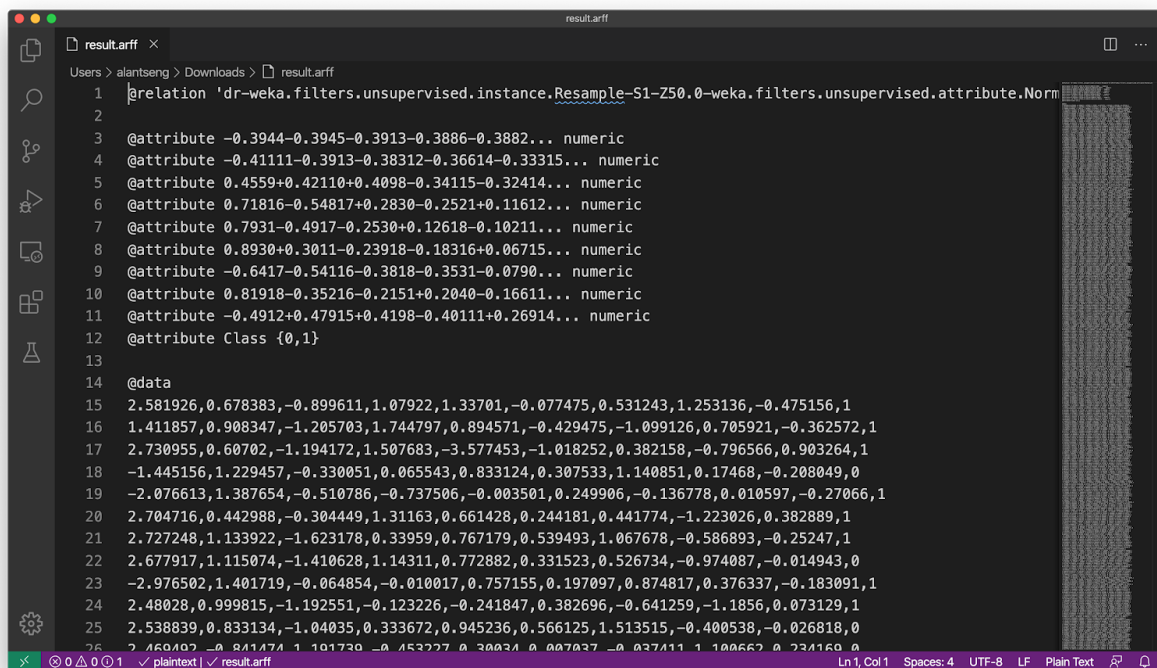


3. 心得

在此次的期中作業中，我們使用了weka學習如何對資料做前處理。從此次練習中可以觀察到，weka是一個很簡單可以快速對資料做處理的資料探勘工具，對於現在正在學習資料探勘技術是非常有幫助的，除了上課學習的理論基礎，更有實作的配合。此外，以前在線性代數老師有提到的主成分分析法(PCA)對於當時的我覺得很抽象，但在做期中專案的時候就了解原來這個是多麼的重要，可以避免一些訓練上的問題，如維度詛咒等等。

綜合上述，我覺得此次期中作業練習除了了解整個做資料探勘的步驟，更對於課堂上的理論及更早線性代數的資料處理演算法有更一步的了解，透過實作能更了解每一個步驟的意義，因此此次作業對於資料探勘是非常重要且有意義的。

3-1 處理完的資料樣式



```
1 relation 'dr-weka.filters.unsupervised.instance.Resample-S1-Z50.0-weka.filters.unsupervised.attribute.Norm
2
3 @attribute -0.3944-0.3945-0.3913-0.3886-0.3882... numeric
4 @attribute -0.41111-0.3913-0.38312-0.36614-0.33315... numeric
5 @attribute 0.4559+0.42110+0.4098-0.34115-0.32414... numeric
6 @attribute 0.71816-0.54817+0.2830-0.2521+0.11612... numeric
7 @attribute 0.7931-0.4917-0.2530+0.12618-0.10211... numeric
8 @attribute 0.8930+0.3011-0.23918-0.18316+0.06715... numeric
9 @attribute -0.6417-0.54116-0.3818-0.3531-0.0790... numeric
10 @attribute 0.81918-0.35216-0.2151+0.2040-0.16611... numeric
11 @attribute -0.4912+0.47915+0.4198-0.40111+0.26914... numeric
12 @attribute Class {0,1}
13
14 @data
15 2.581926,0.678383,-0.899611,1.07922,1.33701,-0.077475,0.531243,1.253136,-0.475156,1
16 1.411857,0.908347,-1.205703,1.744797,0.894571,-0.429475,-1.099126,0.705921,-0.362572,1
17 2.730955,0.60702,-1.194172,1.507683,-3.577453,-1.018252,0.382158,-0.796566,0.903264,1
18 -1.445156,1.229457,-0.330051,0.065543,0.833124,0.307533,1.140851,0.17468,-0.208049,0
19 -2.076613,1.387654,-0.510786,-0.737506,-0.003501,0.249906,-0.136778,0.010597,-0.27066,1
20 2.704716,0.442988,-0.304449,1.31163,0.661428,0.244181,0.441774,-1.223026,0.382889,1
21 2.727248,1.133922,-1.623178,0.33959,0.767179,0.539493,1.067678,-0.586893,-0.25247,1
22 2.677917,1.115074,-1.410628,1.14311,0.772882,0.331523,0.526734,-0.974087,-0.014943,0
23 -2.976502,1.401719,-0.064854,-0.010017,0.757155,0.197097,0.874817,0.376337,-0.183091,1
24 2.48028,0.999815,-1.192551,-0.123226,-0.241847,0.382696,-0.641259,-1.1856,0.073129,1
25 2.538839,0.833134,-1.04035,0.333672,0.945236,0.566125,1.513515,-0.400538,-0.026818,0
26 2.460402,-0.841474,1.101730,-0.453227,0.30034,0.007037,-0.037411,1.100662,0.224160,0
```