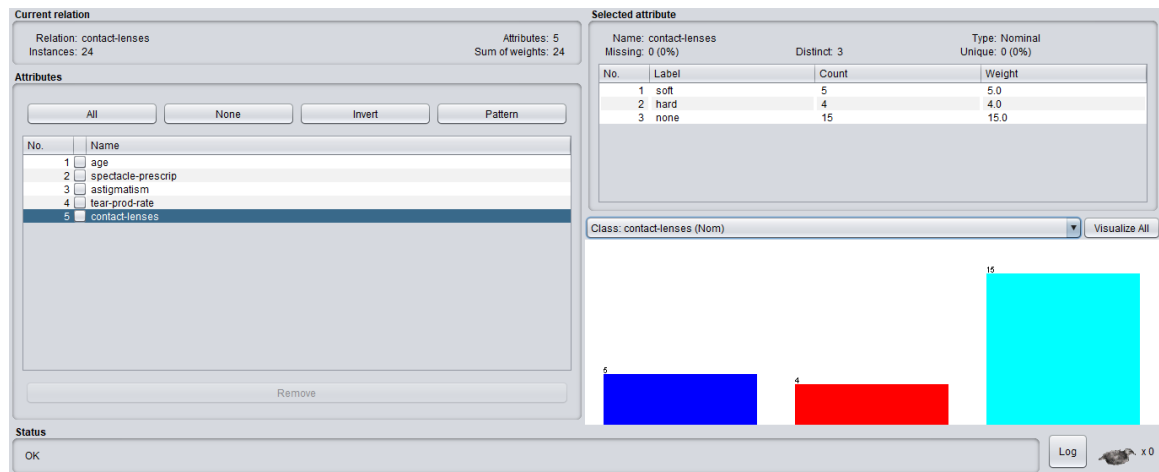


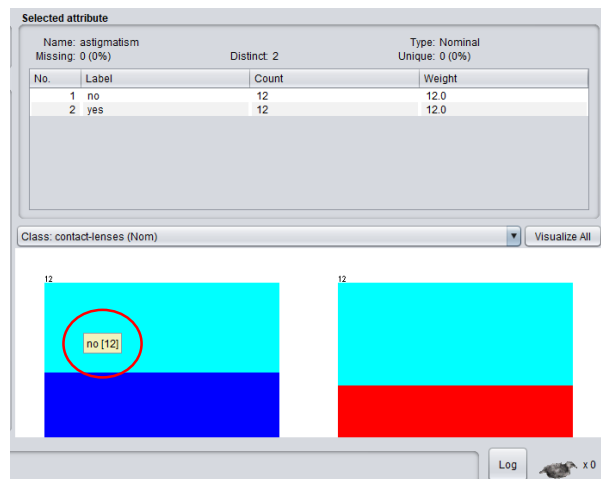
1. 用 Weka 軟體對 contact-lenses.arff 建立 J48 決策樹，選擇 “Use training set”，設定 Attribute: contact-lenses 為 Output，在過程中對重要步驟截圖並加以說明，並回答以下問題：



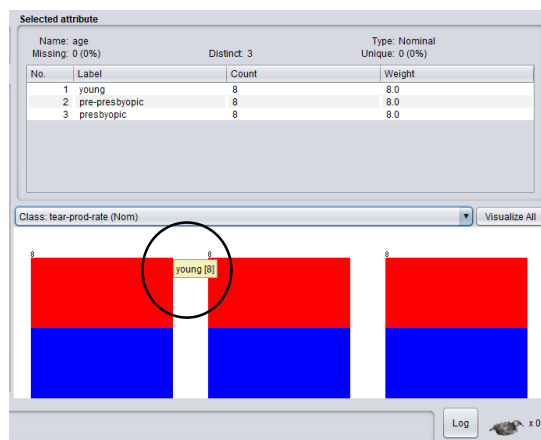
(a) 在前處理部分，右下角選擇不同屬性作為 Class，請解釋長條圖中的數字以及不同顏色意義為何？並說明當選擇不同屬性作為 Class 的長條圖有何不同？ (15%)

在做完前處理後，便會如上圖所示。此為設定 contact-lenses 這個 attribute 為 class attribute 時，它在 contact-lenses 所呈現的分佈。很明顯地，contact-lenses 的 data 分成三種，當游標置於長條圖上時，深藍色為 soft(5)，紅色為 hard(4)，水藍色為 none(15)。長條圖上頭的數字為這個 attribute 中擁有相同 value 的個數及顯示其 value 名稱。

舉例來說：右圖 output 值設定一樣為 contact-lenses，selected attribute 選擇 astigmatism，底下的長條圖所呈現的就是 contact-lenses 的 value 在 astigmatism 的結果下所呈現的分佈。左手邊是 12 筆 no，當中 contact-lenses 的 value 為 none 和 soft 所呈現的比例為圖示，右邊則為 contact-lenses 在 astigmatism = yes 的情況下所呈現的 value 分佈。



那如果選擇不同屬性作為 class 的長條圖就視每一個 attribute 而有不同。左圖為 class attribute : tear-prod-rate 在 age attribute 的三種 value: young, pre-presbyopic, presbyopic 的情況下所呈現的分佈。藍色為 tear-prod-rate = reduced，紅色則為 tear-prod-rate = normal。



(b) 請計算 Output Class=soft 其 F-Measure 為多少？請利用 Confusion matrix 解釋。 (10%)

TP: 預測 soft，實際 soft = 5。

FP: 預測 soft，實際非 soft = 0+1。

FN: 預測非 soft，實際非 soft = 0+0。

Precision Rate = $TP/(TP+FP) = 5 / (5+0+1) = 5/6$

Recall Rate = $TP/(TP+FN) = 5 / (5+0+0) = 1$

F-measure = $(2*5/6*1) / (5/6+1) = 10/11 = 0.90909091$ 此值與圖中相等。

```
=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure
               1.000   0.053   0.833   1.000   0.909
               0.750   0.000   1.000   0.750   0.857
               0.933   0.111   0.933   0.933   0.933
Weighted Avg.  0.917   0.080   0.924   0.917   0.916

=== Confusion Matrix ===
  a  b  c  <-- classified as
  5  0  0  | a = soft
  0  3  1  | b = hard
  1  0 14  | c = none
```

(c) 使用 Visualize Tree 或 Classifier Output 列出三個 Classification Rule 並解釋。(15%)

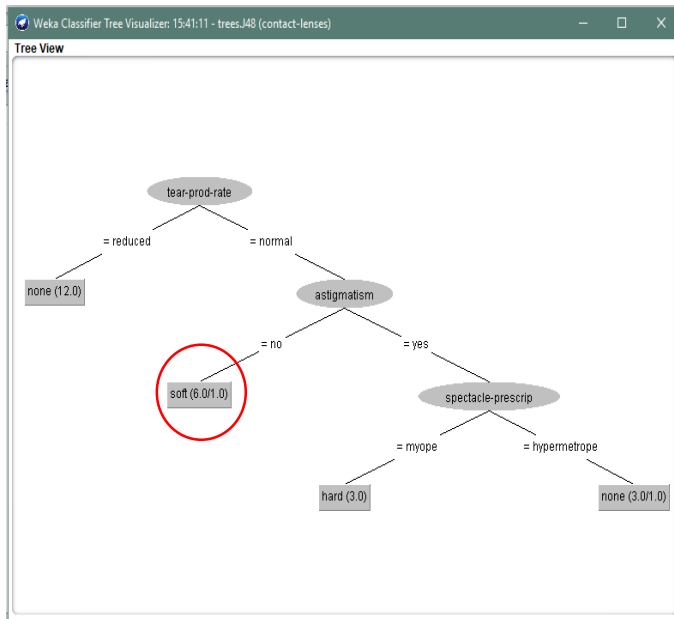
右圖為 visualize tree 所呈現的樹。
以下列出三個 classification rule：

*If tear-prod-rate = reduced
then contact-lenses = none*

*If tear-prod-rate = normal
astigmatism = no
then contact-lenses = soft*

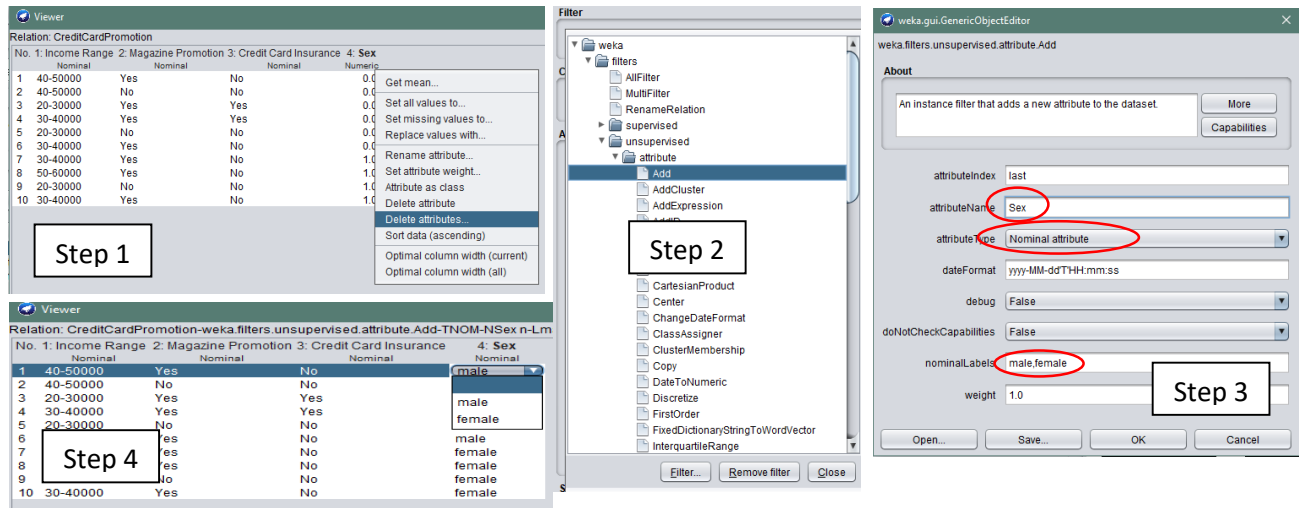
*If tear-prod-rate = normal
astigmatism = yes
spectacle-prescrip = myope
then contact-lenses = hard*

在 leaf 中有一個部分值得說明，在 soft(6.0/1.0) 表示有六個 instance 的 predict 為 soft，但是有一筆是錯誤預測，其 actual 值為 none。



2. 用 Weka 軟體對 CreditCardPromotion.arff 進行 Association Rule ，並使用 Apriori 演算法，設定 confidence = 0.9 、 minimum support = 0.2 ，在過程中對重要步驟截圖並加以說明，並回答以下問題：

(a) 請嘗試著修改 CreditCardPromotion.arff 的欄位與上圖相同，使其可以執行 Association Rule ，請說明使用的方法以及解釋原來的檔案不能執行的原因？(15%)



Step 1: Delete the “Sex” attribute

Step 3: Left click the add filter to edit properties

Step 2: Find the filter called “Add”

Step 4: change the value of all instances respectively

The Apriori algorithm does not work on numeric value; therefore, all variables must be nominal.

(b) 請將 numRule 設成 5 和 10，其各別執行後的 Minimum support 為何，請比較兩者並說明造成其差異的原因。(10%)

Best rules found:

```
1. Income Range=30-40000 4 ==> Magazine Promotion=Yes 4 <conf:(1)> lift:(1.43) lev:(0.12) [1] conv:(1.2)
2. Sex=Female 4 ==> Credit Card Insurance=No 4 <conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0.8)
3. Magazine Promotion=No 3 ==> Credit Card Insurance=No 3 <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
4. Income Range=30-40000 Credit Card Insurance=No 3 ==> Magazine Promotion=Yes 3 <conf:(1)> lift:(1.43) lev:(0.09) [0] conv:(0.9)
5. Magazine Promotion=Yes Sex=Female 3 ==> Credit Card Insurance=No 3 <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
```

When numRule = 5, Minimum support = 0.35.

```
Minimum support: 0.35 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13
```

Best rules found:

```
1. Income Range=30-40000 4 ==> Magazine Promotion=Yes 4 <conf:(1)> lift:(1.43) lev:(0.12) [1] conv:(1.2)
2. Sex=Female 4 ==> Credit Card Insurance=No 4 <conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0.8)
3. Magazine Promotion=No 3 ==> Credit Card Insurance=No 3 <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
4. Income Range=30-40000 Credit Card Insurance=No 3 ==> Magazine Promotion=Yes 3 <conf:(1)> lift:(1.43) lev:(0.09) [0] conv:(0.9)
5. Magazine Promotion=Yes Sex=Female 3 ==> Credit Card Insurance=No 3 <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
6. Income Range=40-50000 2 ==> Credit Card Insurance=No 2 <conf:(1)> lift:(1.25) lev:(0.04) [0] conv:(0.4)
7. Income Range=40-50000 2 ==> Sex=Male 2 <conf:(1)> lift:(1.67) lev:(0.08) [0] conv:(0.8)
8. Credit Card Insurance=Yes 2 ==> Magazine Promotion=Yes 2 <conf:(1)> lift:(1.43) lev:(0.06) [0] conv:(0.6)
9. Credit Card Insurance=Yes 2 ==> Sex=Male 2 <conf:(1)> lift:(1.67) lev:(0.08) [0] conv:(0.8)
10. Income Range=20-30000 Credit Card Insurance=No 2 ==> Magazine Promotion=No 2 <conf:(1)> lift:(3.33) lev:(0.14) [1] conv:(1.4)
```

When numRule = 10, Minimum support = 0.25.

```
Minimum support: 0.25 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15
```

當 Apriori 在做 Association 時，會一直找相符資料直到 numRule 筆，如果沒有的話，Supportmin decrease delta value, which is 0.05 in both cases here，再做 Association 直到有 numRule 筆相符資料。在這邊 numRule=10 為了找出更多規則必須降低比 numRule=5 更多的 support 才能找到，因此 numRule=10 的 Minimum support < numRule5 的 Minimum support， $0.25 < 0.35$ 。

(c) 將 numRule 設成 10，列出前 5 條 Best rule。(10%)

1. *Income Range=30-40000 4 ==> Magazine Promotion=Yes 4*

<conf:(1)> lift:(1.43) lev:(0.12) [1] conv:(1.2)

2. *Sex=female 4 ==> Credit Card Insurance=No 4*

<conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0.8)

3. *Magazine Promotion=No 3 ==> Credit Card Insurance=No 3*

<conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)

4. *Income Range=30-40000 Credit Card Insurance=No 3==>Magazine Promotion=Yes 3*

<conf:(1)> lift:(1.43) lev:(0.09) [0] conv:(0.9)

5. *Magazine Promotion=Yes Sex=female 3 ==> Credit Card Insurance=No 3*

<conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)

(d) 如何在 Associator output 產生 Itemset，請截圖說明並附上 Itemset 結果。(10%)

The screenshot shows the Weka GUI for the Apriori algorithm. The 'outputItemSets' checkbox is checked. The 'Associator output' window displays the results of the algorithm, showing three sets of large itemsets: L(1), L(2), and L(3). Each set is highlighted with a colored box and labeled as 1-ItemSet, 2-ItemSet, and 3-ItemSet respectively.

1-ItemSet

Size of set of large itemsets L(1): 7

Large Itemsets L(1):

- Income Range=20-30000 3
- Income Range=30-40000 4
- Magazine Promotion=Yes 7
- Magazine Promotion=No 3
- Credit Card Insurance=No 8
- Sex=Male 6
- Sex=Female 4

2-ItemSet

Size of set of large itemsets L(2): 8

Large Itemsets L(2):

- Income Range=30-40000 Magazine Promotion=Yes 4
- Income Range=30-40000 Credit Card Insurance=No 3
- Magazine Promotion=Yes Credit Card Insurance=No 5
- Magazine Promotion=Yes Sex=Male 4
- Magazine Promotion=Yes Sex=Female 3
- Magazine Promotion=No Credit Card Insurance=No 3
- Credit Card Insurance=No Sex=Male 4
- Credit Card Insurance=No Sex=Female 4

3-ItemSet

Size of set of large itemsets L(3): 2

Large Itemsets L(3):

- Income Range=30-40000 Magazine Promotion=Yes Credit Card Insurance=No 3
- Magazine Promotion=Yes Credit Card Insurance=No Sex=Female 3

Step 1: Select Apriori Step 2: Adjust OutputItemSet to True Step 3: Click OK

(e) 請自己計算 (記錄在 Word 上或手算拍照附圖皆可)，並與 (d) 小題結果做驗證。(15%)

將不符合 Support min 的資料刪除，剩下的資料從 1-ItemSet extend to 3-ItemSet。
Support min = $0.2 \times 10 \times 0.2 = 2$ (至少要有兩筆重複出現才符合條件)

(紅色畫掉的表示不符合 support min (<2 筆))

1-ItemSet: 7 筆

Income Range = 20-30000	3
Income Range = 30-40000	4
Income Range = 40-50000	2
Income Range = 50-60000	1
Magazine Promotion = yes	7
Magazine Promotion = no	3
Credit Card Insurance = yes	2
Credit Card Insurance = no	8
Sex = male	6
Sex = female	4

2-ItemSet: 8 筆

Income Range = 20-30000 Magazine Promotion = yes	1
Income Range = 20-30000 Magazine Promotion = no	2
Income Range = 30-40000 Magazine Promotion = yes	4
Income Range = 30-40000 Magazine Promotion = no	0
Income Range = 20-30000 Credit Card Insurance = yes	1
Income Range = 20-30000 Credit Card Insurance = no	0
Income Range = 30-40000 Credit Card Insurance = yes	1
Income Range = 30-40000 Credit Card Insurance = no	3
Magazine Promotion = yes Credit Card Insurance = no	5
Magazine Promotion = no Credit Card Insurance = no	3
Magazine Promotion = yes Sex = male	4
Magazine Promotion = yes Sex = female	3
Magazine Promotion = no Sex = male	2
Magazine Promotion = no Sex = female	1
Credit Card Insurance = no Sex = male	4
Credit Card Insurance = no Sex = female	4

3-ItemSet: 2 筆

Income Range = 30-40000 Magazine Promotion = yes Credit Card Insurance = no	3
Income Range = 30-40000 Magazine Promotion = yes Sex = male	2
Income Range = 30-40000 Magazine Promotion = yes Sex = female	2
Magazine Promotion = no Credit Card Insurance = no Sex = male	2
Magazine Promotion = no Credit Card Insurance = no Sex = female	1
Magazine Promotion = yes Credit Card Insurance = no Sex = male	2
Magazine Promotion = yes Credit Card Insurance = no Sex = male	3