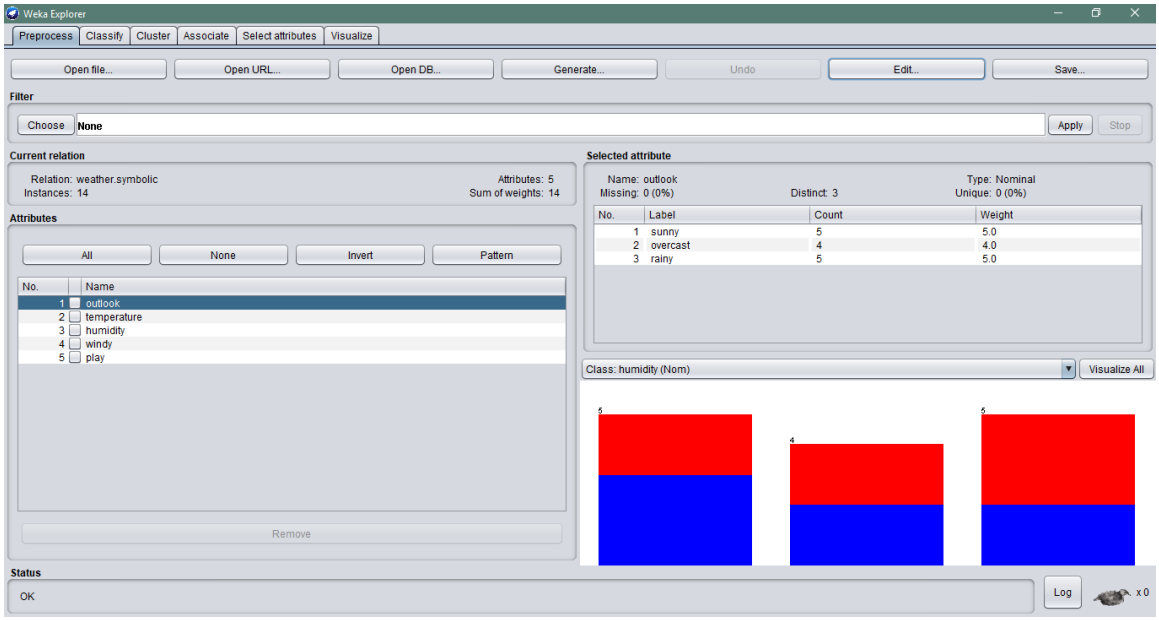
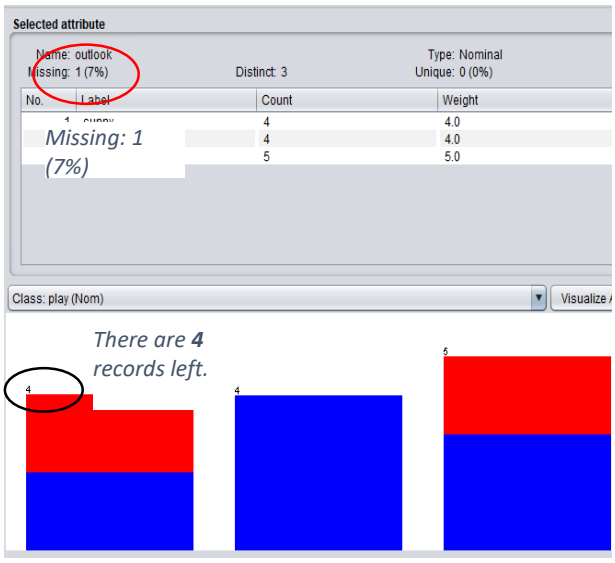
1. 用 Weka 軟體對 weather. nominal. Arff 建立 OneR 規則，選擇 “Use training set”，設定 Attribute: humidity 為 Output，在過程中對重要步驟截圖並加以說明，並回答以下問題：



(a) 在前處理部分，點選 Attribute: outlook ，請問右側的 Selected attribute 中的四個欄位 Type, Missing, Distinct, Unique 分別代表什麼意義? (20%)

Type: 這邊是 nominal，表示每一個 record 的 outlook attribute 是 nominal 而非 numeric 等 type。

Missing: 這邊表示沒有任何 missing data。如果我刪除一筆 sunny 的資料，就會出現**右圖**。他會顯示 number of missing values 以及 ratio of missing values of the whole dataset.



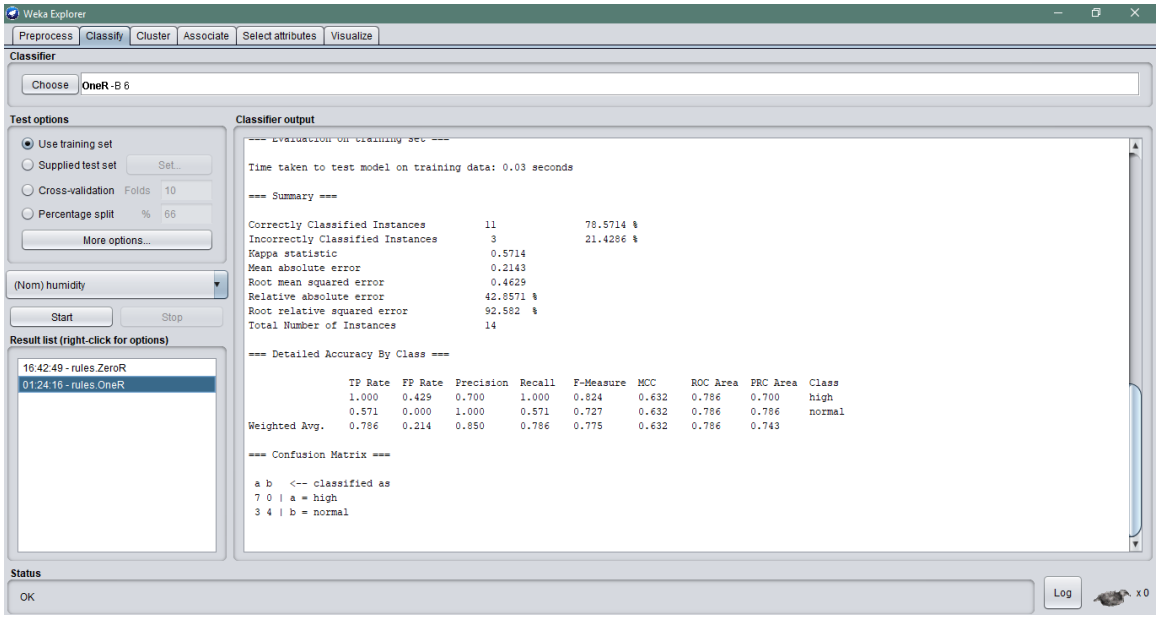
Distinct: 這邊有三種 distinct value。

Unique: 這邊表示 the number of value that appears once，這邊都不只有一次所以是 0。

(b) 請解釋 Test Option 選擇 Use training set 的意義為何 ? (5%)

如果說使用 training set，且用 evaluation 的觀點去看的話，就是一個非常不好的想法。因爲就像是現實生活相對於考試的時候看到一模一樣的題目一樣，就會造成 overfitting。所以千萬不要 testing on training dataset。

(c) 請解釋 Classifier Output 中 Test data 的正確率為多少？有多少筆 Test dataset instances 被分類到 high class 但是實際是屬於 normal class？請解釋 Confusion matrix 和預測結 果之間的關係。 (15%)



Test data 的正確率為 78.5714%。

有 3 個 instances 被分類到 high 但事實上為 normal。

那麼 confusion matrix 和預測之間的關係，我以象限來解釋：

第一象限：data = high but predict normal (true positive)

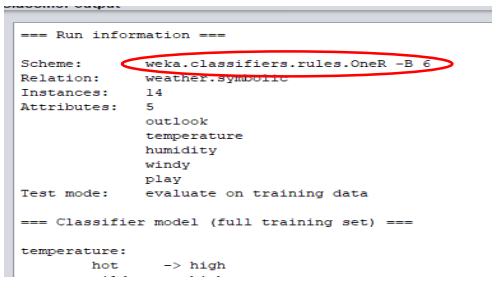
第二象限：data = high and predict high (false negative)

第三象限：data = normal but predict high (false positive)

第四象限：data = normal and predict normal (true negative)

```
Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances          11               78.5714 %
Incorrectly Classified Instances         3               21.4286 %
Kappa statistic                          0.5714
Mean absolute error                      0.2143
Root mean squared error                  0.4629
Relative absolute error                 42.8571 %
Root relative squared error             92.582  %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC
                 1.000     0.429     0.700       1.000   0.824       0.632
                 0.571     0.000     1.000       0.571   0.727       0.632
Weighted Avg.    0.786     0.214     0.850       0.786   0.775       0.632

=== Confusion Matrix ===

 a b   <-- classified as
 7 0 | a = high
 3 4 | b = normal
```

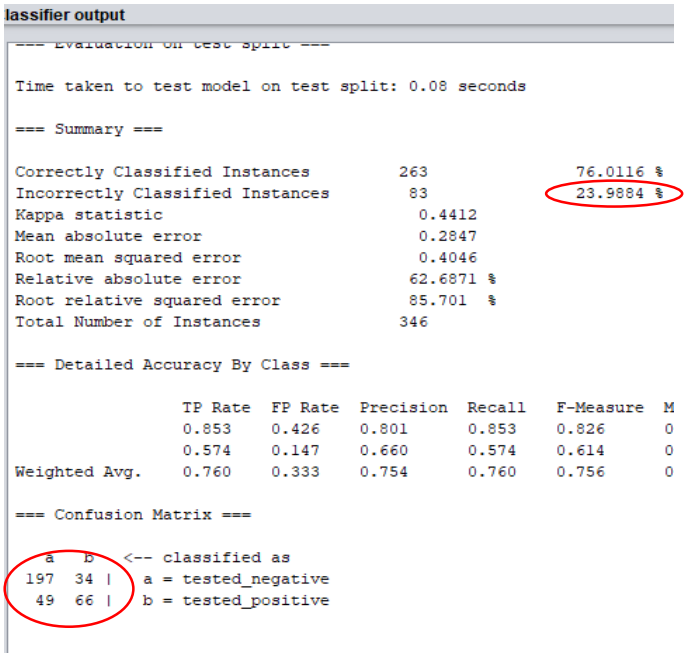
(d) 請將分類規則截圖，並加以說明為何是這個分類規則。 (10%)



這裏使用的是 OneR。因爲他可以針對 nominal 的 data 簡單且準確地對於每一個 predictor 產生 rules，再從其中找出 the smallest total error 的作爲他的 one rule。

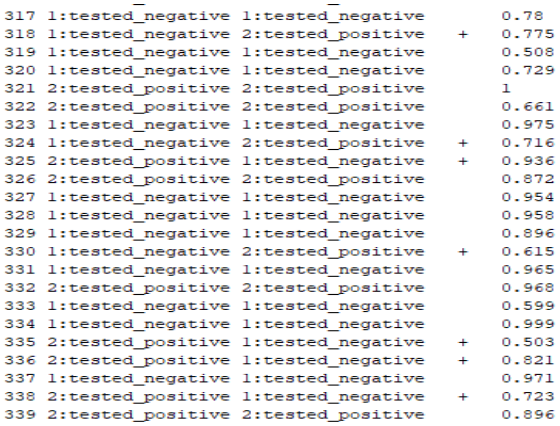
2. 用 Weka 軟體對 diabetes. arff 利用 Naïve Bayes 進行Supervised learning，選擇 “Percentage split: 55%” ，設定 Attribute: class 為 Output ，在過程中對重要步驟截圖並加以說明，並 回答以下問題:

(a) 解釋 Classifier Output，Test data 的錯誤率是多少 ?有多少百分比的 Test dataset instances 被分類到 tested negative class 但實際上屬於 tested positive class ?請利用 Confusion matrix 解釋。 (15%)



錯誤率可由 incorrectly classified instances 看到為 23.9884%。由 confusion matrix 可以得知 49 out of 346 筆資料是被分類到 tested negative class 但實際上屬於 tested positive class，所以答案為 14.1618%。

(b) 在 Output predictions 的結果中，欄位 error 出現 “+” 代表意思為何 ?請截圖並解釋之。 (10%)



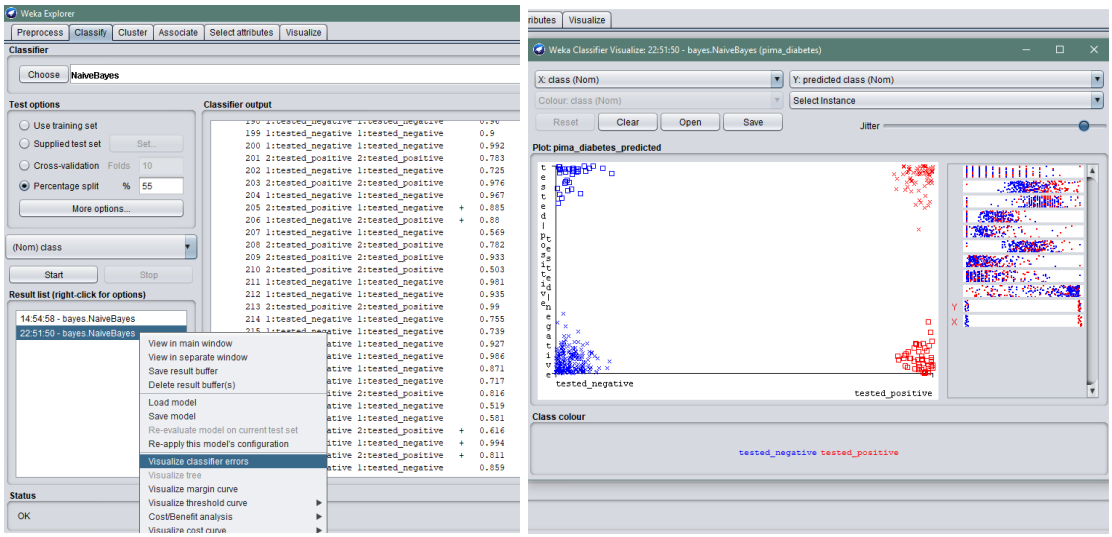
這些有加號的是 Test dataset instances 被分類到 tested negative class 但實際上屬於 tested positive class 或是 Test dataset instances 被分類到 tested positive class 但實際上屬於 tested negative class。

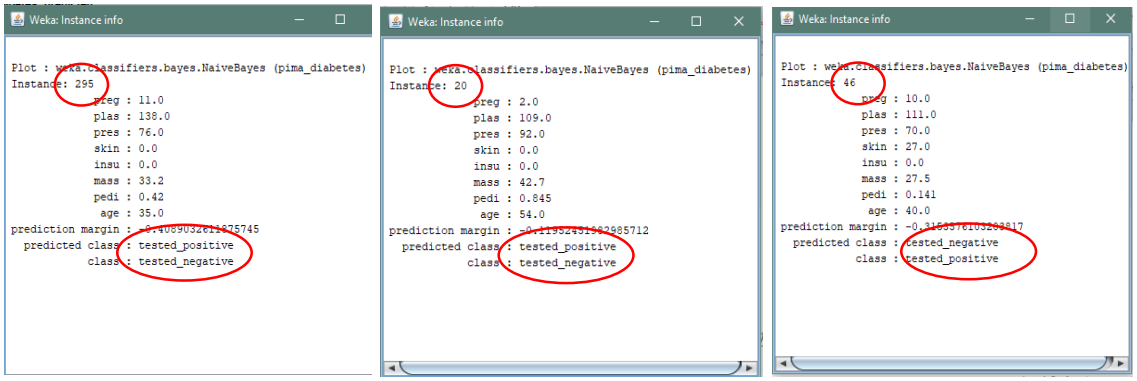
(c) 請利用 Visualize Classifier Errors，找出預測錯誤的資料點 3 個，並寫出各是第幾筆資料，請截圖操作步驟並解釋。 (15%)

Step 1: click the right button on the result and select “visualize classifier errors”.

Step 2: Adjust the jitter button in order to clarify the dots on the plot.

Step 3: Left click on a random dot and it will show you the details.





我所選出來的分別是 instance 295 20 46。而在下方紅圈內都可看出他們是 error。

(d) 請使用 Visualize Classifier Errors，解釋產生的圖以及此圖與 Confusion matrix 之間的 關係。 (10%)

The plot is actually a visualized confusion matrix, which contains every single instances. Therefore, I added the number of instances of every section to make it clearer. Y 軸是 prediction，X 軸是 realistic instances。