CSC662 ปฏิบัติการ ๘

ต้นไม้การตัดสินใจในซอฟต์แวร์ Weka

เขียนโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

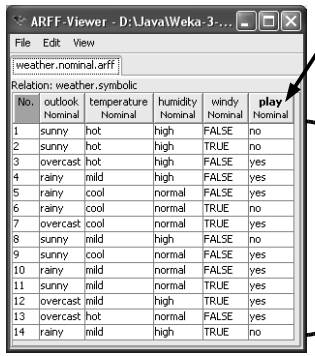
เนื้อหาที่ครอบคลุม

- การทำเหมืองข้อมูลการจัดจำแนกประเภท Classification
- การเตรียมข้อมูลสำหรับการจัดจำแนกประเภท
- การเลือกใช้ต้นไม้การตัดสินใจ
- ผลลัพธ์ที่ได้โดยใช้ ID3 ซึ่งไม่ใช้ลักษณะประจำที่เป็นค่าต่อเนื่อง
- ผลลัพธ์ที่ได้โดยใช้ J48 ซึ่งใช้ได้กับลักษณะประจำที่ต่อเนื่องและไม่ต่อเนื่อง

การทำเหมืองข้อมูลการจัดจำแนกประเภท

- เป็นการสร้างตัวแบบ Classifier ที่สามารถแบ่งแยกข้อมูล (ตัวอย่าง) ออกตามคลาสหรือลักษณะประจำเป้าหมายที่กำหนด
- ตัวแบบที่ต้องการอาจเป็น
  - bayes ใช้หลักของเบย์หรือตัวแบบเชิงความน่าจะเป็น
  - functions ตัวแบบในรูปของฟังก์ชัน
  - lazy ตัวแบบที่เก็บตัวอย่าง การตัดสินใจเกิดเมื่อตัวอย่างใหม่ถูกนำเข้าเท่านั้น
  - meta การทำตัวแบบให้ดีขึ้นโดยการเรียนข้อมูลเมตา
  - misc วิธีการสร้างตัวแบบวิธีอื่น
  - trees การสร้างตัวแบบโดยใช้ต้นไม้
  - rules การสร้างตัวแบบโดยใช้กฎ

แฟ้มตัวอย่าง Weather.nominal.arff



No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

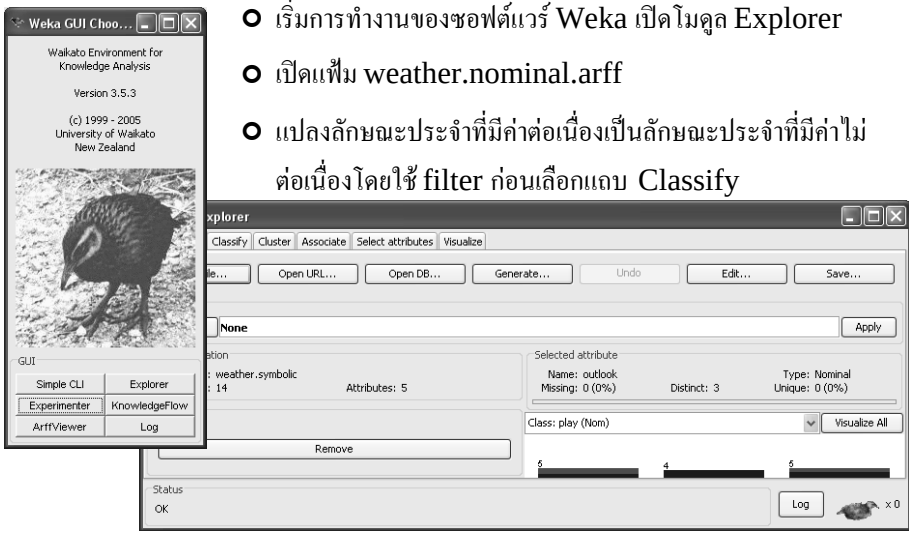
ลักษณะประจำเป้าหมาย
เป็นลักษณะประจำสุดท้ายในตาราง

ตัวอย่างมีทั้งหมด 14 ตัวอย่าง
และมีลักษณะประจำที่ไม่ใช่
ลักษณะประจำเป้าหมาย 4 ตัว

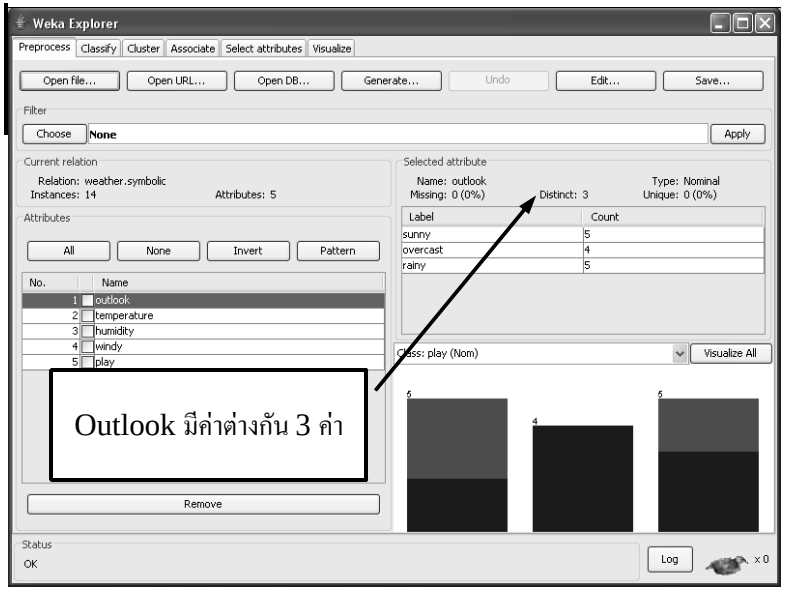
การเตรียมแฟ้มข้อมูล

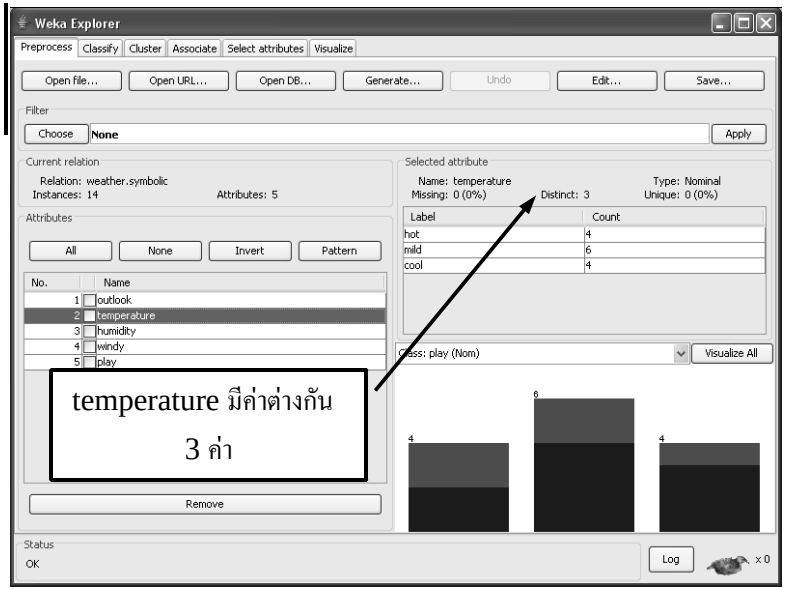
- กำหนดลักษณะประจำเป้าหมายให้เป็นลักษณะประจำสุดท้าย
- บางขั้นตอนวิธีที่ใช้สร้างต้นไม้การตัดสินใจต้องการลักษณะประจำที่มีค่าไม่ต่อเนื่องเท่านั้น ดังนั้นเราจำเป็นต้องเปลี่ยนลักษณะประจำที่มีค่าต่อเนื่องให้เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่อง
- ในกรณีที่มีระเบียนน้อย เราอาจใช้ *k*-fold cross validation หรือ leave-one-out
- ในกรณีที่มีระเบียนมากเพียงพอ เราควรแบ่งกั้นระเบียนบางส่วนเป็น validation, test data และที่เหลือนำมาใช้เป็น training data สัดส่วนที่ใช้อาจเป็น 3/10, 3/10 กับ 4/10

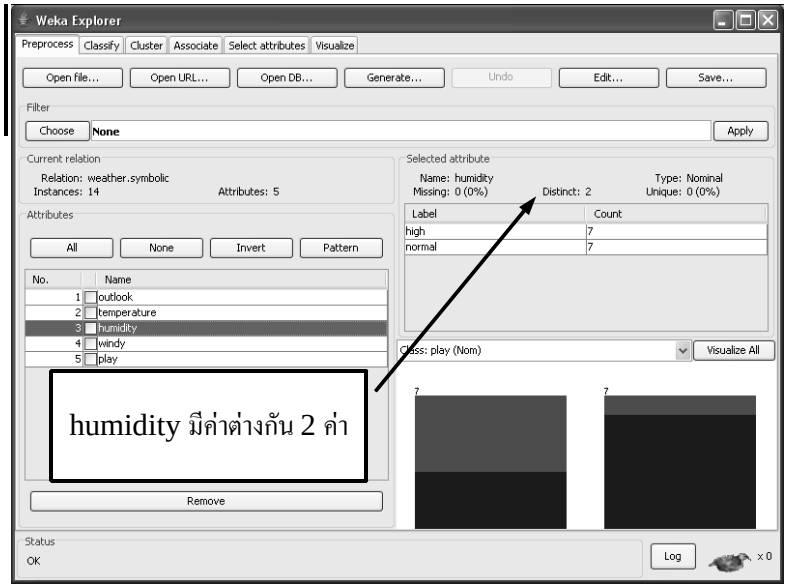
การใช้งานซอฟต์แวร์ Weka explorer

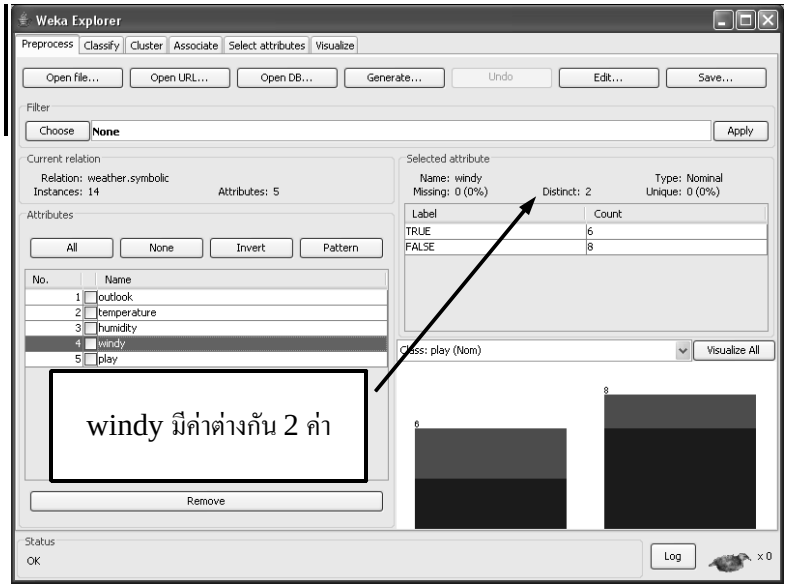


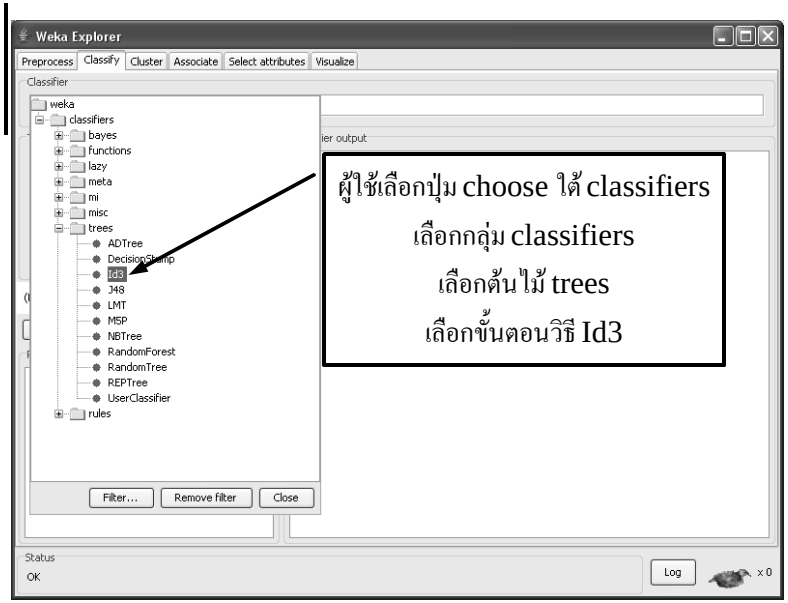
- เริ่มการทำงานของซอฟต์แวร์ Weka เปิดโมดูล Explorer
- เปิดแฟ้ม weather.nominal.arff
- แปลงลักษณะประจำที่มีค่าต่อเนื่องเป็นลักษณะประจำที่มีค่าไม่ต่อเนื่องโดยใช้ filter ก่อนเลือกแถบ Classify

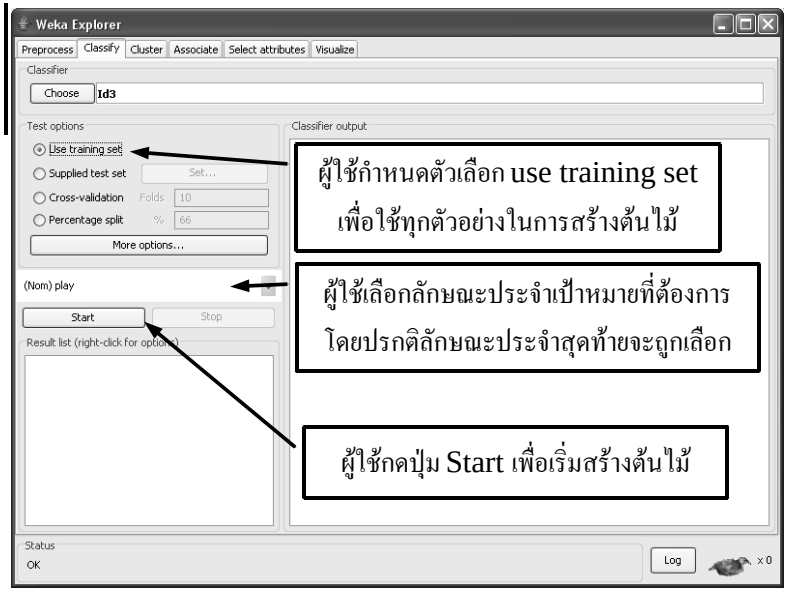


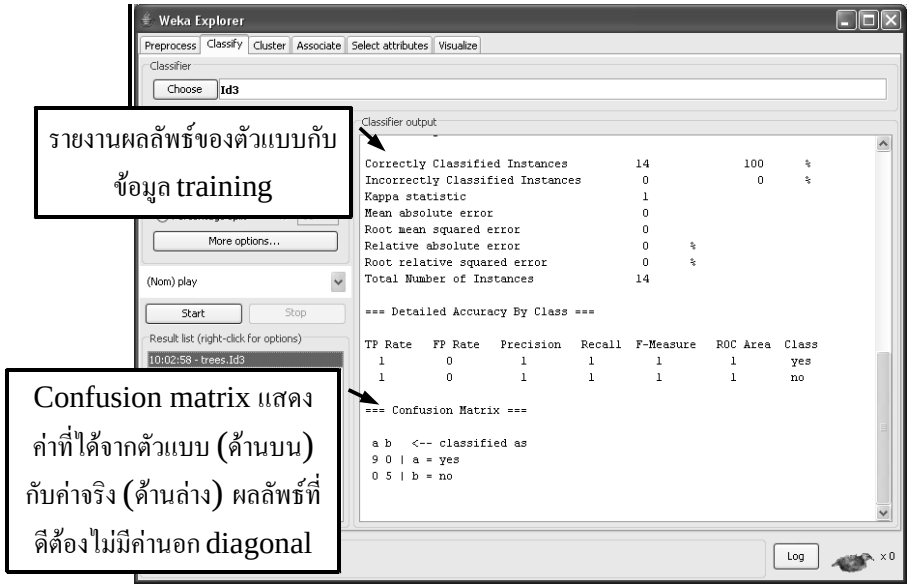



Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None Apply
Current relation
Relation: weather.symbolic
Instances: 14 Attributes: 5
Selected attribute
Name: humidity Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)
Label Count
high 7
normal 7
Attributes
All None Invert Pattern
No. Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Class: play (Nom) Visualize All
Remove
Status
OK
Log
x 0
humidity มีค่าต่างกัน 2 ค่า


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None Apply
Current relation
Relation: weather.symbolic
Instances: 14 Attributes: 5
Selected attribute
Name: windy Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)
Label Count
TRUE 6
FALSE 8
Attributes
All None Invert Pattern
No. Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Class: play (Nom) Visualize All
Remove
Status
OK
Log
x 0
windy มีค่าต่างกัน 2 ค่า


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
mi
misc
trees
ADTree
DecisionStump
Id3
J48
LMT
M5P
NBTree
RandomForest
RandomTree
REPTree
UserClassifier
rules
Filter... Remove filter Close
Status
OK
Log
x 0
ผู้ใช้เลือกปุ่ม choose ใต้ classifiers
เลือกกลุ่ม classifiers
เลือกต้นไม้ trees
เลือกขั้นตอนวิธี Id3


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose Id3
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
Classifier output
(Nom) play
Start Stop
Result list (right-click for options)
Status
OK
Log
x 0
ผู้ใช้กำหนดตัวเลือก use training set
เพื่อใช้ทุกตัวอย่างในการสร้างต้นไม้
ผู้ใช้เลือกลักษณะประจำเป้าหมายที่ต้องการ
โดยปรกติลักษณะประจำสุดท้ายจะถูกเลือก
ผู้ใช้กดปุ่ม Start เพื่อเริ่มสร้างต้นไม้



รายงานผลลัพธ์ของตัวแบบกับข้อมูล training

Confusion matrix แสดงค่าที่ได้จากตัวแบบ (ด้านบน) กับค่าจริง (ด้านล่าง) ผลลัพธ์ที่ดีต้องไม่มีค่านอก diagonal

แฟ้ม weather.arff

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature real

@attribute humidity real

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

rainy,68,80,FALSE,yes

rainy,65,70,TRUE,no

overcast,64,65,TRUE,yes

sunny,72,95,FALSE,no

sunny,69,70,FALSE,yes

rainy,75,80,FALSE,yes

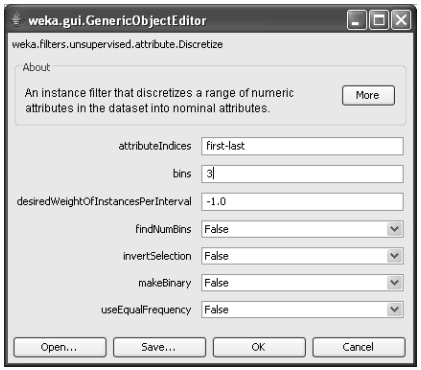
sunny,75,70,TRUE,yes

overcast,72,90,TRUE,yes

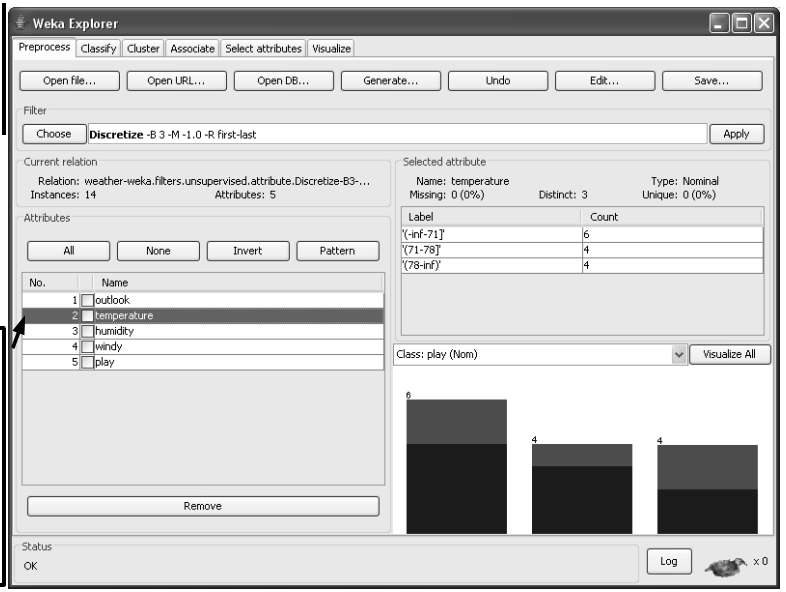
overcast,81,75,FALSE,yes

rainy,71,91,TRUE,no

การเปลี่ยนลักษณะประจำให้เป็นค่าไม่ต่อเนื่อง



- เลือก Discretize ในกล่อง Filter โดยเลือก filters → unsupervised → attribute
- ปรับค่าในกล่อง bins ให้เหมาะสม เช่นกำหนดให้เป็น 3 กล่อง
- กดปุ่ม OK
- แล้วกด Apply



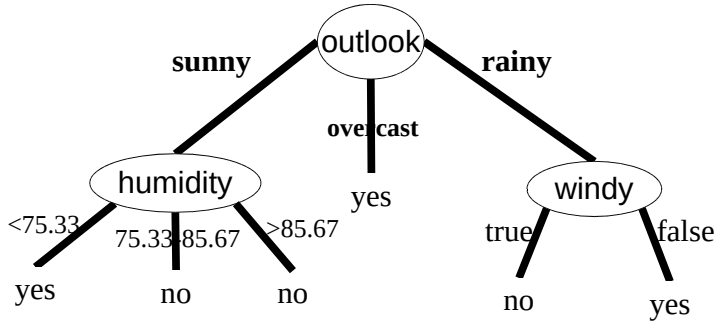
ผลที่ได้จากการแปลงเป็นค่าไม่ต่อเนื่อง

การทำเหมืองข้อมูลแบบจัดจำแนกประเภท ID3

```
Classifier output

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  1         0          1         1         1          1        yes
  1         0          1         1         1          1        no

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

- เลือก Id3 ในกล่อง Classifier ใต้แถบ Classify โดย classifiers → trees → Id3
- เลือก Use training set ในกล่อง Test options
- กดปุ่ม Start
- จะได้ผลลัพธ์ดังรูปด้านซ้าย

ต้นไม้ที่ได้จาก ID3



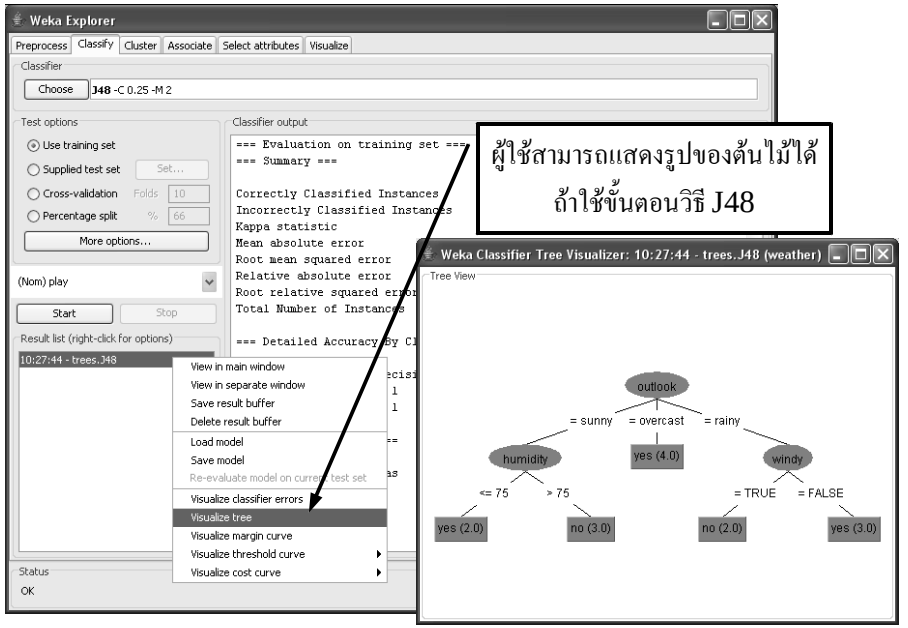
ต้นไม้การตัดสินใจดังกล่าวสามารถจำแนก play ถูกต้อง 100%

```
=== Confusion Matrix ===
 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

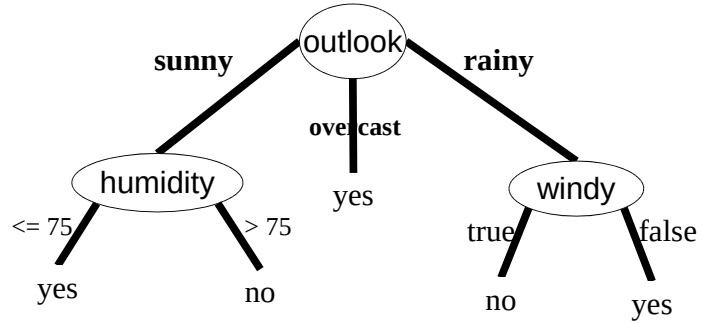

การทำเหมืองข้อมูลแบบจัดจำแนกประเภท J48

```
Classifier output

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  1         0          1         1         1          1        yes
  1         0          1         1         1          1        no

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

- เลือก J48 ในกล่อง Classifier ใต้แถบ Classify โดย classifiers → trees → J48
- เราไม่จำเป็นต้องเปลี่ยนลักษณะประจำให้เป็นชนิดที่มีค่าไม่ต่อเนื่อง
- เลือก Use training set ในกล่อง Test options
- กดปุ่ม Start
- จะได้ผลลัพธ์ดังรูปด้านซ้าย



ต้นไม้ที่ได้จาก J48



ต้นไม้การตัดสินใจดังกล่าวสามารถจำแนก play ถูกต้อง 100%

```
=== Confusion Matrix ===
 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```


สรุป

- การทำเหมืองข้อมูลแบบจัดจำแนกประเภท มีขั้นตอนวิธีในการสร้างตัวแบบมากมาย
- การใช้ต้นไม้ในการบ่งบอกตัวแบบก็เป็นหนึ่งในวิธีดังกล่าว
- สำหรับขั้นตอนวิธี Id3 ลักษณะประจำทุกตัวต้องมีค่าไม่ต่อเนื่อง
- แต่ขั้นตอนวิธี J48 ลักษณะประจำไม่จำเป็นต้องมีค่าไม่ต่อเนื่อง