

# CLASSIFICATION STEPS (1)

1

ขั้นตอนการสร้างโมเดล  
(Classification model building)

ข้อมูลการเรียนรู้  
(Train data)

เทคนิค  
Classification

แอลทริบิวต์ 1	แலทริบิวต์ 1	แอลทริบิวต์ 1
1	0	A
0	1	B

2

ขั้นตอนการวัดประสิทธิภาพ  
(Evaluation)

ข้อมูลทดสอบ  
(Evaluation)

โมเดล  
Model

คลาสที่  
ทำนาย  
A

คลาสที่  
ทำนาย  
A

แอลทริบิวต์ 1	แอลทริบิวต์ 2
1	0

เปรียบเทียบผลที่ได้จาก  
โมเดลและคำตอบจริง

# CLASSIFICATION STEPS (2)

1

ขั้นตอนการสร้างโมเดล  
(Classification model building)

ข้อมูลการเรียนรู้  
(Trainy data)

เทคนิค  
Classification

แอทริบิวต์ 1	แอทริบิวต์ 1	แอทริบิวต์ 1
1	0	A
0	1	B

2

ขั้นตอนการวัดประสิทธิภาพ  
(Evaluation)

ข้อมูลทดสอบ  
(Evaluation)

โมเดล  
Model

คลาสที่  
ทํานาย  
A  
คลาสที่  
ทํานาย  
A

แอทริบิวต์ 1	แอทริบิวต์ 2
1	0

3

การใช้งานจริง

Unseen data

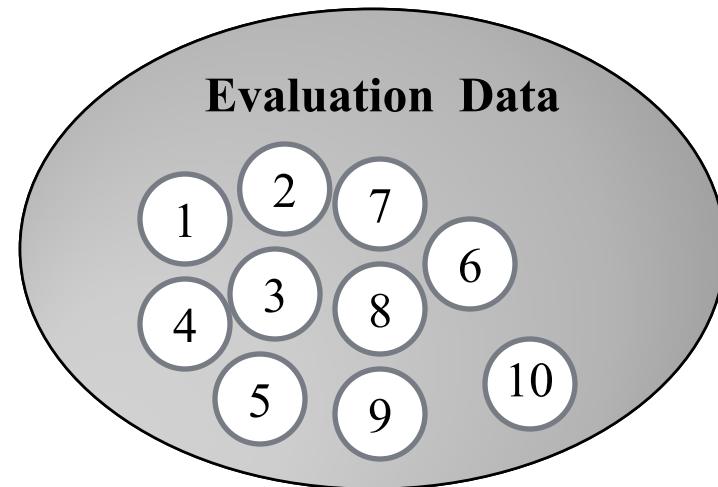
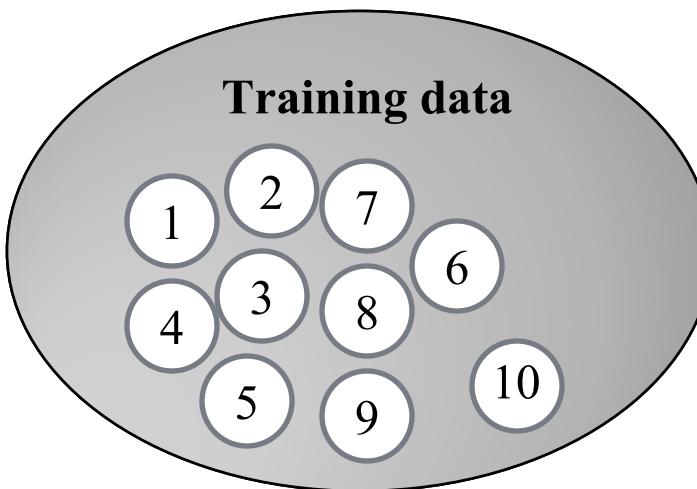
แอทริบิวต์ 1	แอทริบิวต์ 2
1	0

โมเดล  
Model

คลาส  
A

# EVALUATION MODEL :SELF CONSISTENCY TEST

- การวัดประสิทธิภาพของโมเดลโดยใช้ข้อมูลเรียนรู้ training data
  - ใช้ข้อมูลเรียนรู้ทั้งหมดเพื่อ Evaluate โมเดล (Self Consistency Test)
  - ควรจะได้ค่าความถูกต้องที่สูงหรือเท่ากับ 100 %

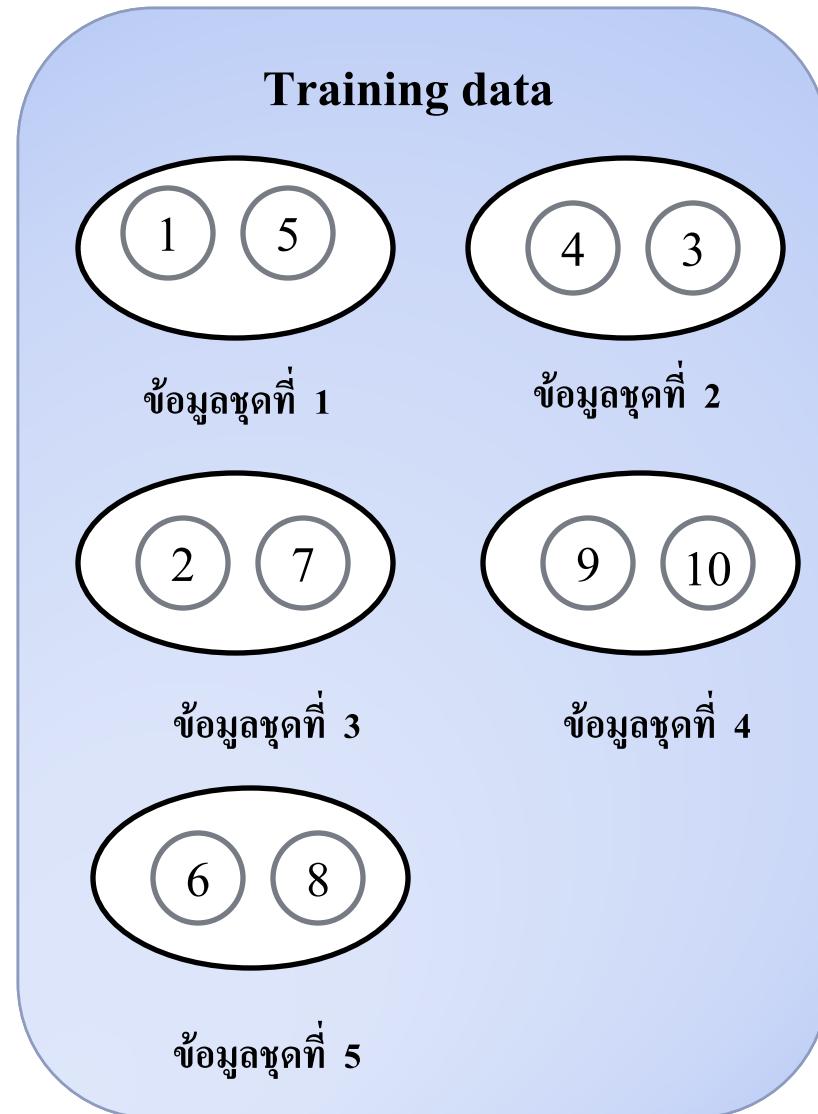


# EVALUATION MODEL :CROSS VALIDATION

## ○ การวัดประสิทธิภาพของโมเดลแบบ

### cross-validation

- แบ่งข้อมูลเรียนรู้ออกเป็น  $k$  ชุดเท่า ๆ กัน
- ใช้ข้อมูลส่วนที่เหลือ ( $k - 1$  ชุด) เพื่อทำการสร้างโมเดล
- เก็บข้อมูลที่แบ่งไว้ 1 ชุด เพื่อทำการ Evaluate
- วนทำซ้ำจนข้อมูลทุกส่วนถูกนำมาทดสอบ



## EVALUATION MODEL :CROSS VALIDATION(2)

○ การวัดประสิทธิภาพของโมเดลแบบ cross – validation ที่นิยมได้แก่

1. 5-fold cross-validation
2. 10-fold cross-validation
3. Leave – One – Out cross-validation



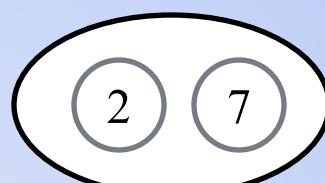
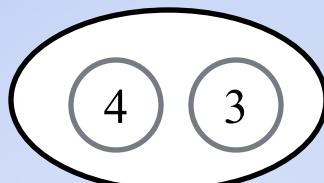
# EVALUATION MODEL :CROSS VALIDATION (3)

- 5 – fold cross -validation

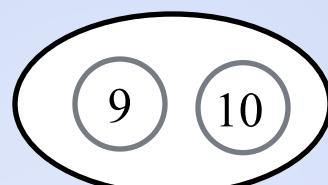
รอบที่ 1

Accuracy รอบที่ 1 = 80.74 %

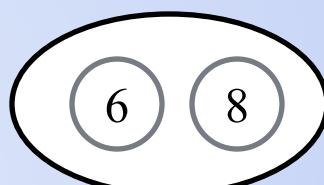
Training data



ข้อมูลชุดที่ 2

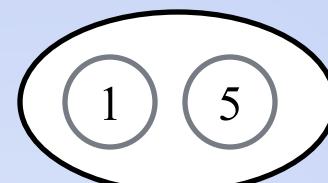


ข้อมูลชุดที่ 4



ข้อมูลชุดที่ 5

Evaluate data



ข้อมูลชุดที่ 1

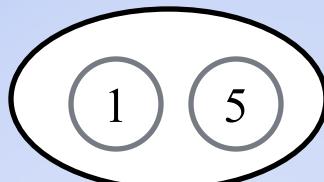
# EVALUATION MODEL :CROSS VALIDATION (4)

- 5 – fold cross -validation

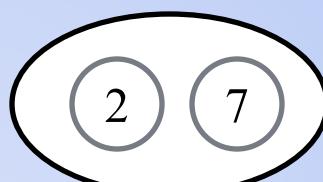
รอบที่ 2

Accuracy รอบที่ 2 = 82.39 %

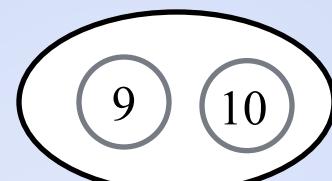
Training data



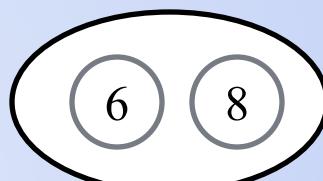
ข้อมูลชุดที่ 1



ข้อมูลชุดที่ 3

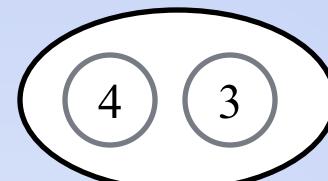


ข้อมูลชุดที่ 4



ข้อมูลชุดที่ 5

Evaluate data



ข้อมูลชุดที่ 2

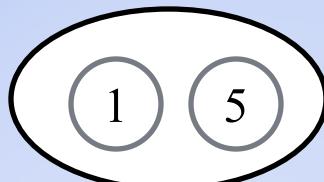
# EVALUATION MODEL :CROSS VALIDATION (5)

- 5 – fold cross -validation

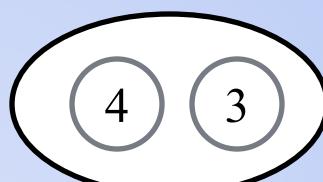
รอบที่ 3

Accuracy รอบที่ 3 = 80.69 %

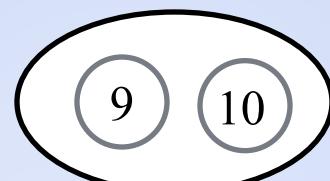
Training data



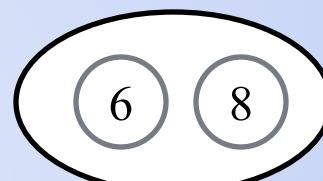
ข้อมูลชุดที่ 1



ข้อมูลชุดที่ 2

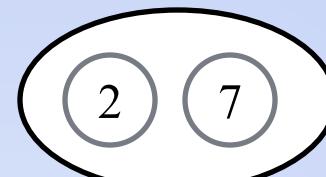


ข้อมูลชุดที่ 4



ข้อมูลชุดที่ 5

Evaluate data



ข้อมูลชุดที่ 3

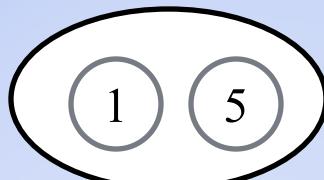
# EVALUATION MODEL :CROSS VALIDATION (6)

- 5 – fold cross -validation

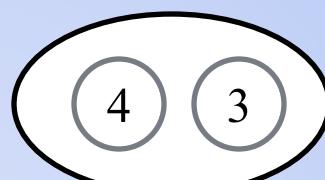
รอบที่ 4

Accuracy รอบที่ 4 = 81.24 %

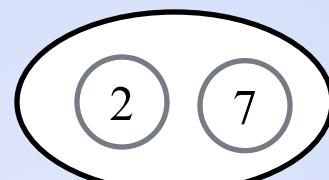
Training data



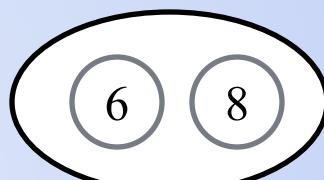
ข้อมูลชุดที่ 1



ข้อมูลชุดที่ 2

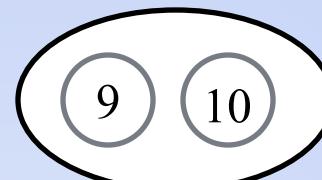


ข้อมูลชุดที่ 3



ข้อมูลชุดที่ 5

Evaluate data



ข้อมูลชุดที่ 4

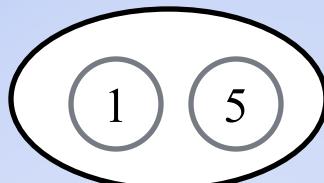
# EVALUATION MODEL :CROSS VALIDATION (7)

- 5 – fold cross -validation

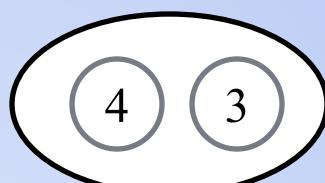
รอบที่ 5

Accuracy รอบที่ 5 = 80.72 %

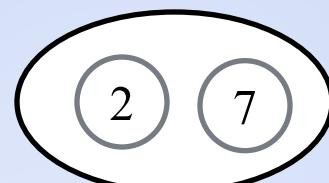
Training data



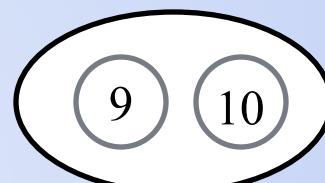
ข้อมูลชุดที่ 1



ข้อมูลชุดที่ 2

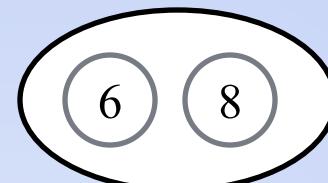


ข้อมูลชุดที่ 3



ข้อมูลชุดที่ 4

Evaluate data



ข้อมูลชุดที่ 5

## EVALUATION MODEL :CROSS VALIDATION (8)

- ค่าความถูกต้องของ Model

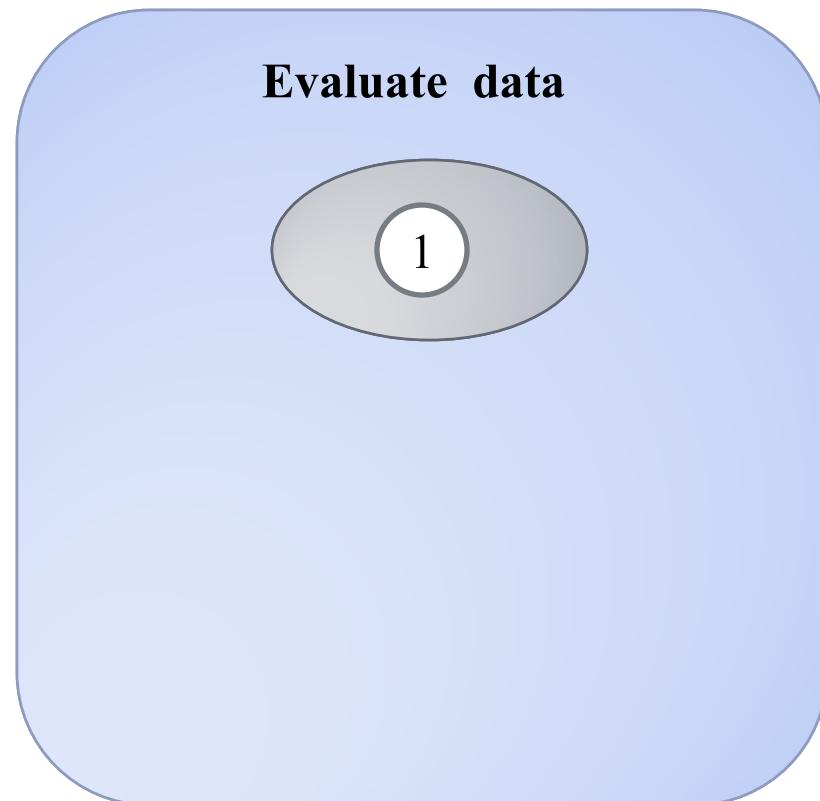
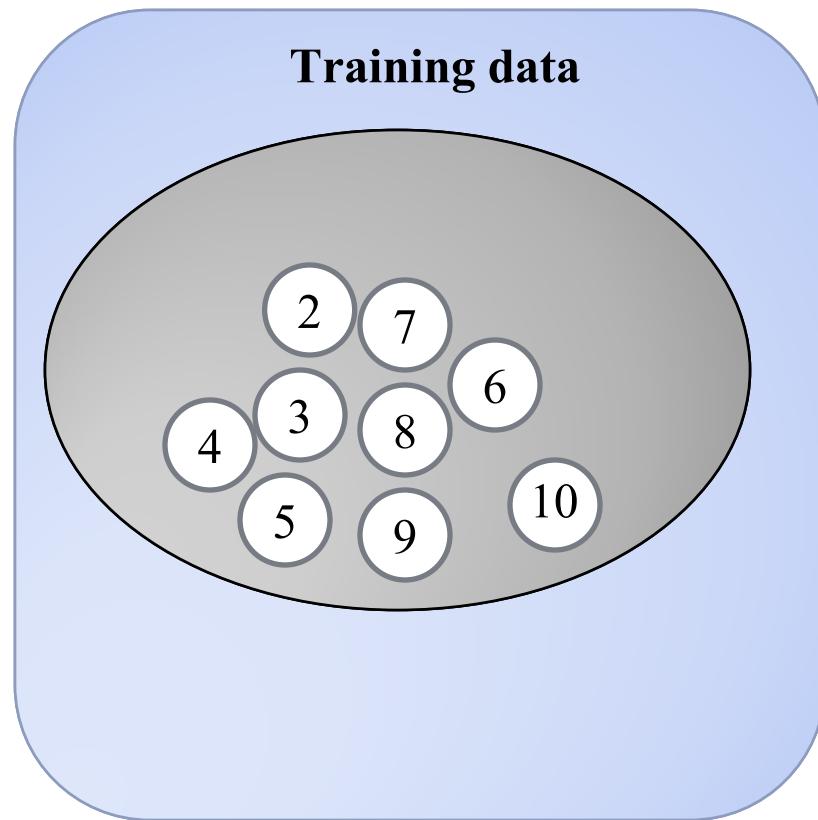
$$\text{overall accuracy} = \text{Average}(\sum_{i=1}^k \text{Accuracy}_i)$$



# EVALUATION MODEL :CROSS VALIDATION (9)

- Leave-One-Out cross-validation

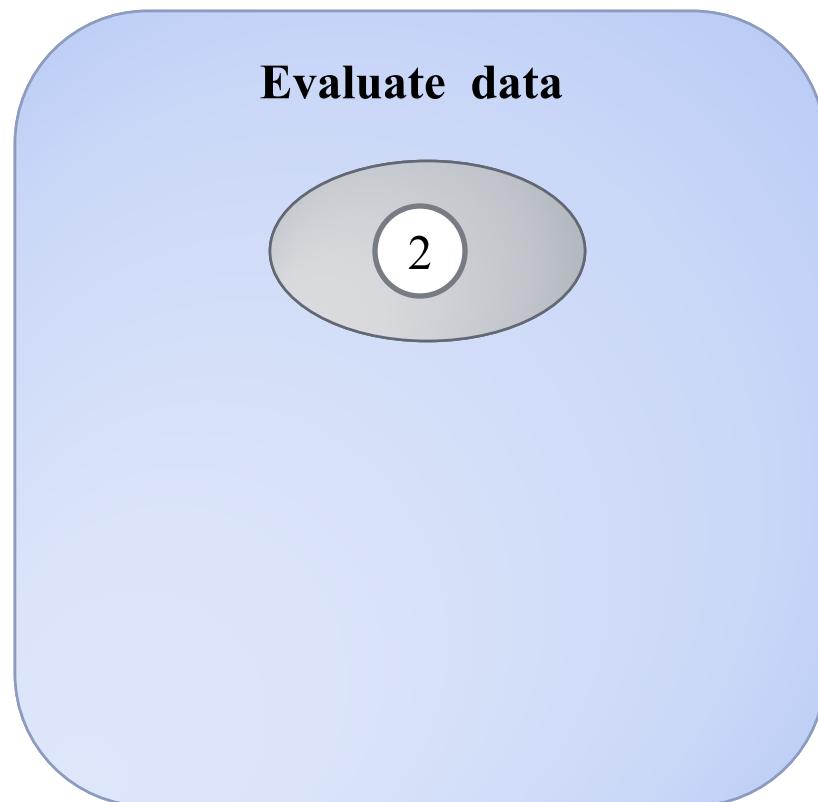
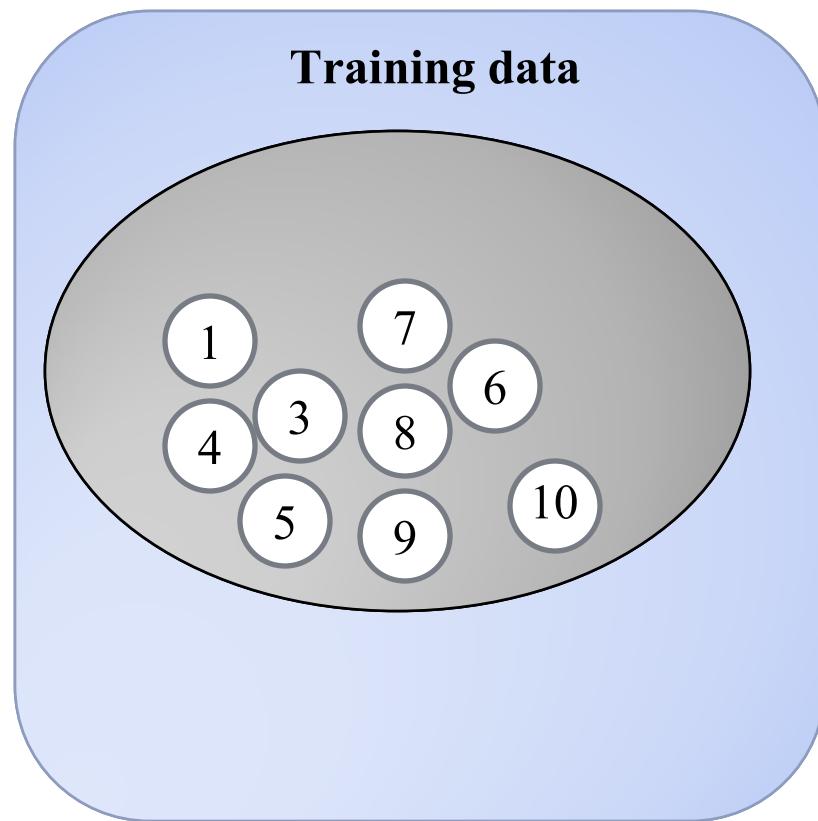
รอบที่ 1



# EVALUATION MODEL :CROSS VALIDATION (10)

- Leave-One-Out cross-validation

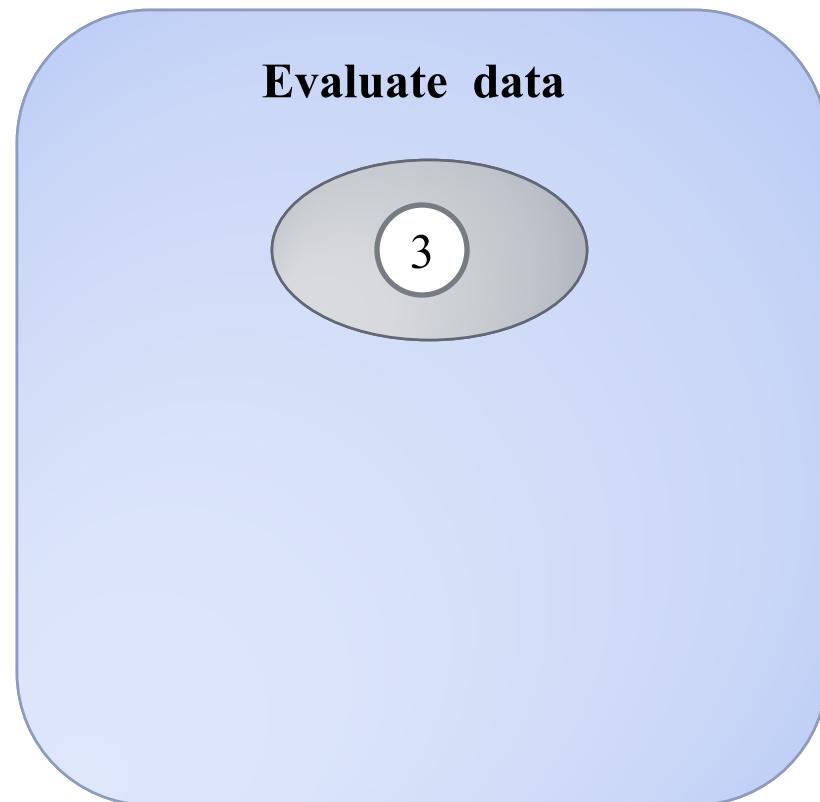
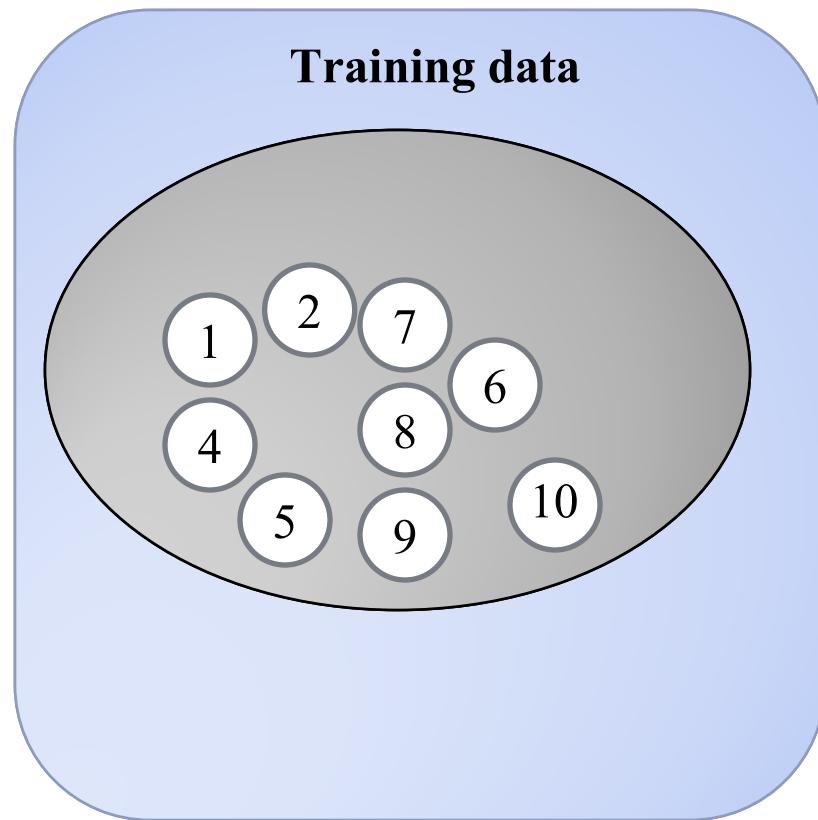
รอบที่ 2



# EVALUATION MODEL :CROSS VALIDATION (11)

- Leave-One-Out cross-validation

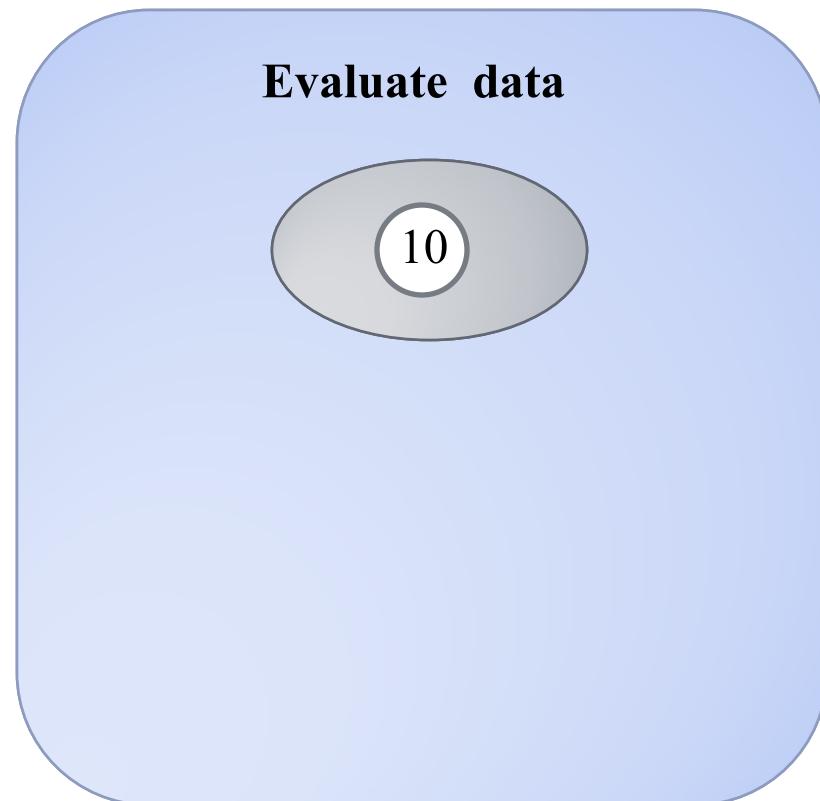
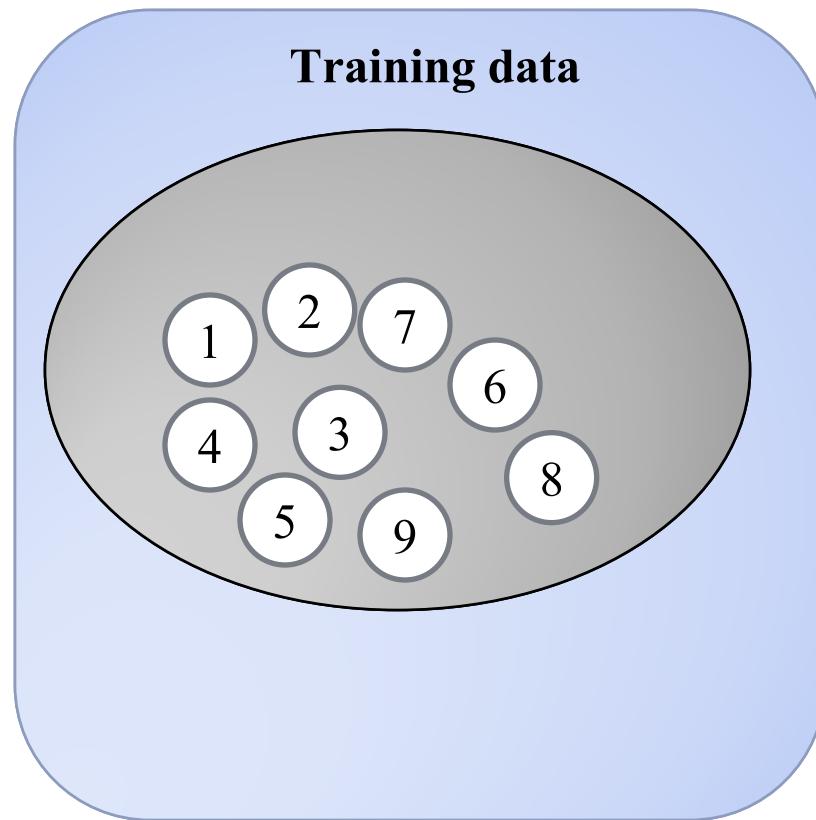
รอบที่ 3



# EVALUATION MODEL :CROSS VALIDATION (12)

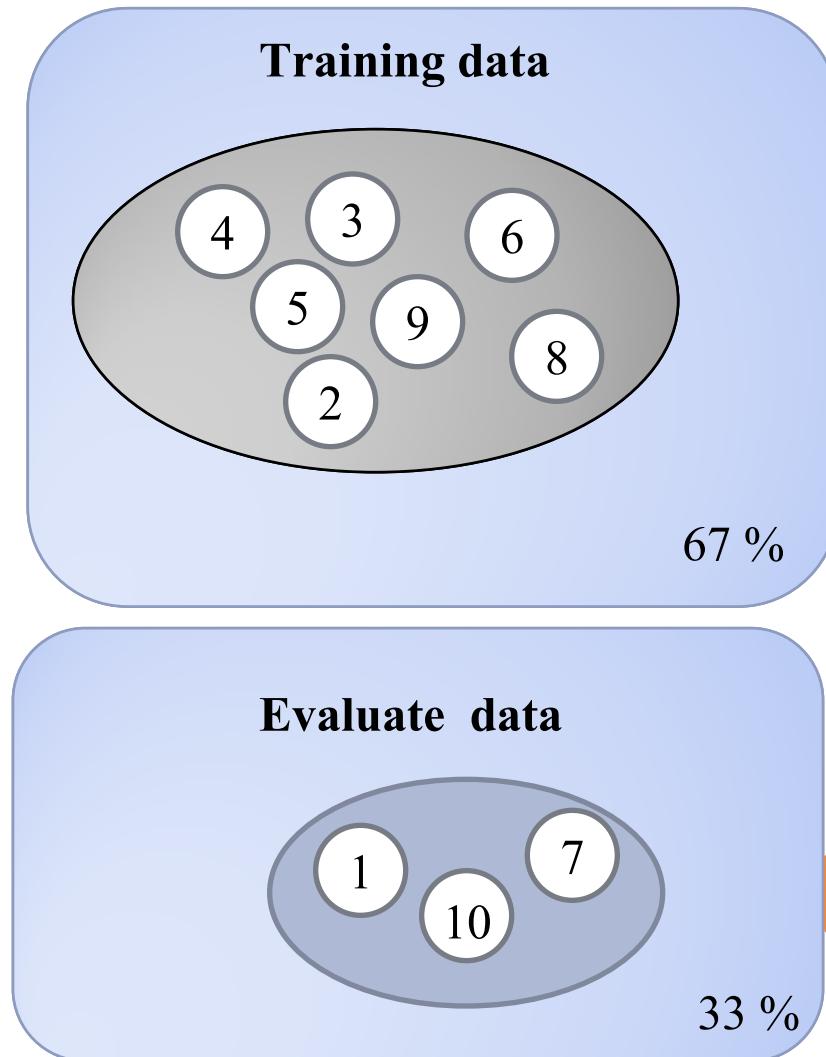
- Leave-One-Out cross-validation

รอบที่ N (N = 10)



# EVALUATION MODEL :SPLIT DATA

- การวัดประสิทธิภาพของโมเดลการใช้ข้อมูลเรียนรู้บางส่วน
  - แบ่งข้อมูลออกเป็น 2 ส่วน
  - ส่วนที่ 1 สำหรับใช้ในการเรียนรู้ (training)
    - เป็นข้อมูลส่วนใหญ่
    - ประมาณ 2 ใน 3 หรือ 66 % ของข้อมูลเรียนรู้
  - ส่วนที่ 1 สำหรับใช้ในการทดสอบ (testing)
    - เป็นข้อมูลส่วนน้อยที่เหลือในข้อมูลเรียนรู้
    - ประมาณ 1 ใน 3 หรือ 33 % ของข้อมูลเรียนรู้



## EVALUATE STEP 2 (2)

- ตัววัดประสิทธิภาพโมเดล

- ค่าความถูกต้อง (Accuracy) สำหรับข้อมูลประเภท (nominal)
- ค่าความคลาดเคลื่อน (Error) สำหรับข้อมูลตัวเลข (numeric)



# EVALUATE STEP 2 (3)

## ○ ค่าความถูกต้อง (accuracy)

- % ของจำนวนข้อมูลที่ตอบถูก
  - ค่าที่ตรงกันระหว่างข้อมูลจริงและข้อมูลที่ทำนายได้

ข้อมูลจริง	ข้อมูลที่ทำนายได้
A	A
D	D
A	D
D	A
A	A
D	D

$$\text{ค่าความถูกต้อง} = (4/6) \times 100 = 66.67\%$$

## ○ ค่าความคลาดเคลื่อน (error)

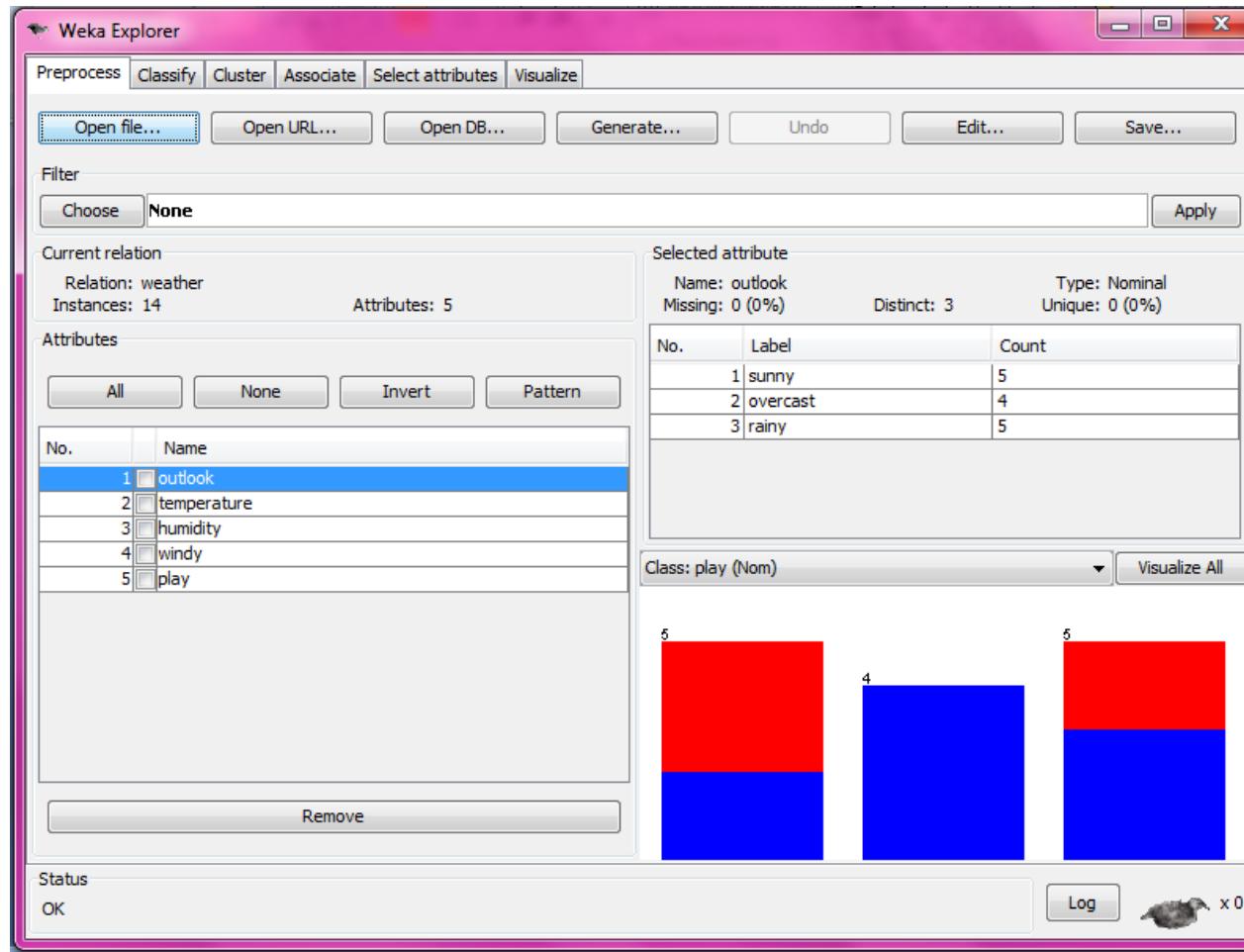
- ค่าที่แตกต่างจากค่าจริง
  - ค่าที่คาดเดาได้ (predicted)- ค่าจริง

ข้อมูลจริง	ข้อมูลที่ทำนายได้
10	10
15	16
17	17.1
20	19.8
25	25
30	30

$$\begin{aligned}\text{ค่าความผิดพลาดรวม} \\ = & |15-16| + |17.1-17| + |20-19.8| = 1.3\end{aligned}$$

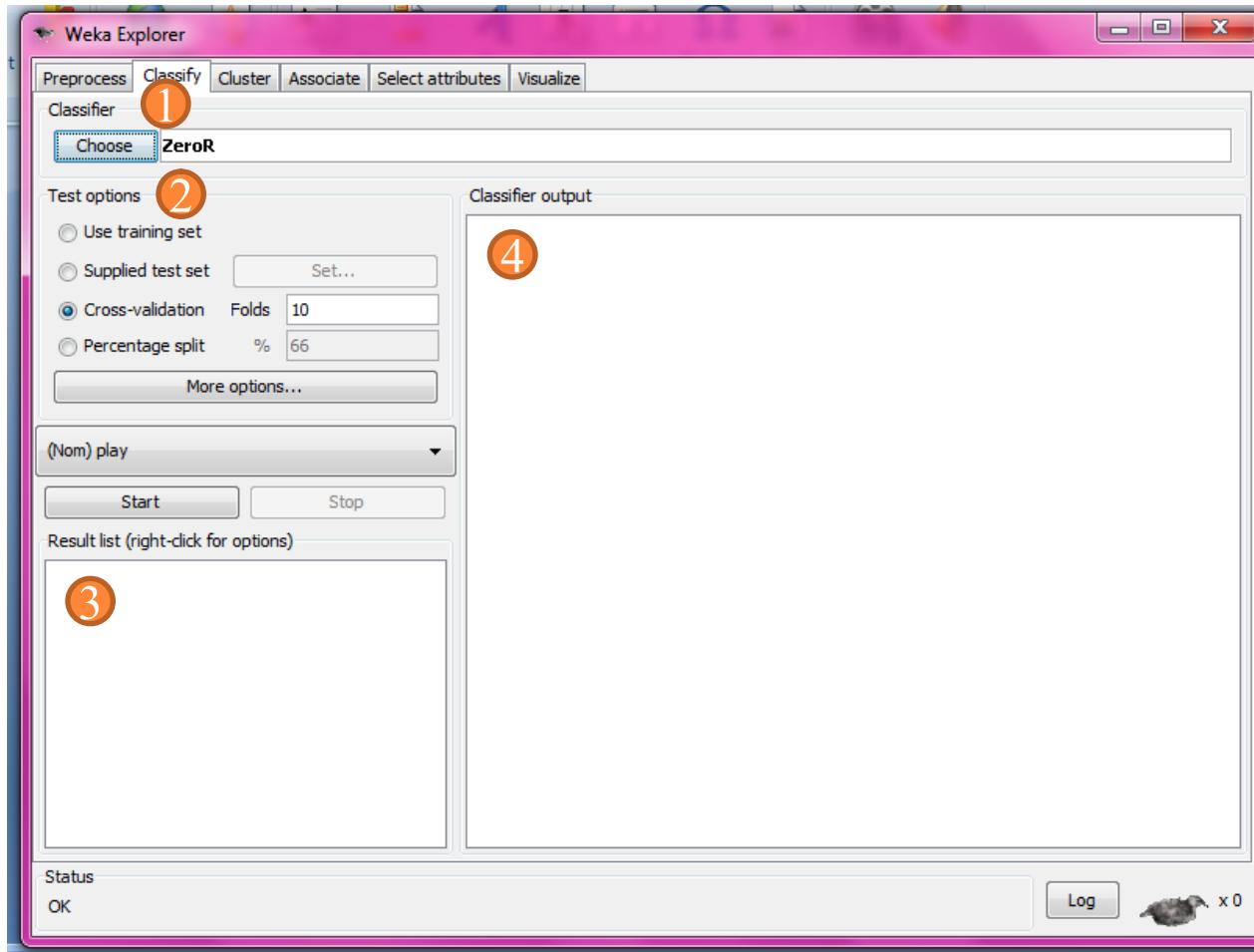
# CLASSIFICATION IN WEKA

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file.... > เลือกไฟล์ data\weather.arff



# CLASSIFICATION IN WEKA (CONT’)

- คลิกที่ tab Classify



# 1: CLASSIFIER

Weka Explorer

Preprocess Classify Cluster Associate Select attributes

Classifier

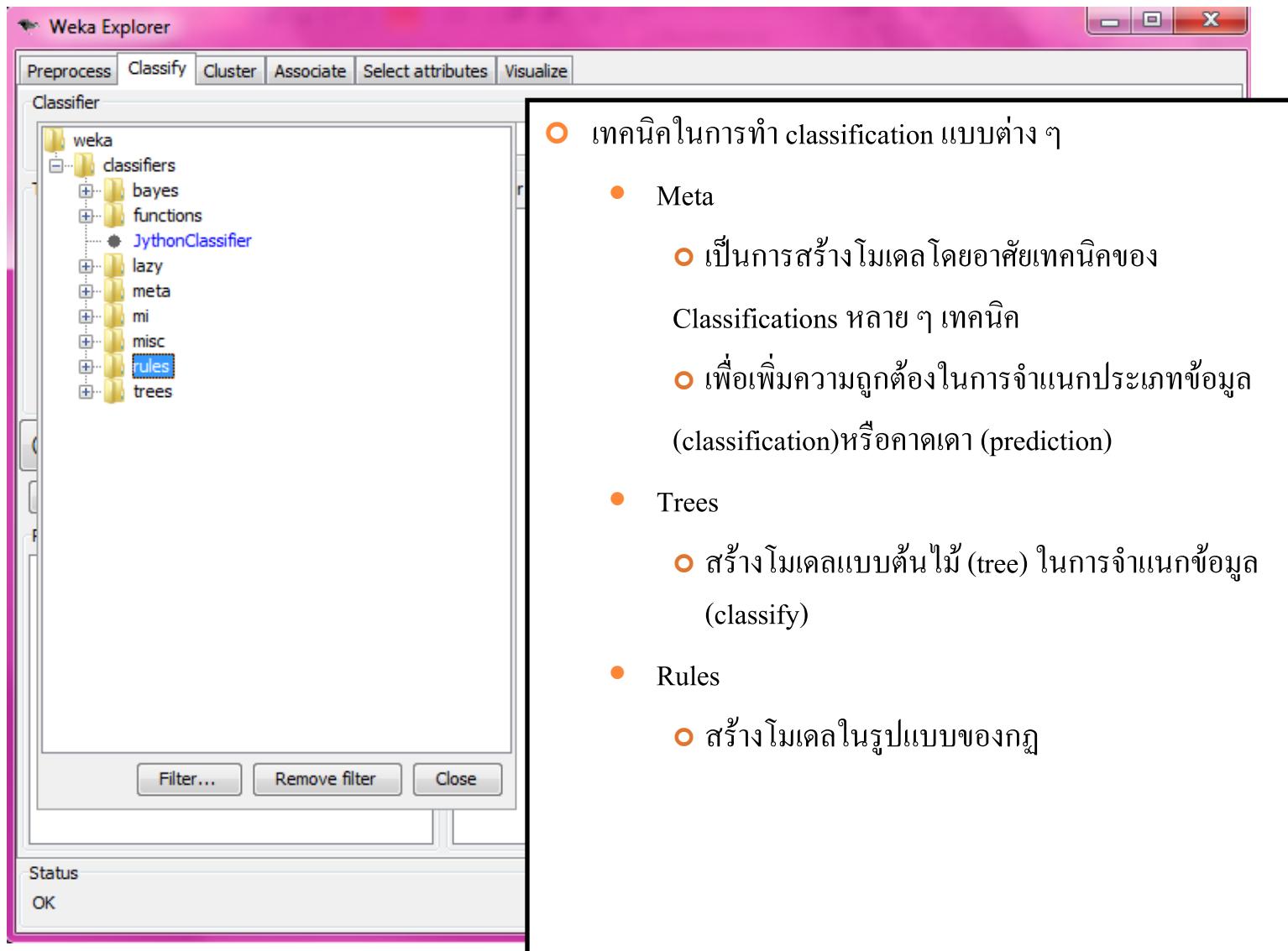
weka  
└ classifiers  
 └ rules

Filter... Remove filter Close

Status OK

- เทคนิคในการทำ classification แบบต่าง ๆ
  - Bayes
    - สร้างโมเดลโดยอาศัยการคำนวณความน่าจะเป็นของข้อมูลต่าง ๆ
  - Functions
    - สร้างโมเดลโดยอาศัยการคำนวณทางคณิตศาสตร์
    - โมเดลเป็นรูปแบบของสมการ
  - Lazy
    - ต่างจากเทคนิค classification แบบอื่น ๆ
    - ไม่มีการสร้างโมเดลไว้ก่อน
    - ใช้ข้อมูลเรียนรู้เพื่อจำแนกประเภทข้อมูลของข้อมูลใหม่ได้เลย

# 1: CLASSIFIER (CONT’)



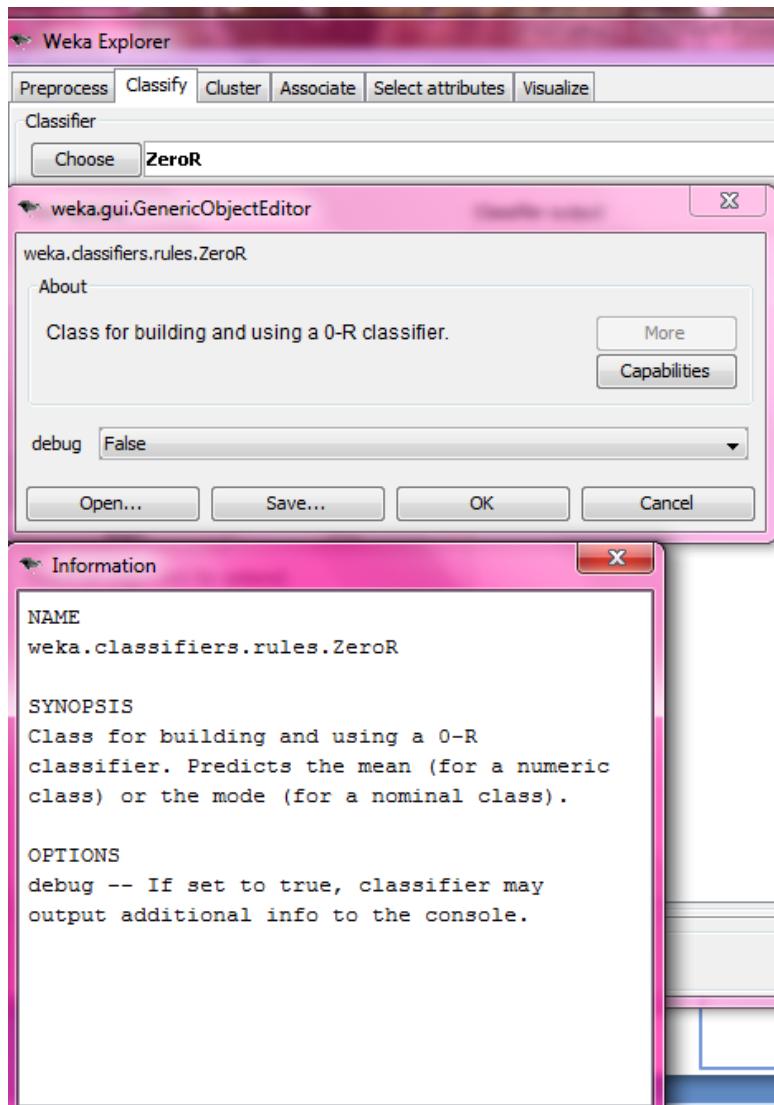
The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The left pane displays a tree view of classifiers under the 'weka.classifiers' package, including 'bayes', 'functions', 'lazy', 'meta', 'mi', 'misc', 'rules', and 'trees'. A callout box highlights the 'Classifier' section and provides the following text:

○ เทคนิคในการทำ classification !!แบบต่าง ๆ

- Meta
  - เป็นการสร้างโมเดลโดยอาศัยเทคนิคของ Classifications หลาย ๆ เทคนิค
  - เพื่อเพิ่มความถูกต้องในการจำแนกประเภทข้อมูล (classification) หรือคาดเดา (prediction)
- Trees
  - สร้างโมเดลแบบต้นไม้ (tree) 在การจำแนกข้อมูล (classify)
- Rules
  - สร้างโมเดลในรูปแบบของกฎ

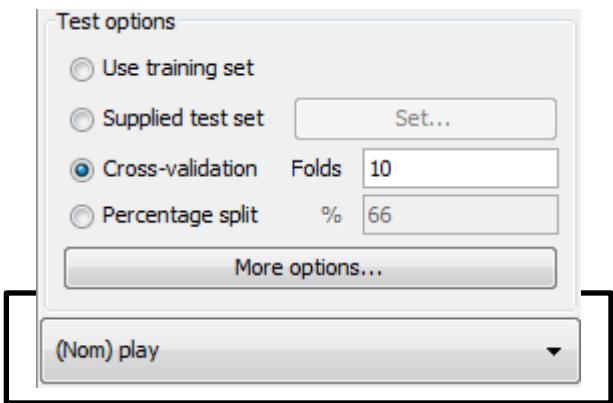
At the bottom of the interface, there are buttons for 'Filter...', 'Remove filter', and 'Close'. The status bar at the bottom left shows 'Status OK'.

# 1: CLASSIFIER (CONT’)



- คลิกที่บริเวณชื่อของเทคนิค
  - pragquhnāต่างให้กำหนดค่า พารามิเตอร์ ต่าง ๆ
  - More เพื่อดูรายละเอียดของพารามิเตอร์ ต่าง ๆ
  - Save...บันทึกค่าพารามิเตอร์ที่เซตไว้
  - Open...เปิดไฟล์ที่เก็บค่าพารามิเตอร์ที่เซตไว้

# 1: CLASSIFIER (CONT’)

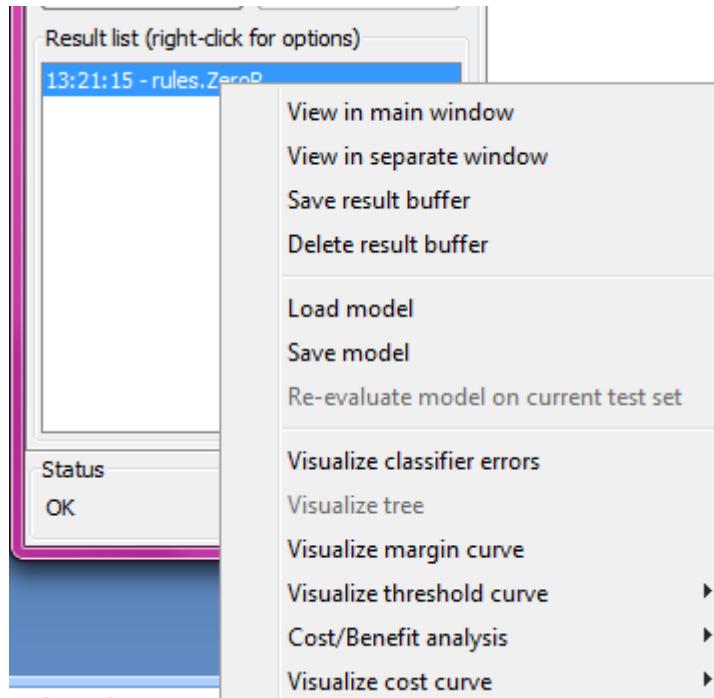


แอตทริบิวต์ที่จะใช้เป็นคลาส (class) ในการจำแนก ประเภทข้อมูล (ปกติจะเป็นแอตทริบิวต์สุดท้าย)

## ○ เลือกข้อมูลเพื่อใช้ทดสอบ

- Use training set
  - ใช้ข้อมูลเรียนรู้ทั้งหมดเพื่อเป็นตัวทดสอบ ประสิทธิภาพ
- Supplied test set
  - ใช้ข้อมูลใหม่ (unseen data) เพื่อทำการทดสอบ โมเดลที่สร้างขึ้น
- Cross-validation ( default 10 folds )
  - แบ่งข้อมูลเรียนรู้ออกเป็น  $k$  ส่วนเท่า ๆ กัน (folds) เพื่อใช้ในการทดสอบ
  - ระบุจำนวน  $k$  ในช่อง Folds
- Percentage split
  - แบ่งข้อมูลเรียนรู้ออกเป็น  $x\%$  เพื่อใช้ในการสร้าง โมเดล ส่วนที่เหลือใช้เป็นข้อมูลทดสอบ
  - กำหนดในช่อง %

# 3: RESULT LIST



## ○ แสดงผลการทำงานครั้งก่อน ๆ

- แสดงเวลา
- เทคนิคที่ใช้
- คลิกขวาที่ผลจะแสดง option เพิ่มเติม
  - Load model เปิดโมเดลที่ได้เคยเก็บไว้
  - Save model บันทึกโมเดลไว้ใช้ในคราวต่อไป
  - Visualize tree (มีเฉพาะเทคนิค J48)  
ใช้แสดงโมเดล (decision tree) ในรูปแบบต้นไม้

# 4: CLASSIFIER OUTPUT (2)

```
Classifier output
==== Run information ====
Scheme: weka.classifiers.rules.ZeroR
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation

==== Classifier model (full training set) ====
ZeroR predicts class value: yes

Time taken Classifier output
Time taken
==== Stratif
==== Summary
-----.
Correctly Classified Instances      9      64.2857 %
Incorrectly Classified Instances   5      35.7143 %
Kappa statistic                   0
Mean absolute error               0.4762
Root mean squared error          0.4934
Relative absolute error           100 %
Root relative squared error      100 %
Total Number of Instances         14

==== Detailed Accuracy By Class ====
           TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area
           1        1        0.643     1        0.783    0.178
           0        0        0          0        0        0.178
Weighted Avg.  0.643    0.643    0.413    0.643    0.503    0.178

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

## แสดงผลการจำแนกประเภทข้อมูล (classify)

### Run information

- แสดงรายละเอียดของข้อมูลที่ใช้
- เทคนิคและพารามิเตอร์ที่เลือก
- การทดสอบประสิทธิภาพ

### Classifier model (full training set)

- สร้างโมเดล เช่น tree ที่สร้างได้จากข้อมูลเรียนรู้ทั้งหมด

### Summary

- ค่าความถูกต้อง (Accuracy)
  - กรณีที่คลาสเป็นข้อมูลแบบประเภท Nominal
- ค่าความคลาดเคลื่อน (error)
  - กรณีที่คลาสเป็นข้อมูลแบบตัวเลข (numeric)

# 4: CLASSIFIER OUTPUT

```
Classifier output
==== Run information ====
Scheme: weka.classifiers.rules.ZeroR
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation

==== Classifier model (full training set) ====
ZeroR predicts class value: yes

Time taken Classifier output
Time taken    correctly classified instances      9        64.2857 %
==== Stratified Summary
Incorrectly Classified Instances      5        35.7143 %
Kappa statistic                      0
Mean absolute error                  0.4762
Root mean squared error              0.4934
Relative absolute error              100 %
Root relative squared error         100 %
Total Number of Instances           14

==== Detailed Accuracy By Class ====
          TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area
          1         1         0.643      1         0.783      0.178
          0         0         0          0         0          0.178
Weighted Avg.  0.643     0.643     0.413     0.643     0.503     0.178

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

## แสดงผลการจำแนกประเภทข้อมูล (classify)

### Detailed accuracy By Class

- ค่าทางสถิติของเมื่อแยกตามคลาส
- TP Rate (True Positive) ค่าที่ทายถูก
- FP Rate (False Positive) ค่าที่ทายผิด

### Confusion Matrix

- colum : ค่าที่ทำนายได้
  - row : ค่าจริง
- ข้อมูล 2 ส่วนท้ายนี้จะไม่มีเมื่อคลาสเป็นตัวเลข !!!

## 4: CLASSIFIER OUTPUT (3)

### Classifier output

```
== Stratified cross-validation ==
== Summary ==

Correctly Classified Instances      9          64.2857 %
Incorrectly Classified Instances   5          35.7143 %
Kappa statistic                   0
Mean absolute error              0.4762
Root mean squared error          0.4934
Relative absolute error           100       %
Root relative squared error      100       %
Total Number of Instances        14
```

### == Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	1	0.643	1	0.783	0.178	yes
0	0	0	0	0	0	0.178	no
Weighted Avg.	0.643	0.643	0.413	0.643	0.503	0.178	

### == Confusion Matrix ==

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

# CONFUSION MATRIX

		Predicted	
		A	B
Actual	A	8	2
	B	3	4

True Positive

○ พิจารณาที่ class A

- True Positive (TP) คือจำนวนที่ classify ถูกว่าเป็น class A

# CONFUSION MATRIX (2)

Actual	Class	Predicted	
		A	B
A	A	8	2
B	B	3	4

False Positive

## พิจารณาที่ class A

- True Positive (TP) คือ จำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)

# CONFUSION MATRIX (3)

Actual	Class	Predicted	
		A	B
A	A	8	2
B	B	3	4

True Positive

TP Rate = TP/#data in class X

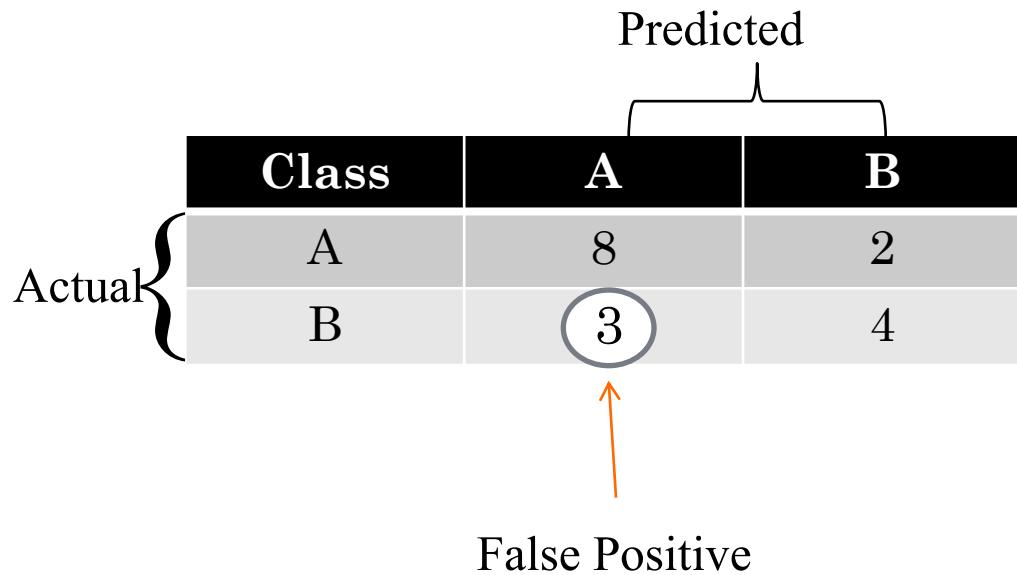
## พิจารณาที่ class A

- True Positive (TP) คือ จำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)
- TP Rate =  $8/10 = 0.8$

# CONFUSION MATRIX (4)

Actual	Class	Predicted	
		A	B
A	A	8	2
B	B	3	4

False Positive



$$\text{TP Rate} = \text{TP}/\#\text{data in class X}$$

$$\text{FP Rate} = \text{TP}/\#\text{data in class X}$$

## พิจารณาที่ class A

- True Positive (TP) คือ จำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)
- TP Rate =  $8/10 = 0.8$
- FP Rate =  $3/10 = 0.428$

# CONFUSION MATRIX (5)

Actual	Predicted	
	A	B
Class	A	B
A	8	2
B	3	4

$$\text{TP Rate} = \text{TP}/\#\text{data in class } X$$

$$\text{FP Rate} = \text{FP}/\#\text{data in class } \sim X$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

○ พิจารณาที่ class A

- True Positive (TP) คือจำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)
- TP Rate =  $8/10 = 0.8$
- FP Rate =  $3/10 = 0.428$
- Precision =  $8/(8+3) = 0.727$

# CONFUSION MATRIX (6)

Actual	Predicted	
	A	B
Class	A	B
A	8	2
B	3	4

$$\text{TP Rate} = \text{TP}/\#\text{data in class } X$$

$$\text{FP Rate} = \text{FP}/\#\text{data in class } \sim X$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{Recall} = \text{TP Rate}$$

○ พิจารณาที่ class A

- True Positive (TP) คือจำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)
- TP Rate =  $8/10 = 0.8$
- FP Rate =  $3/10 = 0.428$
- Precision =  $8/(8+3) = 0.727$
- Recall =  $8/(8+2) = 0.8 = \text{TP Rate}$

# CONFUSION MATRIX (7)

Actual	Predicted	
	A	B
Class	A	B
A	8	2
B	3	4

$$\text{TP Rate} = \text{TP}/\#\text{data in class } X$$

$$\text{FP Rate} = \text{FP}/\#\text{data in class } \sim X$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

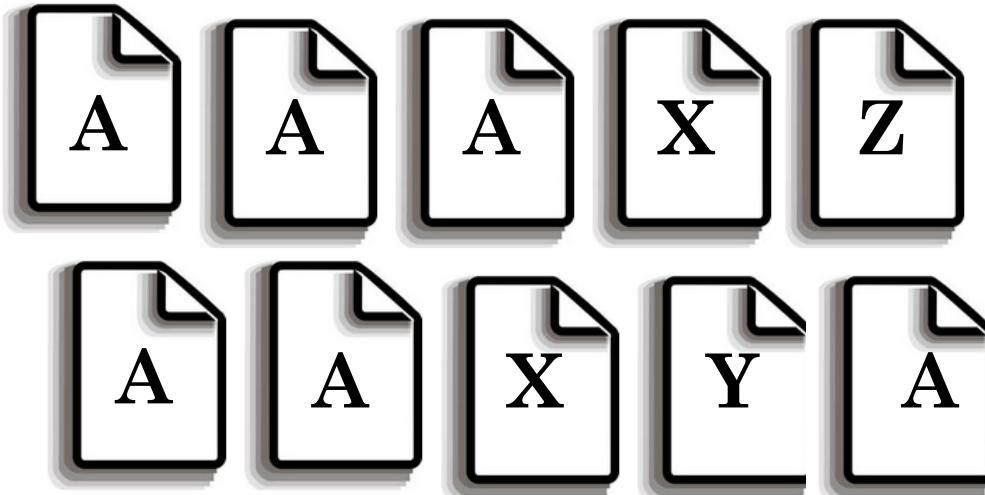
$$\text{Recall} = \text{TP Rate}$$

○ พิจารณาที่ class A

- True Positive (TP) คือจำนวนที่ classify ถูกว่าเป็น class A
- False Positive (FP) คือ จำนวนที่ classify ผิดว่าเป็น class A (ซึ่งจริง ๆ แล้วเป็น class B)
- TP Rate =  $8/10 = 0.8$
- FP Rate =  $3/10 = 0.428$
- Precision =  $8/(8+3) = 0.727$
- Recall =  $8/(8+2) = 0.8 = \text{TP Rate}$
- Accuracy =  $(8+4)/(8+2+3+4) = 0.706$
- =70.6 %

$$\text{Accuracy} = \frac{\#\text{correctly classified instances}}{\#\text{instances}}$$

# PRECISION & RECALL



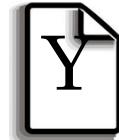
# PRECISION & RECALL(2)



= 6 ฉบับ



= 2 ฉบับ

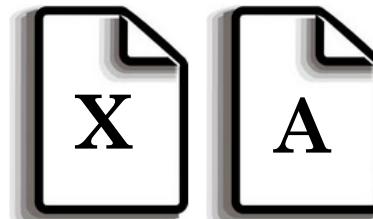
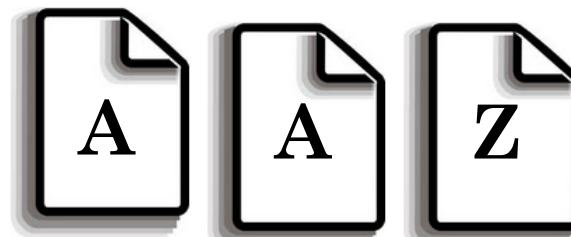


= 1 ฉบับ



= 1 ฉบับ

A screenshot of a Google search results page. The search bar contains the letter 'A'. Below the search bar are two buttons: 'ค้นหาด้วย Google' (Search with Google) and 'ติดตาม ค้นแล้วเจอเลย' (Follow-up, Found it). The main search results area has a red underline under the word 'เว็บ' (Web). Below the results, a text box displays: 'ผลการค้นหาประมาณ 25,270,000,000 รายการ (0.23 วินาที)' (Approximate search results: 25,270,000,000 items found in 0.23 seconds).



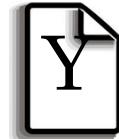
# PRECISION & RECALL(3)



= 6 ฉบับ



= 2 ฉบับ



= 1 ฉบับ



= 1 ฉบับ



- Precision =  $\frac{\text{จำนวนเอกสารที่ค้นเจอและมีเนื้อหาสัมพันธ์กับคำค้น}}{\text{จำนวนเอกสารทั้งหมดที่สืบค้นได้}}$

# PRECISION & RECALL(4)



= 6 ฉบับ



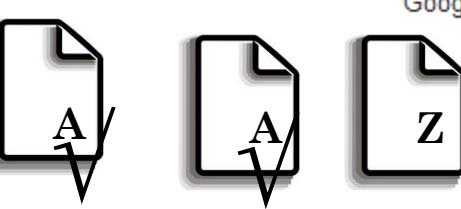
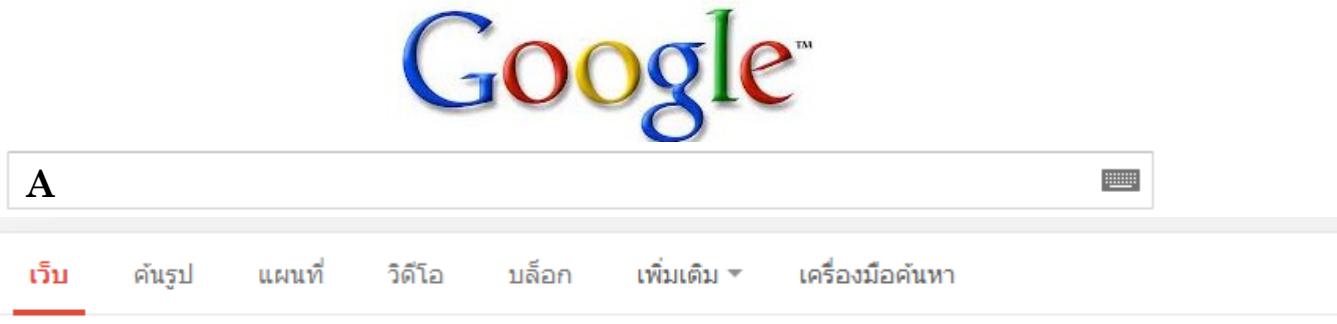
= 2 ฉบับ



= 1 ฉบับ



= 1 ฉบับ

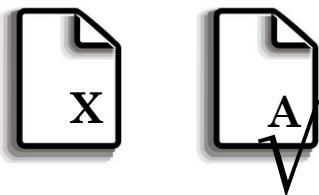


• Recall = จำนวนเอกสารที่ค้นเจอและมีเนื้อหาสัมพันธ์กับคำค้น

จำนวนเอกสารทั้งหมดที่มีเนื้อหาสัมพันธ์กับคำค้นได้

$$= 3/6$$

$$= 0.5$$



## 4 : CLASSIFIER OUTPUT(3)

```
Classifier output

    === Stratified cross-validation ===
    === Summary ===

    Correctly Classified Instances      9          64.2857 %
    Incorrectly Classified Instances   5          35.7143 %
    Kappa statistic                   0

    Mean absolute error              0.4762
    Root mean squared error         0.4934
    Relative absolute error          100        %
    Root relative squared error     100        %
    Total Number of Instances       14

    === Detailed Accuracy By Class ===

           TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
             1          1        0.643      1        0.783      0.178    yes
             0          0        0           0        0           0.178    no
    Weighted Avg.   0.643      0.643      0.413      0.643      0.503      0.178

    === Confusion Matrix ===

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

# KAPPA STATISTIC - MEASURE OF AGREEMENT



# KAPPA STATISTIC (2)

- สมาชิกบ้าน AF 12 คน
- Raters มี 2 คน คือ ครูเงาะ และครูเมย์
- คุณครูทำแบบสอบถามเพื่อพิจารณาว่า AF ไหนควรอยู่ในบ้านต่อหรือควรออกจากบ้าน
  - Yes = ควรอยู่ในบ้านต่อ
  - No = ควรออกจากบ้าน
- พิจารณาว่าความเห็นของครูทั้งสองสอดคล้องกันหรือไม่

		ครูเมย์	
		Yes	No
ครูเงาะ	Yes	6	1
	No	2	3



# KAPPA STATISTIC (3)

គ្រូមេរោគ

		Yes	No
គ្រូការណ៍	Yes	6	1
	No	2	3
Total		8	4

- $\text{Pr}(a) = \text{observed percentage agreement}$

$$= (6+3)/12 = 0.75$$

- $\text{Pr}(e) = \text{obed percentage agreement} = (0.3889 + 0.1389) = 0.5278$  គាំនុវត្តនាក់
  - គ្រូមេរោគ vote “Yes” = 8 , Vote “No” = 4 ដែលនៃន័យ ឯកាសពួកខ្លួន Yes គិតបាន  $8/12 = 0.6667$
  - គ្រូការណ៍ vote “Yes” = 7 , Vote “No” = 5 ដែលនៃន័យ ឯកាសពួកខ្លួន Yes គិតបាន  $7/12 = 0.5833$
  - ឯកាសពួកខ្លួនទាំងអស់ Vote “Yes” គិតបាន  $0.6667 \times 0.5833 = 0.3889$
  - ឯកាសពួកខ្លួនទាំងអស់ Vote “No” គិតបាន  $(1 - 0.6667) \times (1 - 0.5833) = 0.1389$



# KAPPA STATISTIC (4)

ករណីមេរី

		Yes	No
ករណីនៅក្រោម	Yes	6	1
	No	2	3
Total	8	4	

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$$= (0.75 - 0.5278) / (1 - 0.5278)$$

$$= 0.4706$$

**Note :** Fleiss's guidelines characterize kappas over .75 as excellent, .40 to .75 as fair to good, and below .40 as poor.

## 4: CLASSIFIER OUTPUT (4)

```
Classifier output

==== Stratified cross-validation ====
==== Summary ====

    Correctly Classified Instances      9          64.2857 %
    Incorrectly Classified Instances   5          35.7143 %
    Kappa statistic                   0
    Mean absolute error              0.4762
    Root mean squared error          0.4934
    Relative absolute error          100      %
    Root relative squared error     100      %
    Total Number of Instances        14

==== Detailed Accuracy By Class ====

           TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
           1          1          0.643       1          0.783      0.178     yes
           0          0          0            0          0            0.178     no
    Weighted Avg.  0.643     0.643       0.413     0.643     0.503      0.178

==== Confusion Matrix ====

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

ROC Area	Class
0.178	yes
0.178	no
0.178	

# ROC CURVE

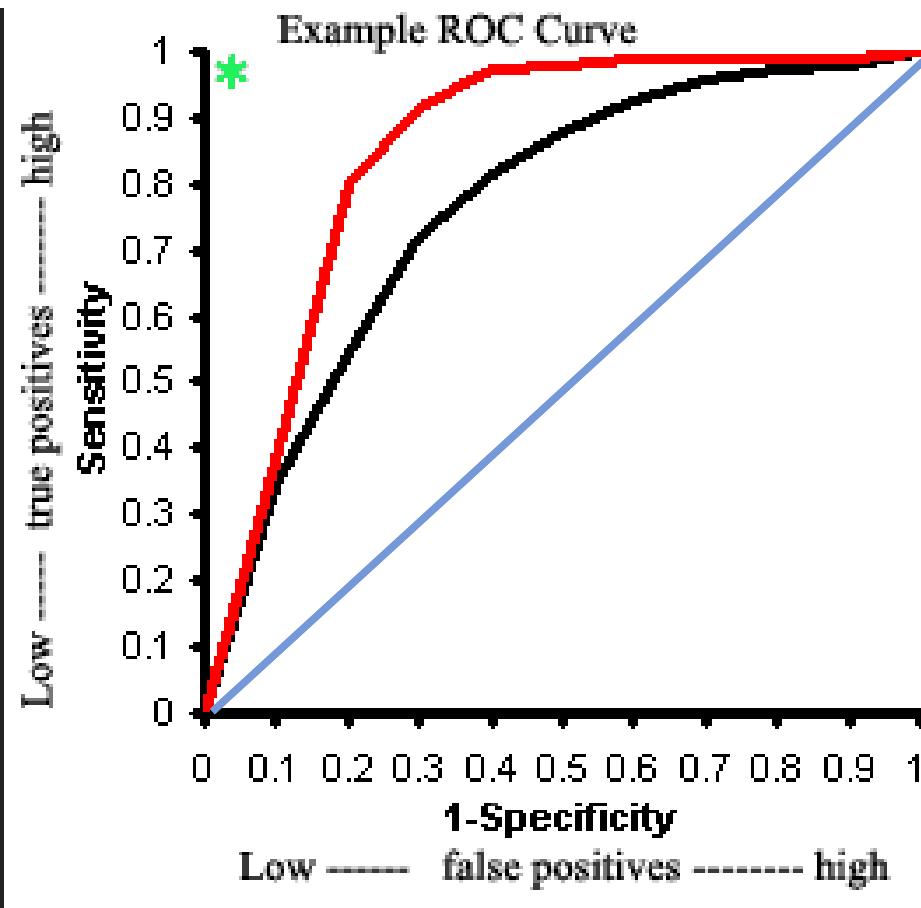
- Class ที่กำลังพิจารณา (Positive Class) คือ Class “Yes”

Rank	Predicted Probability	Rank	Rank	Predicted Probability	Rank
1	0.95	Yes	11	0.77	No
2	0.93	Yes	12	0.76	Yes
3	0.93	No	13	0.73	Yes
4	0.88	Yes	14	0.65	No
5	0.86	Yes	15	0.63	Yes
6	0.85	Yes	16	0.58	No
7	0.82	Yes	17	0.56	Yes
8	0.80	Yes	18	0.49	No
9	0.80	No	19	0.48	Yes
10	0.79	Yes	...	...	...



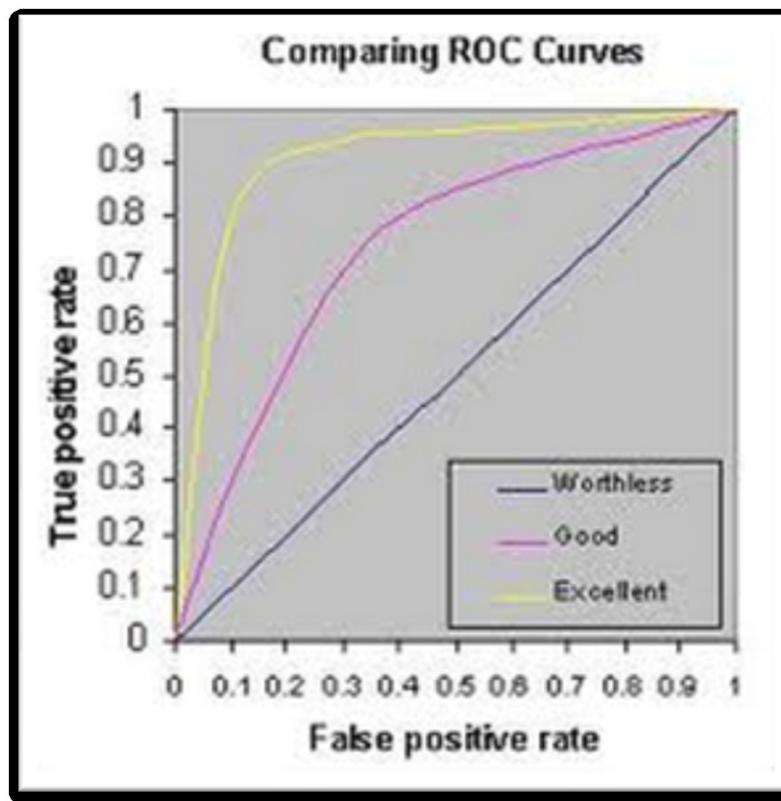
# ROC CURVE

- Plot ค่า TP vs FP



# ROC CURVE

- พื้นที่ใต้กราฟ ROC (AUC-Area Under Curve) แสดง performance ในการจำแนกคลาสของ Model



# DECISION TREE APPLICATIONS

- ตอบคำถามที่ต้องการจัดจำแนกประเภทข้อมูล (classification) ที่ต้องการเข้าใจประกอบ
  - ใช้ในการพิจารณาให้สินเชื่อแก่บุคคลต่าง ๆ
  - ใช้ในการทำนายว่าลูกค้าคนไหนบ้างที่มีโอกาสจะยกเลิกการใช้บริการและเหตุผลอะไร



## EXAMPLE 2 : DECISION TREE

- Weather.arff

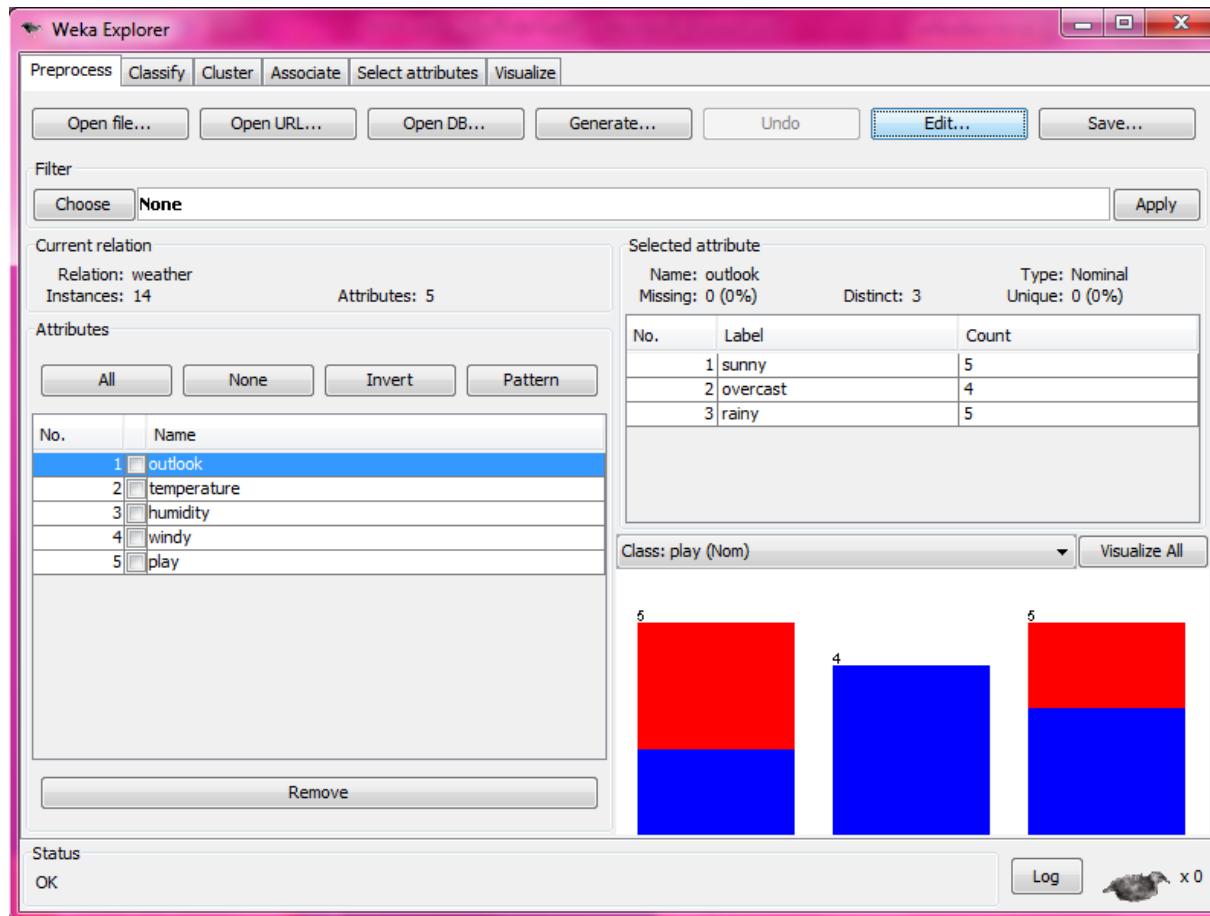
outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
sunny	85.0	85.0	FALSE	no
sunny	80.0	90.0	TRUE	no
overcast	83.0	86.0	FALSE	yes
rainy	70.0	96.0	FALSE	yes
rainy	68.0	80.0	FALSE	yes
rainy	65.0	70.0	TRUE	no
overcast	64.0	65.0	TRUE	yes
sunny	72.0	95.0	FALSE	no

ข้อมูลที่ใช้ในการวิเคราะห์

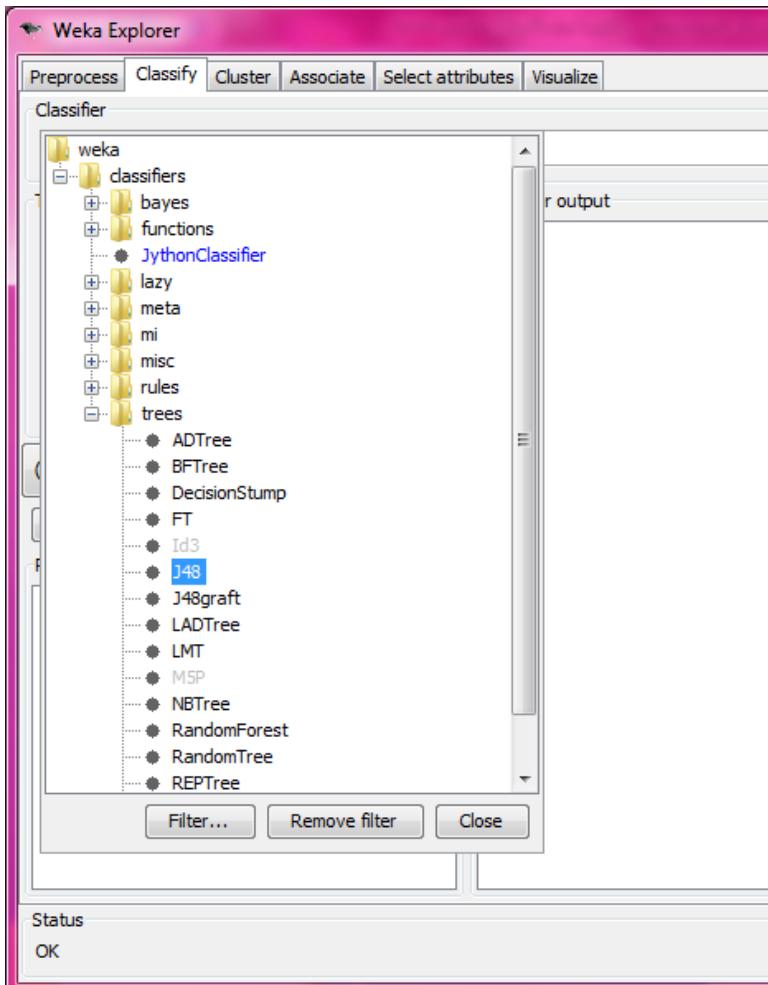
ค่าที่ต้องการคาดเดา (class)

# EXAMPLE 2 : DECISION TREE (2)

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ data\weather.arff



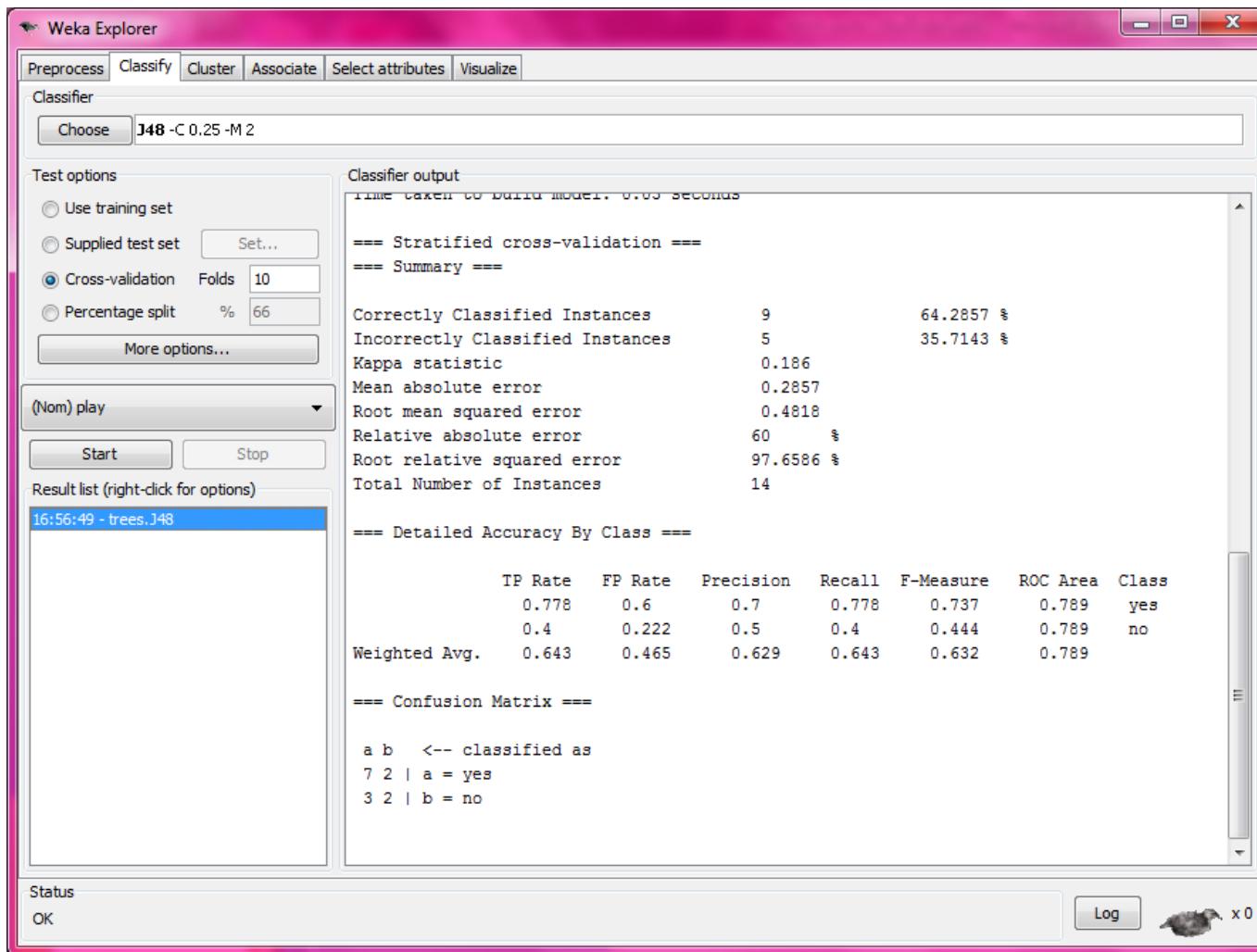
# EXAMPLE 2 : DECISION TREE (3)



- คลิกที่ tab Classify
- กดปุ่ม Choose
  - เลือก Classifiers
  - เลือก trees
  - เลือก J48
- กดปุ่ม Start

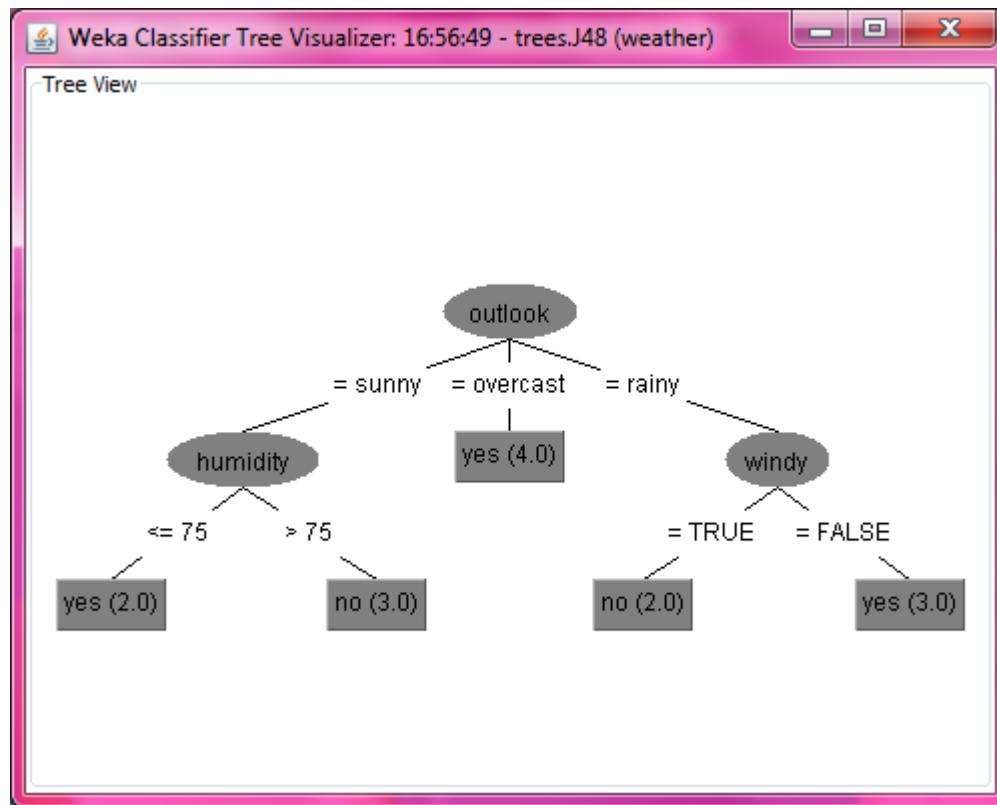
# EXAMPLE 2 : DECISION TREE (4)

○ ผลที่ได้

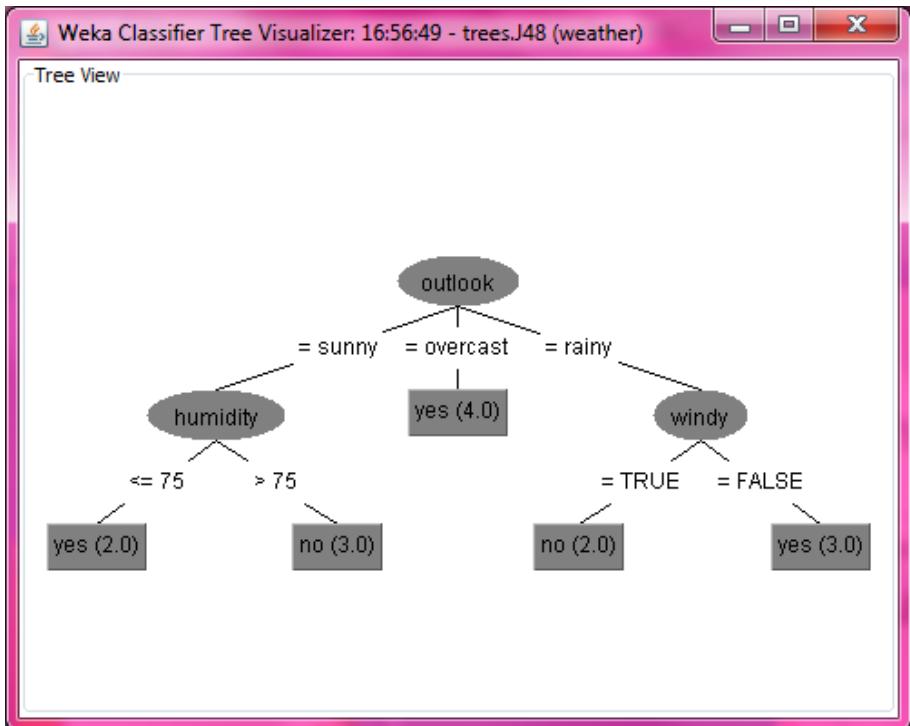


## EXAMPLE 2 : DECISION TREE (5)

- คลิกขวาที่โมเดลในช่อง Result list
- เลือก Visualize tree



## EXAMPLE 2 : DECISION TREE (6)

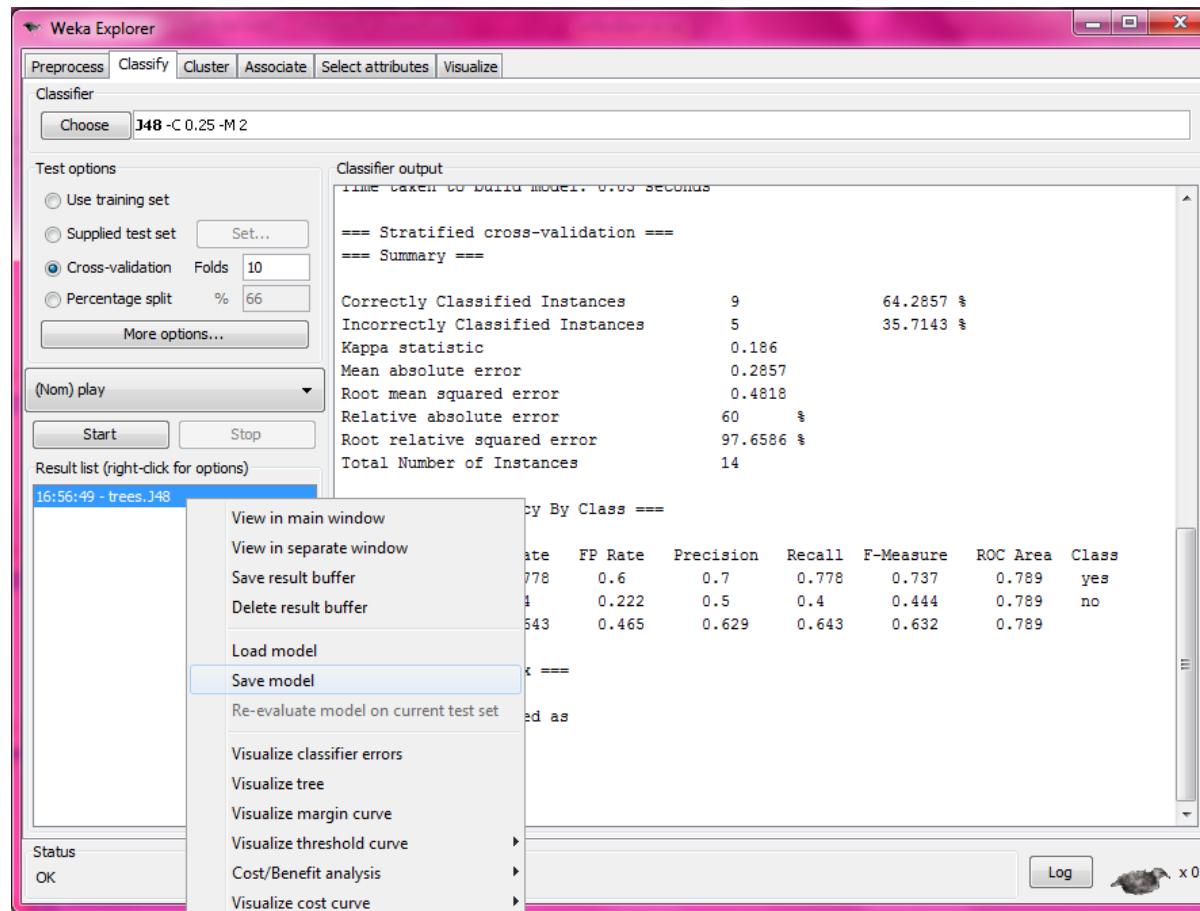


○ กฎการตัดสินใจจาก decision tree

- IF outlook = sunny AND humidity  $\leq 75$   
Then play  
IF outlook = sunny AND  
humidity  $\leq 75$  Then play
- IF outlook = sunny AND humidity  $> 75$   
Then no play
- IF outlook = overcast Then play
- IF outlook = rainy AND windy = TRUE  
Then play
- IF outlook = rainy AND windy = FALSE  
Then no play

# EXAMPLE 2 : DECISION TREE (7)

- บันทึกโมเดลที่ได้สร้างไว้



คลิกขวาเลือก

Save model

## EXAMPLE 2 : DECISION TREE (8)

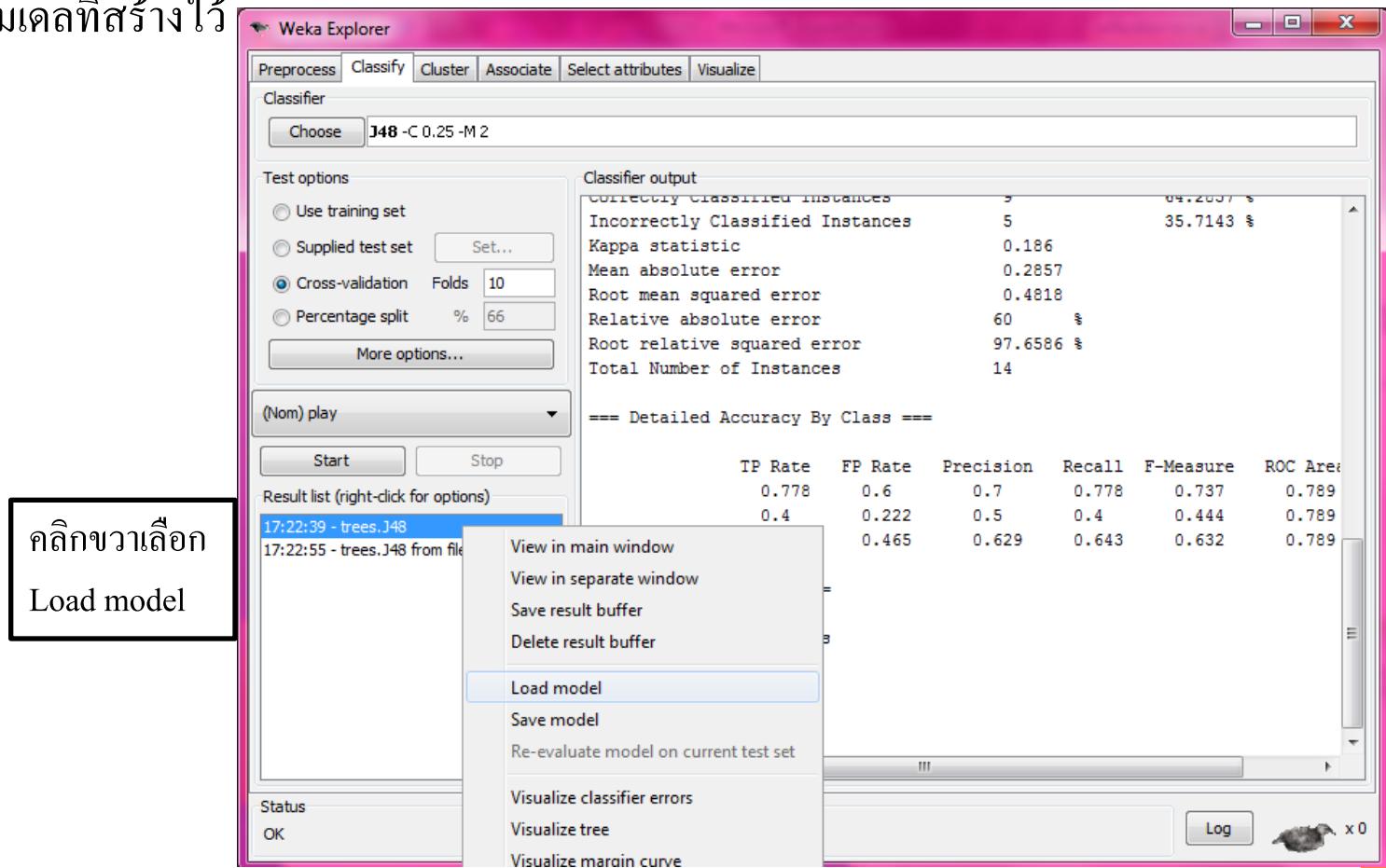
- Testing file
- Weather-test.arff

outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	Play
sunny	85.0	85.0	FALSE	?
sunny	80.0	90.0	TRUE	?
overcast	83.0	86.0	FALSE	?
rainy	70.0	96.0	FALSE	?
rainy	68.0	80.0	FALSE	?
rainy	65.0	70.0	TRUE	?
overcast	64.0	65.0	TRUE	?
sunny	72.0	95.0	FALSE	?

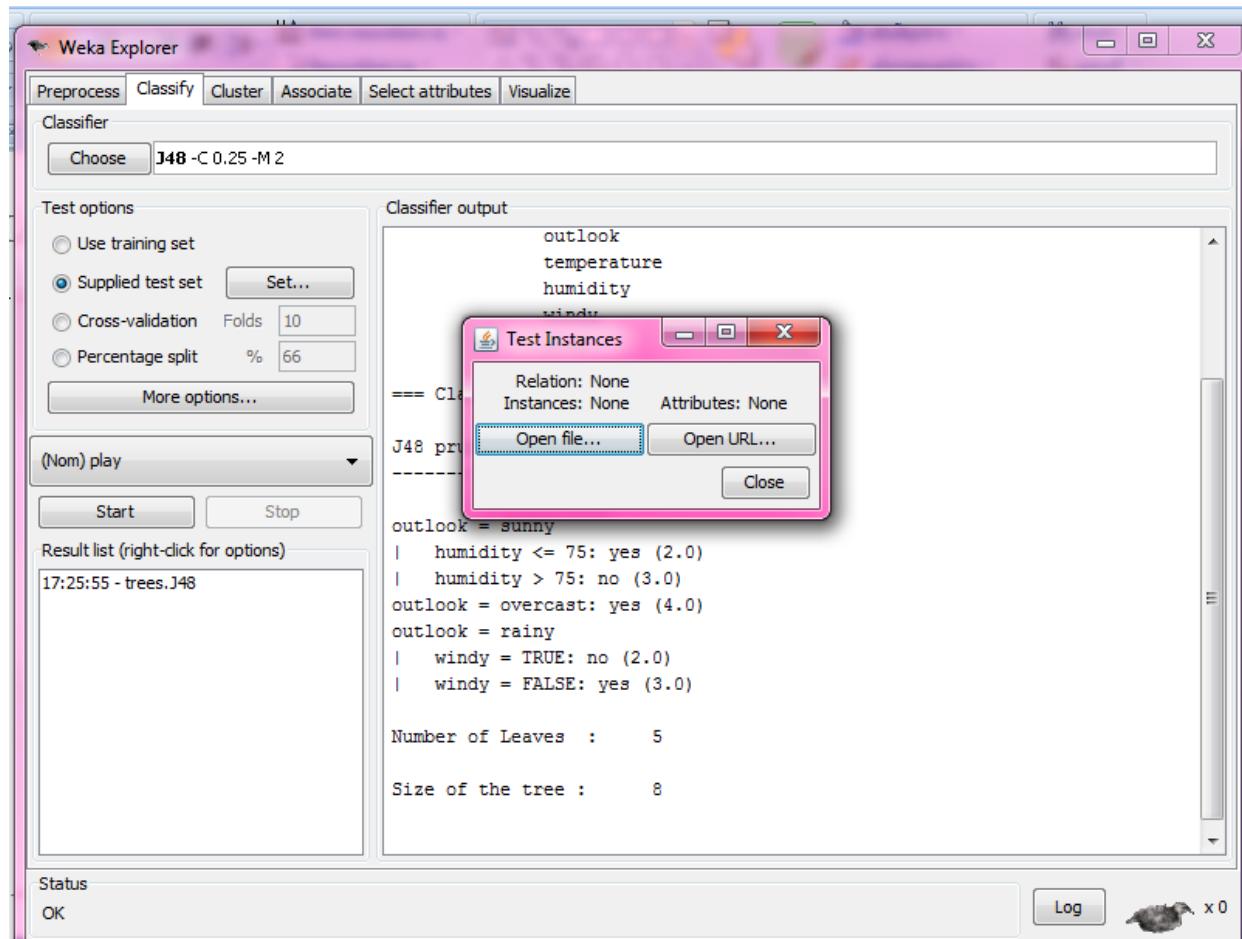


# EXAMPLE 2 : DECISION TREE (9)

- เปิดโมเดลที่สร้างไว้

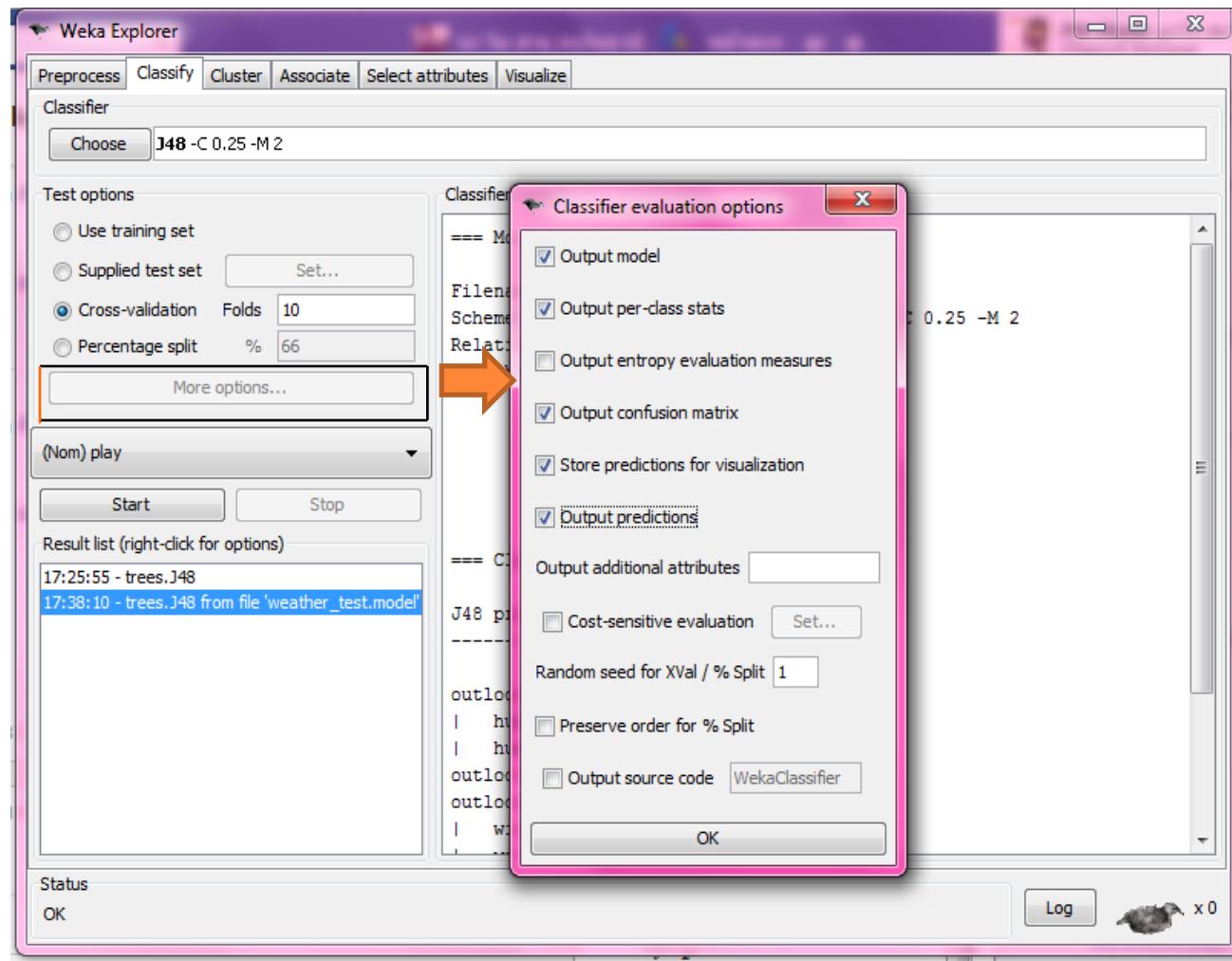


# EXAMPLE 2 : DECISION TREE (11)



- เลือก Supplied test set
- กดปุ่ม Set... เลือกไฟล์ Weather-test.arff

# EXAMPLE 2 : DECISION TREE (11)

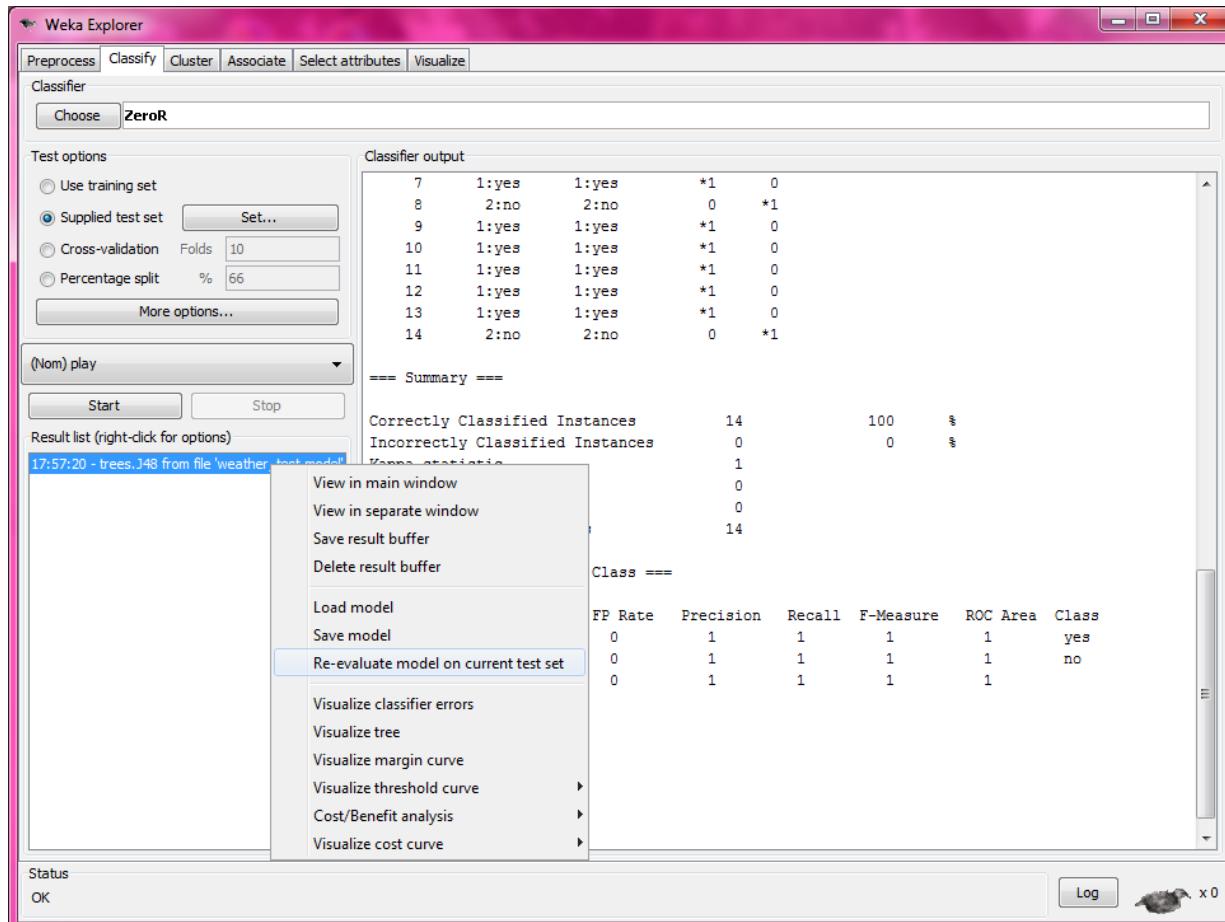


- เลือก Supplied test set
- กดปุ่ม Set... เลือกไฟล์

Weather-test.arff

- กดปุ่ม More options...
  - เลือก output predictions

# EXAMPLE 2 : DECISION TREE (12)



- เลือก Supplied test set
- กดปุ่ม Set... เลือกไฟล์

Weather-test.arff

- กดปุ่ม More options
- เลือก output predictions
- คลิกขวาที่โมเดลเลือก Re-evaluate model on current test set

# EXAMPLE 3 : CUSTOMERS

○ ผลที่ได้

ผลที่ทำนายได้

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'ZeroR'. In the 'Test options' section, 'Supplied test set' is selected with 10 folds. The 'Classifier output' pane displays the predictions for a test set of 14 instances. The first 13 instances are correctly classified (predicted 'no'), while instance 14 is predicted 'yes'. The 'Summary' section shows 14 correctly classified instances and 0 incorrectly classified instances.

inst#	actual	predicted	error	probability distribution
1	2: no	2: no	0	* 1
2	2: no	2: no	0	* 1
3	1: yes	1: yes	* 1	0
4	1: yes	1: yes	* 1	0
5	1: yes	1: yes	* 1	0
6	2: no	2: no	0	* 1
7	1: yes	1: yes	* 1	0
8	2: no	2: no	0	* 1
9	1: yes	1: yes	* 1	0
10	1: yes	1: yes	* 1	0
11	1: yes	1: yes	* 1	0
12	1: yes	1: yes	* 1	0
13	1: yes	1: yes	* 1	0
14	2: no	2: no	0	* 1

Instances: unknown (yet). Reading incrementally  
Attributes: 5

==== Predictions on test set ===

inst#, actual, predicted, error, probability distribution

==== Summary ===

Correctly Classified Instances 14 100  
Incorrectly Classified Instances 0 0  
Kappa statistic 1  
Mean absolute error 0

Status OK

# EXAMPLE 3 : CUSTOMERS

- แก้ไขไฟล์ customer\_wmissing.arff ที่เคยสร้างไว้ใน Lab 4-3

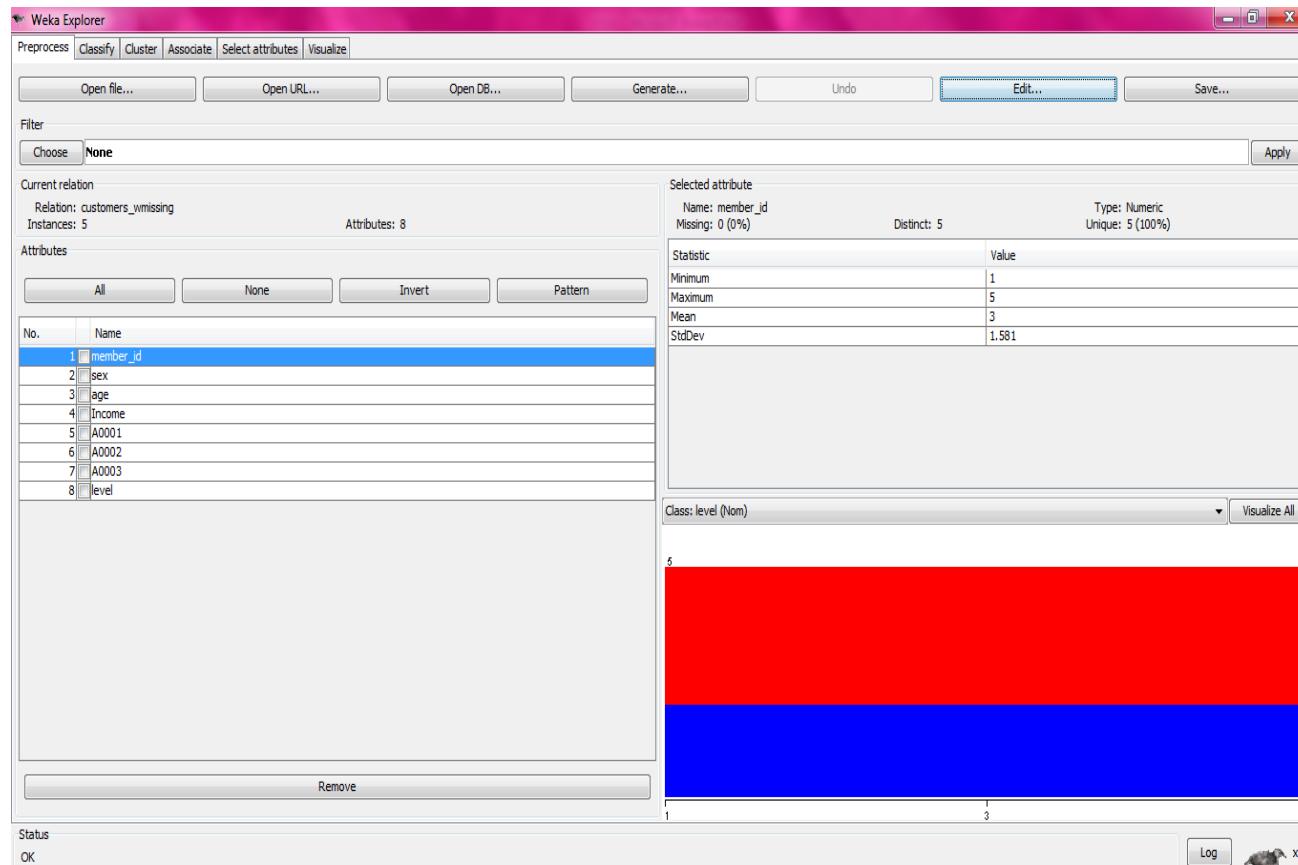
หมายเลข สมาชิก	เพศ	อายุ	รายได้	สินค้า A	สินค้า B	สินค้า C	ระดับของ สมาชิก
1	ชาย	20	12,000	0	0	0	D
2	หญิง	18	7,000	1	1	1	A
3	หญิง	35	35,000	0	0	0	D
4	เด็ก	16	4,000	1	1	0	A
5	หญิง	300	20,000	0	0	0	D

```
%customers_wmissing.arff
@relation customers_wmissing
%The following is attributes
@attribute member_id real
@attribute sex {male, female}
@attribute age real
@attribute Income numeric
@attribute A0001 {0,1}
@attribute A0002 {0,1}
@attribute A0003 {0,1}
@attribute level {A,D}
%The following is attributes
@data
1,male,20,12000,0,0,0,D
2,female,18,7000,1,1,1,A
3,female,35,35000,0,0,0,D
4,?,16,4000,1,1,0,A
5,female,?,20000,0,0,0,D
```



# EXAMPLE 3 : CUSTOMERS (2)

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ customers\_wmissing.arff



# EXAMPLE 3 : CUSTOMERS (3)

- กดปุ่ม Choose > เลือก filters > เลือก unsupervised > เลือก attribute > เลือก ReplaceMissingValues
- กดปุ่ม Apply
- ดูข้อมูลที่แก้ไขโดยการกดปุ่ม Edit...



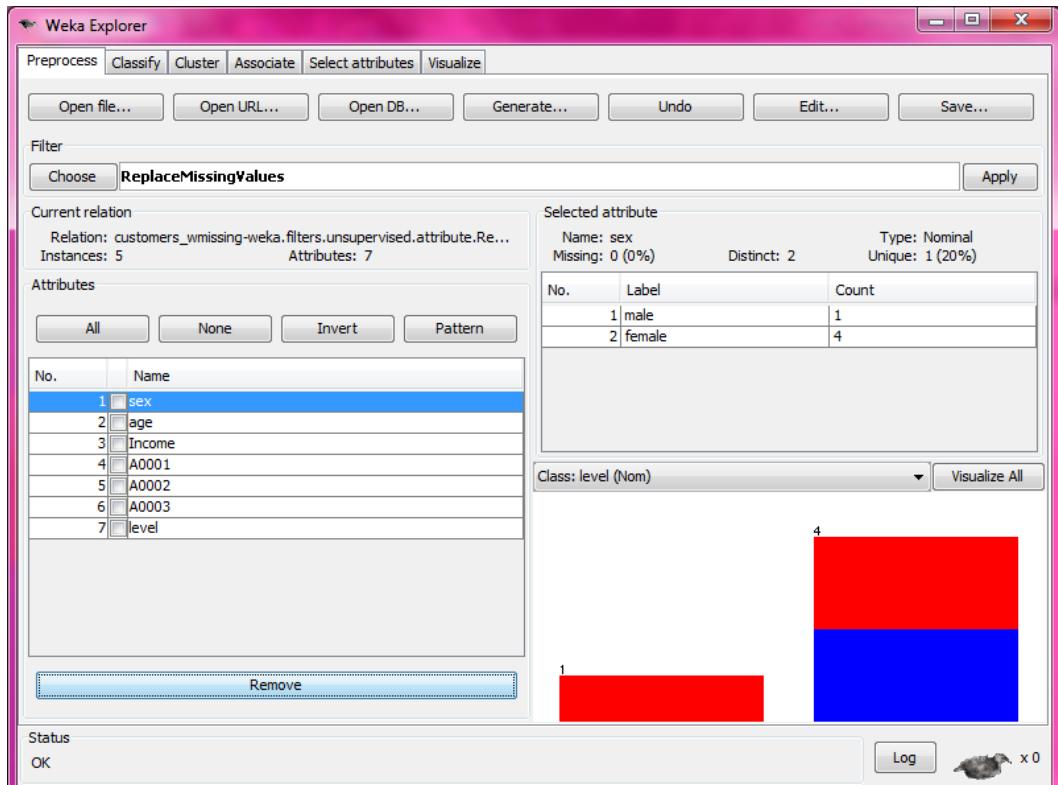
Viewer

Relation: customers\_wmissing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters...

No.	member_id Numeric	sex Nominal	age Numeric	Income Numeric	A0001 Nominal	A0002 Nominal	A0003 Nominal	level Nominal
1	1.0	male	20.0	12000.0	0	0	0	D
2	2.0	female	18.0	7000.0	1	1	1	A
3	3.0	female	35.0	35000.0	0	0	0	D
4	4.0	female	16.0	4000.0	1	1	0	A
5	5.0	female	22.25	20000.0	0	0	0	D

Undo OK Cancel

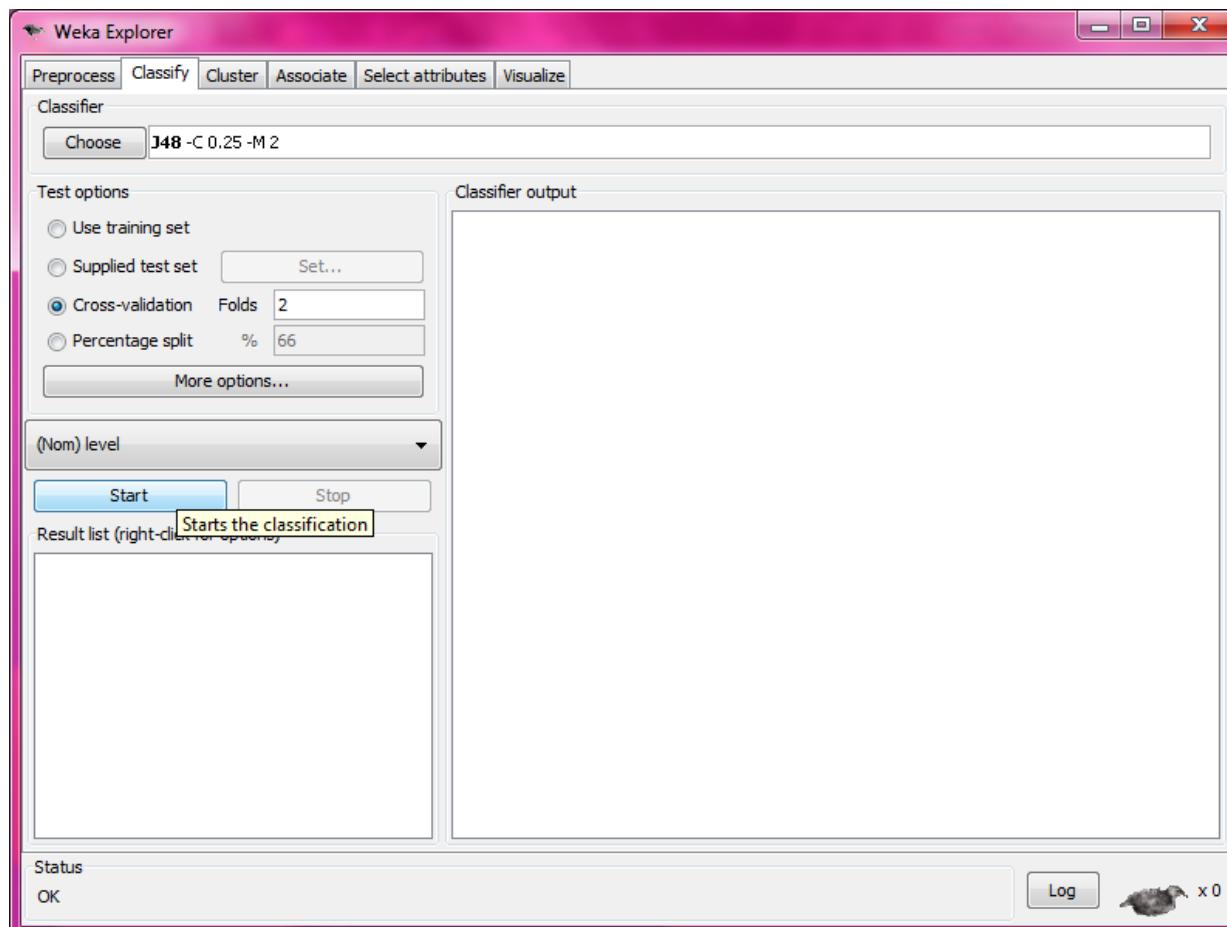
# EXAMPLE 3 : CUSTOMERS (4)



- นำข้อมูลที่ได้มีการแก้ไขด้วยวิธี Replace missing value แล้วสร้าง Decision tree

- Remove ออกทริบิวต์ที่ไม่จำเป็น เช่น member\_id ทั้งหมด
- คลิกที่ช่องหน้าซ้ายของทริบิวต์
- กดปุ่ม Remove

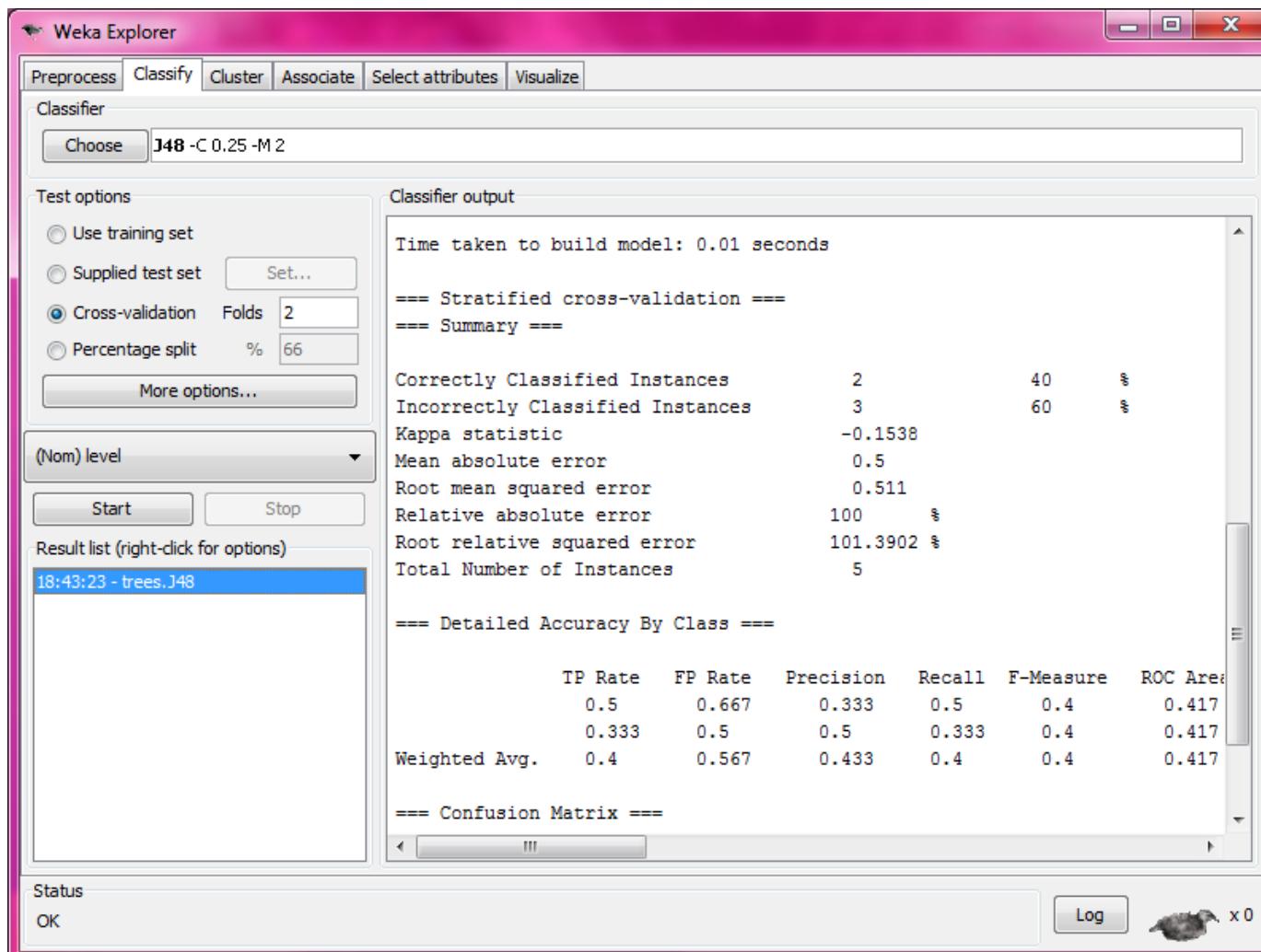
# EXAMPLE 3 : CUSTOMERS (5)



- คลิกที่ tab Classify
- กดปุ่ม Choose
  - เลือก Classifiers
  - เลือก trees
  - เลือก J48
  - เลือก Folds เป็น 2
- กดปุ่ม Start

# EXAMPLE 3 : CUSTOMERS (6)

ผลที่ได้



# EXAMPLE 3 : CUSTOMERS (7)

- คลิกขวาที่โมเดลในช่อง Result list
- เลือก Visualize tree

The screenshot shows the Weka interface with two windows open. The bottom window is titled "Weka Classifier Tree Visualizer: 18:43:23 - trees.J48 (customers\_w...)" and displays a decision tree structure under "Tree View". The root node is labeled "A0001". It has two branches: one labeled "= 0" leading to a leaf node "D (3.0)", and one labeled "= 1" leading to a leaf node "A (2.0)". The top window is titled "Viewer" and shows a table titled "Relation: customers\_wmissing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters....". The table has columns: No., member\_id, sex, age, Income, A0001, A0002, A0003, and level. The data consists of 6 rows:

No.	member_id	sex	age	Income	A0001	A0002	A0003	level
	Numeric	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal
1	1.0	male	20.0	12000.0	0	0	0	D
2	7000.0	1		1	1	1	1	A
3	35000.0	0		0	0	0	0	D
4	4000.0	1		1	0	0	0	A
5	20000.0	0		0	0	0	0	D

Buttons at the bottom of the viewer window include Undo, OK, and Cancel.

# LAB 5-2 : GERMAN CREDIT CARD

## ○ Business Understanding

- การอนุมัติบัตรเครดิตของธนาคารต่าง ๆ จำเป็นจะต้องพิจารณาปัจจัยหลาย ๆ ด้านของลูกค้าผู้ขออนุมัติทั้งนี้ เพราะความเสี่ยงที่อาจจะเกิดขึ้นจากการใช้บัตรเครดิตของลูกค้าอาจจะทำให้ธนาคารสูญเสียเงินเป็นจำนวนมาก
- การสร้างระบบช่วยการตัดสินใจ (decision support system) ในการอนุมัติบัตรเครดิตแบบอัตโนมัติจะช่วยให้ธนาคารสามารถทำงานได้เร็วขึ้น

## ○ Data Understanding

- ธนาคารได้เก็บรวบรวมข้อมูลการขออนุมัติบัตรเครดิตจากลูกค้าเก่าจำนวน 600 คน
- โดยธนาคารจะเก็บคุณลักษณะของลูกค้าแต่ละคนไว้ เช่น จำนวนเงินในบัญชี เป็นต้น
- รวบรวมเก็บไว้ในไฟล์ GermanCreditBalance.arff ซึ่งอยู่ในแผ่น CD โฟลเดอร์ dataset/GermanCredit
- รายละเอียดของแอตทริบิวต์ต่าง ๆ ดูได้จากไฟล์ GermanCredit.pdf



# LAB 5-2 : GERMAN CREDIT CARD

## ○ Data Understanding

- ตัวอย่างของบางแอ็ตทริบิวต์

ลำดับ	แอ็ตทริบิวต์	ชื่อย่อ	ประเภท
1	Observation No.	OBS#	Categorical
2	Checking Account Status	CHK_ACCT	Categorical
3	Duration on credit in months	DURATION	Numerical
4	Credit history	HISTORY	Categorical
5	Purpose of credit	NEW_CAR	Binary
6	Purpose of credit	USED_CAR	Binary
7	Purpose of credit	FURNITURE	Binary
.....			
32	Credit rating is good	RESPONSE	Binary

# LAB 5-2 : GERMAN CREDIT CARD

- Data Preparation
  - .....
- Modeling
  - ใช้เทคนิคการสร้าง Decision Tree เพื่อช่วยในการอนุมัติ/ไม่อนุมัติเครดิตแบบอัตโนมัติ
- Evaluation
  - ทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10 – folds cross – validation
  - ทดสอบด้วยข้อมูลของลูกค้าคนใหม่โดยใช้ไฟล์



# LAB 5-3 : CHURN

## ○ Business understanding

- การยกเลิกการใช้บริการ โทรศัพท์เคลื่อนที่ (Churn) เป็นปัญหาสำคัญสำหรับบริษัทผู้ให้บริการเครือข่ายโทรศัมนาคมในประเทศต่าง ๆ
- การที่สามารถคาดการณ์ได้ว่าลูกค้าคนใดมีแนวโน้มที่จะยกเลิกบริการและทราบถึงปัญหาที่ทำให้ลูกค้าไม่พอใจได้ก่อนที่จะเสียลูกค้าคนนั้นไปจะช่วยให้บริษัทไม่ต้องสูญเสียรายได้

## ○ Data Understanding

- บริษัทโทรศัมนาคมแห่งหนึ่งในประเทศสหรัฐอเมริกาได้เก็บข้อมูลการใช้บริการโทรศัพท์เคลื่อนที่ของลูกค้าจำนวน 3,333 คน
- ข้อมูลของลูกค้านี้แบ่งออกเป็น 2 กลุ่ม
  - กลุ่มที่ยกเลิกการใช้บริการ (Churn) จำนวน 483 คน
  - กลุ่มที่ไม่ยกเลิกการใช้บริการ (No Churn) จำนวน 2,850 คน



# LAB 5-3 : CHURN (2)

## ○ Data Understanding

- มีแอตทริบิวต์ทั้งหมด .... แอตทริบิวต์
- .....
- .....
- .....
- .....
- .....
- ข้อมูลรวมไว้ในไฟล์ churn.csv ในแผ่น CD โฟลเดอร์ dataset/churn

## ○ Data Preparation

- .....



# LAB 5-3 : CHURN (3)

- Modeling

- ใช้เทคนิคการสร้าง Decision Tree เพื่อช่วยในการหาคาดการณ์ว่าลูกค้าคนไหนมีโอกาสย้ายเลิกงานใช้บริการและเพราะเหตุได้

- Evaluation

- ทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10 – folds cross - validation



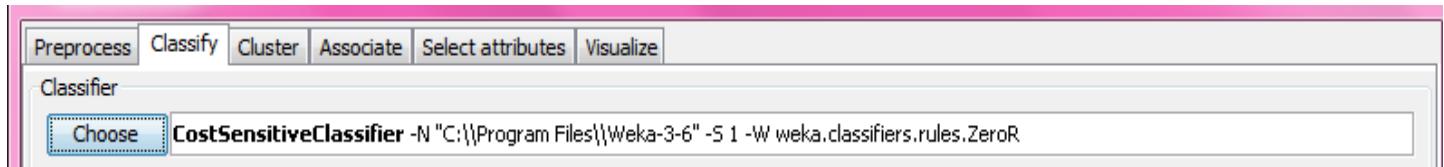
# COST SENSITIVE

- เป็นค่า Cost ที่เกิดขึ้น สำหรับแต่ละคลาสคำตอบ
- ตัวอย่าง ปัญหาการยกเลิกบริการ การใช้โทรศัพท์มือถือของลูกค้า (Churn) พบว่า คลาสที่ต้องการพิจารณา มี 2 class คือ
  - ยกเลิกบริการ (churn)
  - ไม่ยกเลิกบริการ (no churn)
- การพิจารณา Cost Sensitivity ให้พิจารณาดังนี้
  - กรณียกเลิกบริการ (churn)-Cost ของการที่ลูกค้ายกเลิกบริการ คือการพิจารณาว่าบริษัทจะสูญเสียรายได้เท่าไรหากลูกค้า 1 คนทำการยกเลิกบริการ
  - กรณีไม่ยกเลิกบริการ(No churn) – Cost ในกรณีนี้จะเกิดขึ้นก็ต่อเมื่อระบบทำนายว่าลูกค้าจะมีการยกเลิกแต่ความเป็นจริงแล้วลูกค้าไม่ต้องการยกเลิกบริการ ดังนั้นจะเสียค่าใช้จ่ายส่วนหนึ่งในการทำ Campaign กับลูกค้าคนนี้



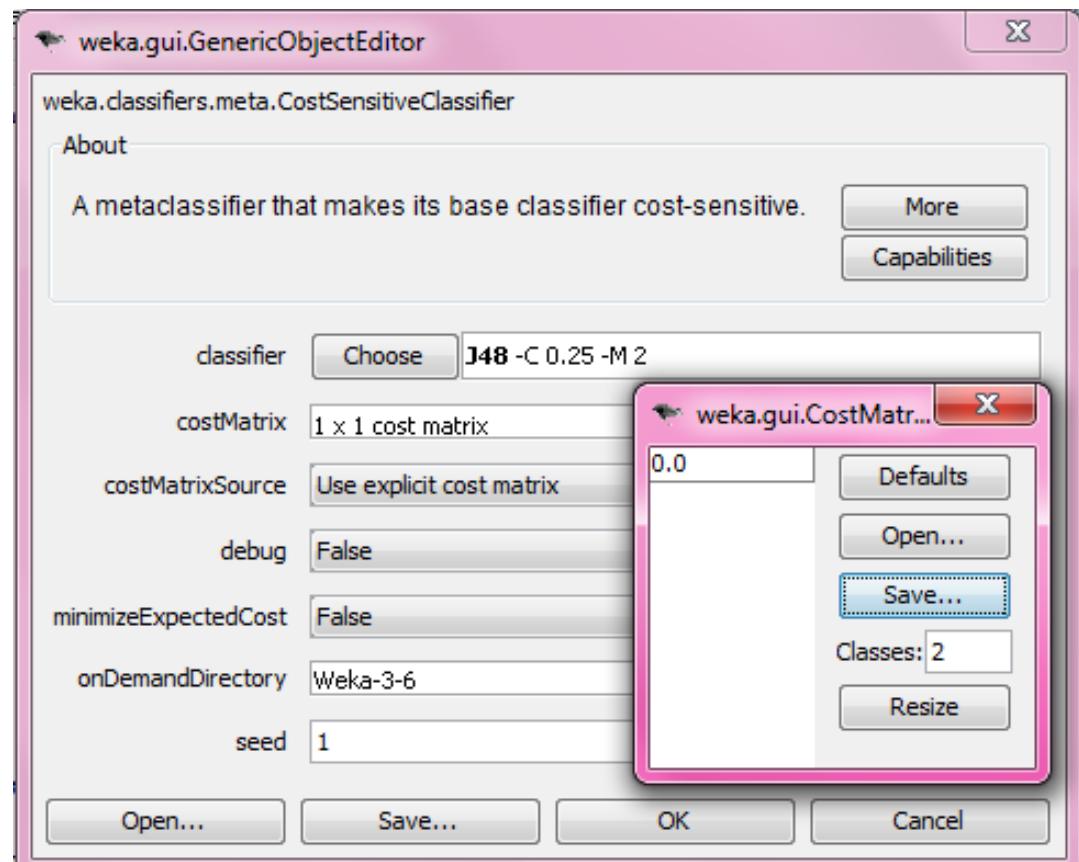
# LAB 5-3 : CHURN WITH COST SENSITIVE

- ใช้ข้อมูล Churn จาก LAB 5-3
- สำหรับค่า cost นี้ สมมติบริษัทได้มีการคำนวณแล้วดังนี้
  - ค่า cost ที่เกิดจากการที่ลูกค้า 1 คนยกเลิกบริการมีค่าเท่ากับ 180 บาท
  - ค่า cost ที่บริษัทต้องใช้ในการจัด Campaign ให้กับกลุ่มลูกค้าที่ลูกทำนายว่าจะยกเลิกบริการ คิดเป็น 18 บาทต่อคน
- ให้ทำการสร้าง Model โดยใช้เทคนิค Decision tree ตามขั้นตอนดังนี้
  - Tab Classify > คลิก Choose
  - เลือก weka > classifiers > meta > CostSensitiveClassifier
  - คลิกที่



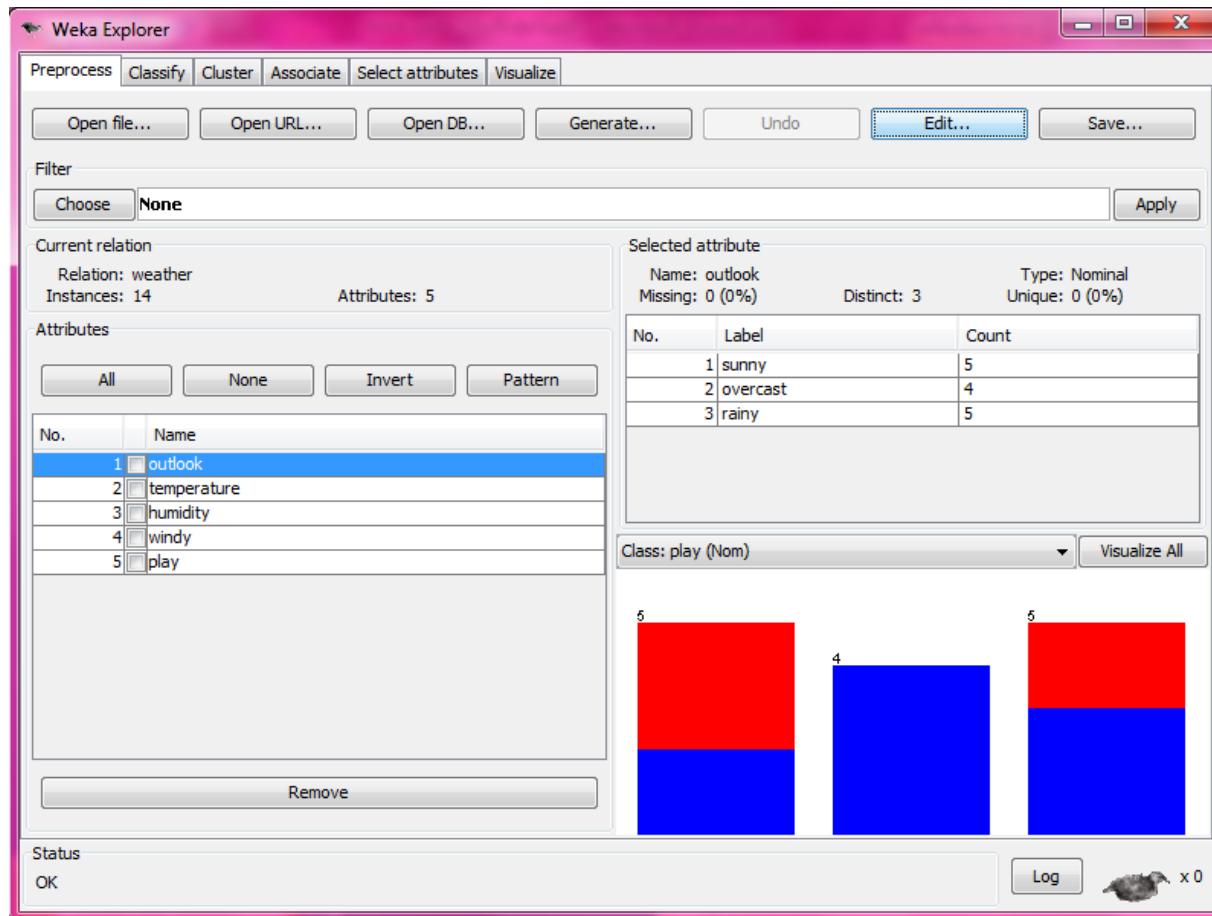
# LAB 5-4 : CHURN WITH COST SENSITIVE(2)

- Classifier เลือก J48
- CostMatrix
  - classes = 2
  - ค่า Cost ของการทำนายผิดว่าลูกค้าจะยกเลิกบริการแต่ความจริงลูกค้าไม่ต้องการยกเลิก = 1
  - ค่า Cost ของการทำนายผิดว่าลูกค้าจะไม่ยกเลิกบริการแต่ความจริงลูกค้าต้องการยกเลิก = 10

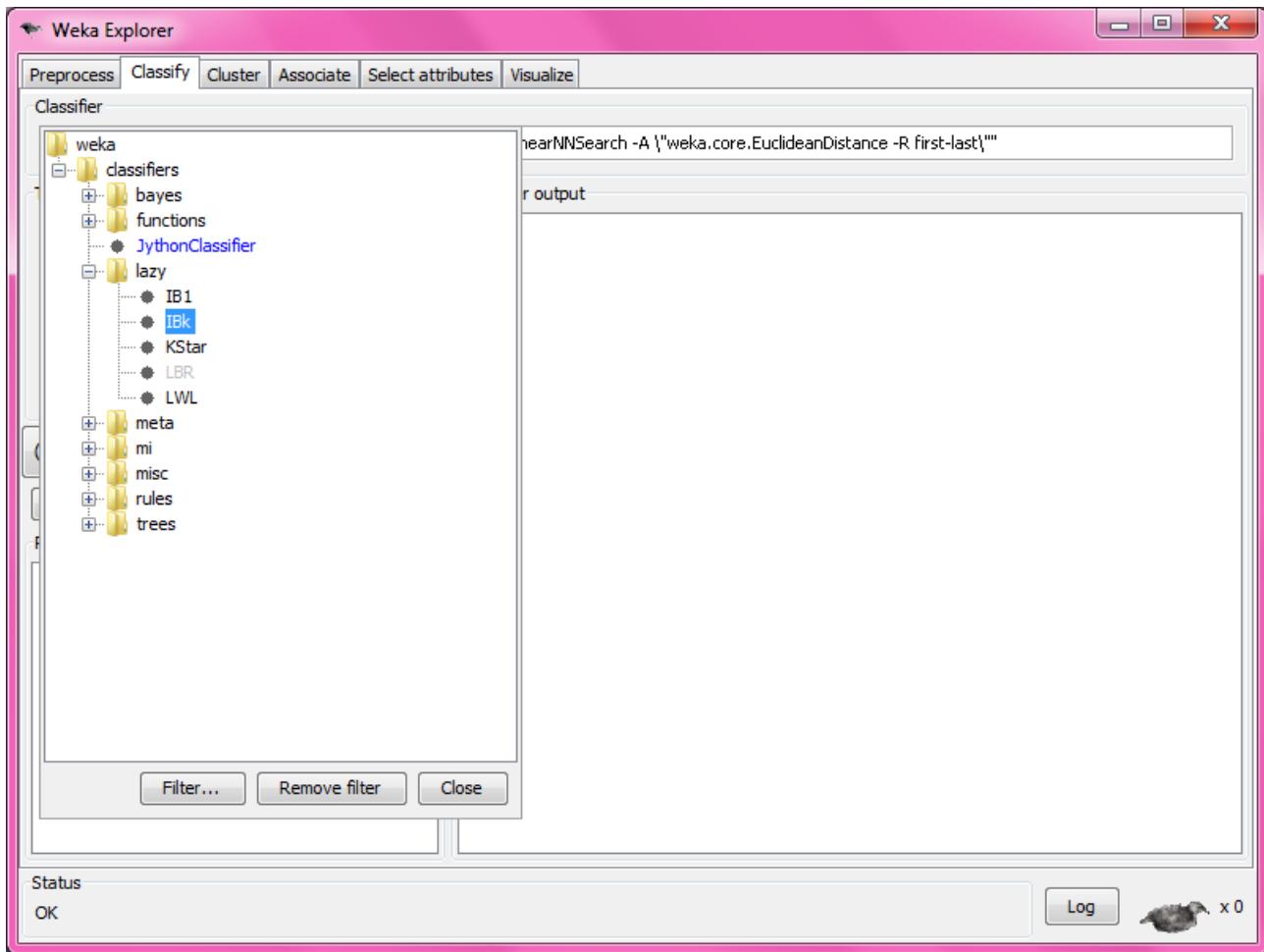


# EXAMPLE 4 : K-NEAREST NEIGHBORS

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ data\weather.arff



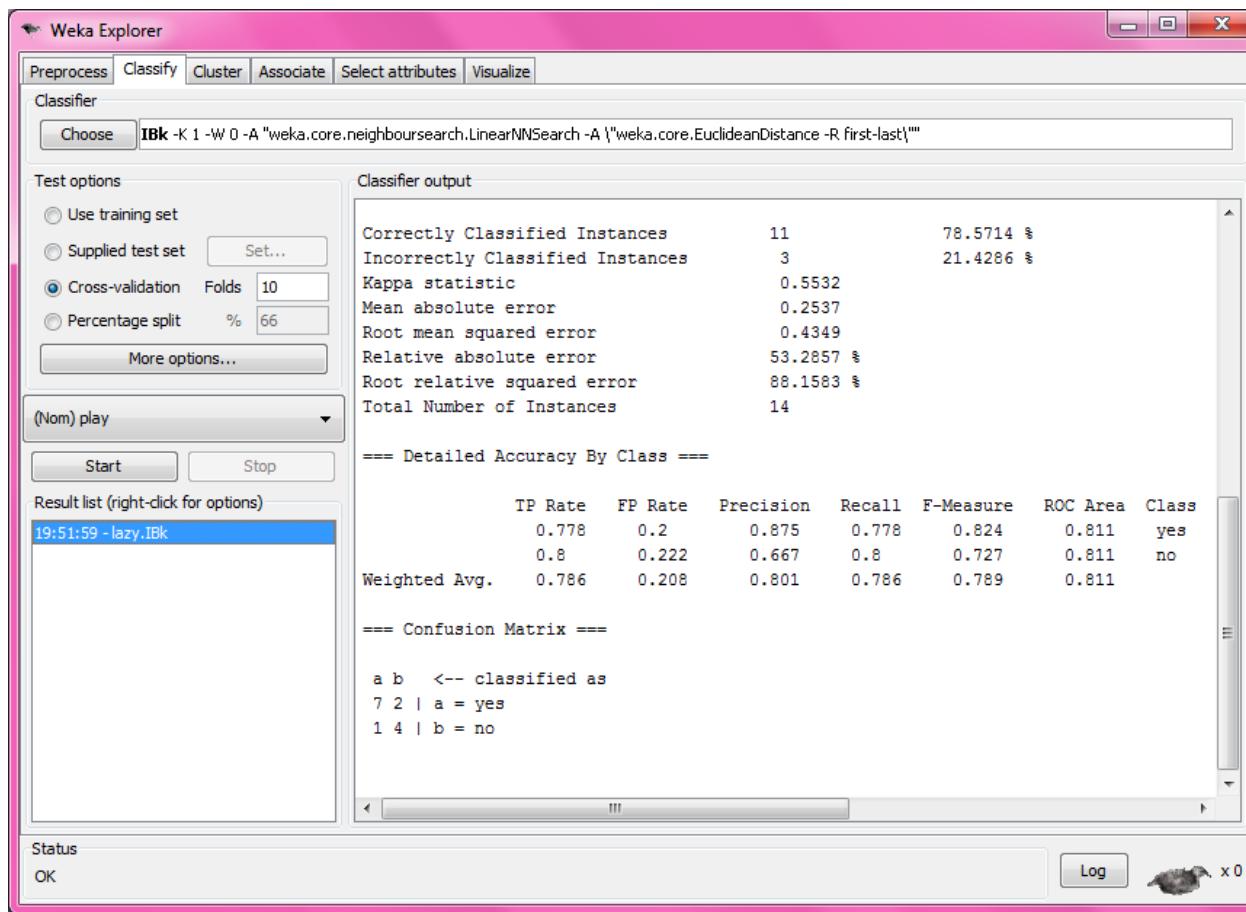
# EXAMPLE 4 : K-NEAREST NEIGHBORS(2)



- คลิกที่ tab classify
- กดปุ่ม choose
  - เลือก classifiers
  - เลือก lazy
  - เลือก ibk
- กดปุ่ม Start

# EXAMPLE 4 : K-NEAREST NEIGHBORS(3)

ผลที่ได้



# LAB 5-5 : BMW PROMOTION CAMPAIGN

## ○ Business Understanding

- ตัวแทนจำหน่ายรถยนต์ BMW ได้จะทำการออกแบบโปรโมชั่นเพื่อขายประกันเพิ่มอีก 2 ปี ให้กับลูกค้าที่เคยซื้อรถ BMW จึงอยากรู้ว่าควรจะนำเสนอโปรโมชั่นนี้ให้กับลูกค้าแบบใด

## ○ Data Understanding

- ตัวแทนจำหน่ายรถยนต์ BMW ได้เคยใช้โปรโมชั่นนี้มาก่อนหน้านี้แล้วทำให้มีลูกค้าเก่าอยู่จำนวน 4,500 คน
- มีจำนวนแอตทริบิวต์ทั้งหมด 4 แอตทริบิวต์
- ข้อมูลรวมเก็บไว้ในไฟล์ bmw-training.arff ในแผ่น CD โฟลเดอร์ dataset/BMW



# LAB 5-5 : BMW PROMOTION CAMPAIGN

## ○ Data Understanding

- ข้อมูลที่เก็บของแต่ละคนมีแอดทริบิวต์ดังนี้

ลำดับ	แอดทริบิวต์	ชื่อย่อ	ประเภท
1	Income Bracket(0=\$0-\$30k, 1=\$31k-\$40k,2=\$41k-\$60k,3=\$61k-\$75k,4=\$76-\$100k,5=\$101k-\$150k,6=\$151-\$500,7=\$501k+)	IncomeBraket	Categorical
2	Year/month first BMW Bought	FirstPurchase	Numeric
3	Year/month most recent BMW bought	LastPurchase	Numeric
4	Responding to the extended warranty	Responded	Categorical



# LAB 5-5 BMW PROMOTION CAMPAIGN (3)

## ○ Data Preparation

- ไม่มีข้อมูลที่ขาดหายไป (missing Values)

## ○ Modeling

- ใช้เทคนิคการทำ K-Nearest neighbors

## ○ Evaluation

- ทดสอบประสิทธิภาพของโมเดลด้วยวิธี Self Consistency Test
- ทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10-folds cross-validation



# REFERENCE

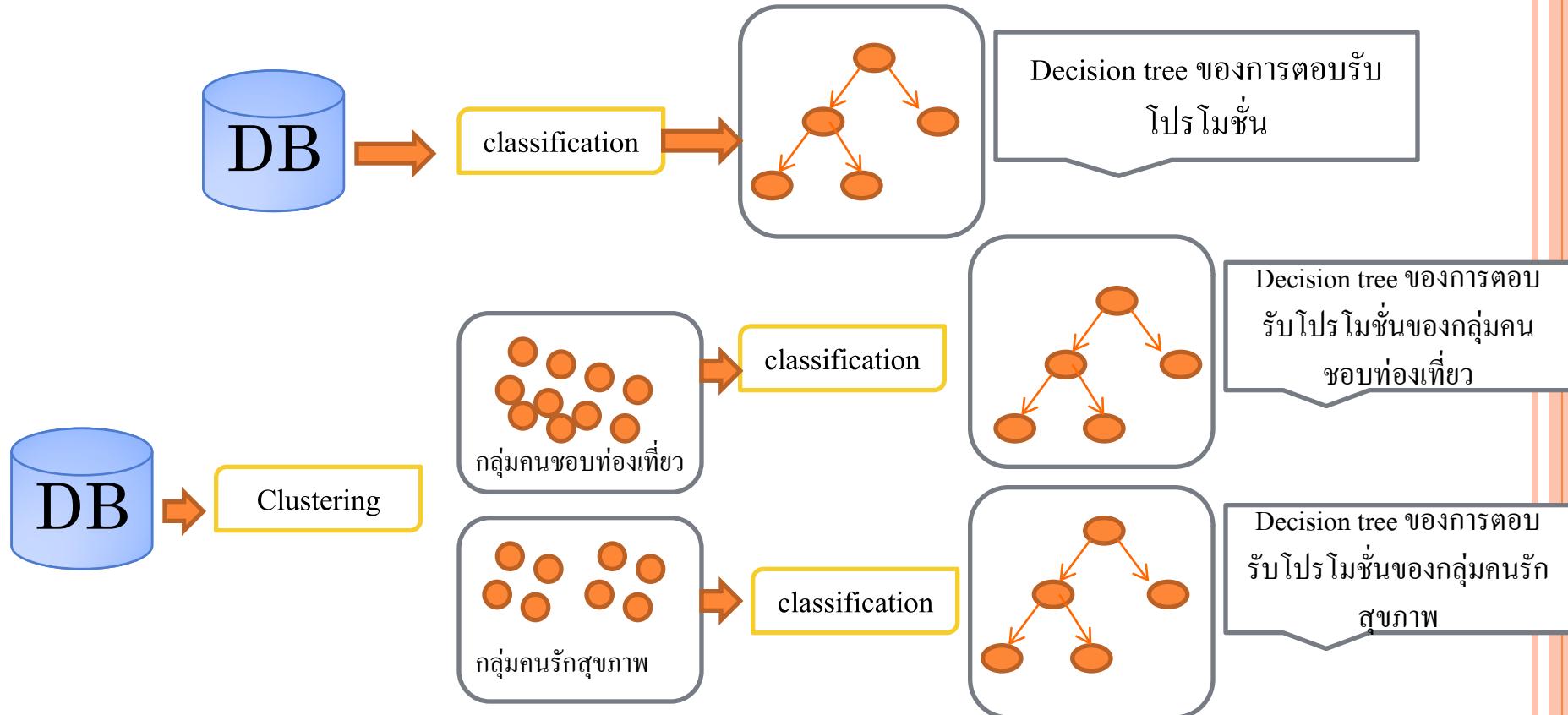
- Discovering Spatial Relationships Between Approximately Equivalent Patterns in contact Maps, BIOKDD04



# CLUSTERING AS PREPROCESSING STEP

- ใช้แบ่งข้อมูลก่อนการวิเคราะห์ข้อมูลด้วยเทคนิคอื่น ๆ

- Classification
- Association rule discovery



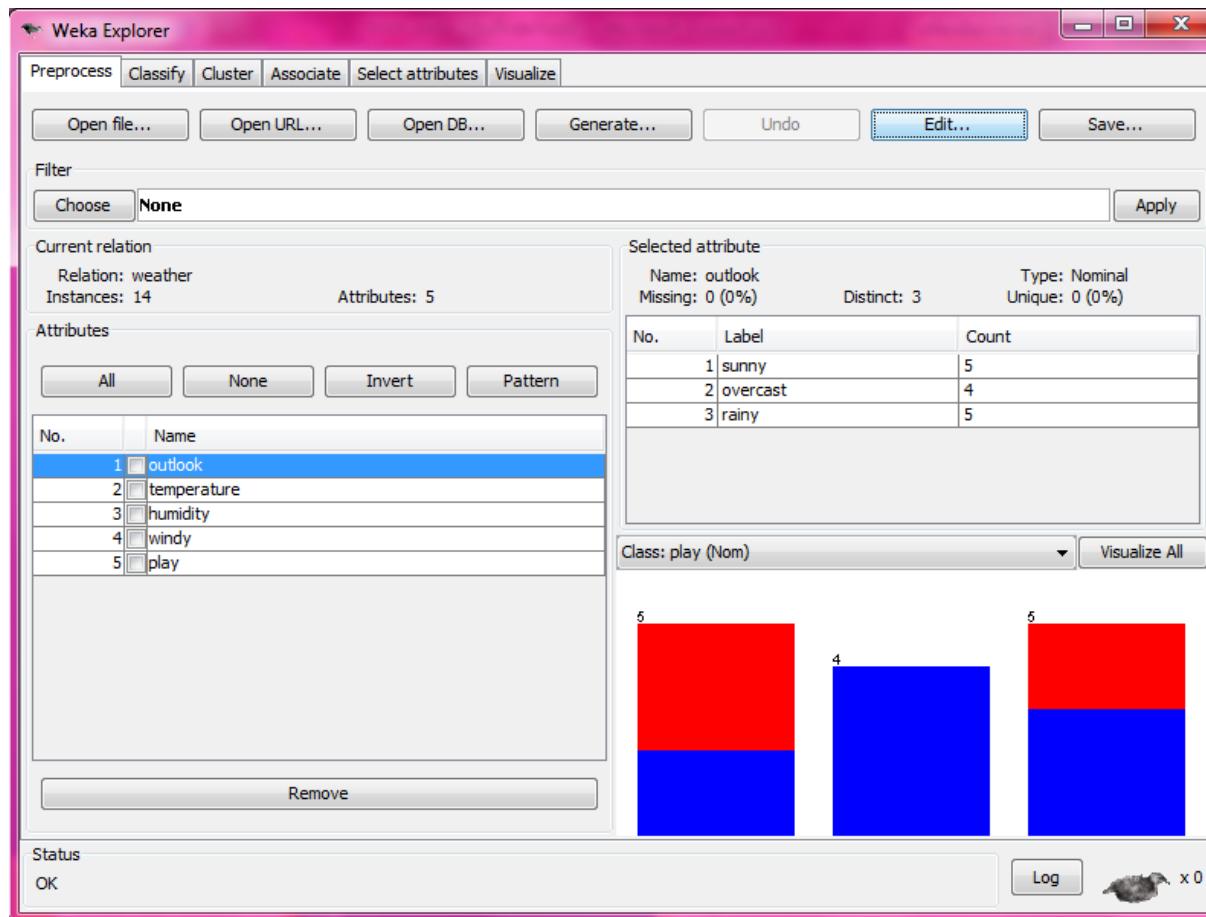
# CLUSTERING APPLICATIONS

- เข้าใจลักษณะของลูกค้าในแต่ละกลุ่ม
- ส่งโปรโมชั่นหรือโฆษณาให้ตรงกับกลุ่มเป้าหมายมากขึ้น
- วางแผน positioning ของสินค้าในบริษัทว่าควรเจาะกลุ่มเป้าหมายใด
- ตัวอย่าง : ร้านเช่า DVD Blockbuster
  - แบ่งกลุ่มข้อมูลลูกค้าตามพฤติกรรมการเช่า DVD เพื่อใช้ในการแนะนำ DVD ให้ตรงกับความต้องการของลูกค้า
- ตัวอย่าง : การแบ่งกลุ่มเอกสาร
  - มีการสร้างอภิธานศัพท์
  - พิจารณาความคล้ายคลึงของคำในเนื้อหาที่ตรงกับอภิธานศัพท์
  - สามารถแบ่งกลุ่มเอกสาร หรือสืบค้นเอกสารที่มีเนื้อหาประเภทเดียวกัน



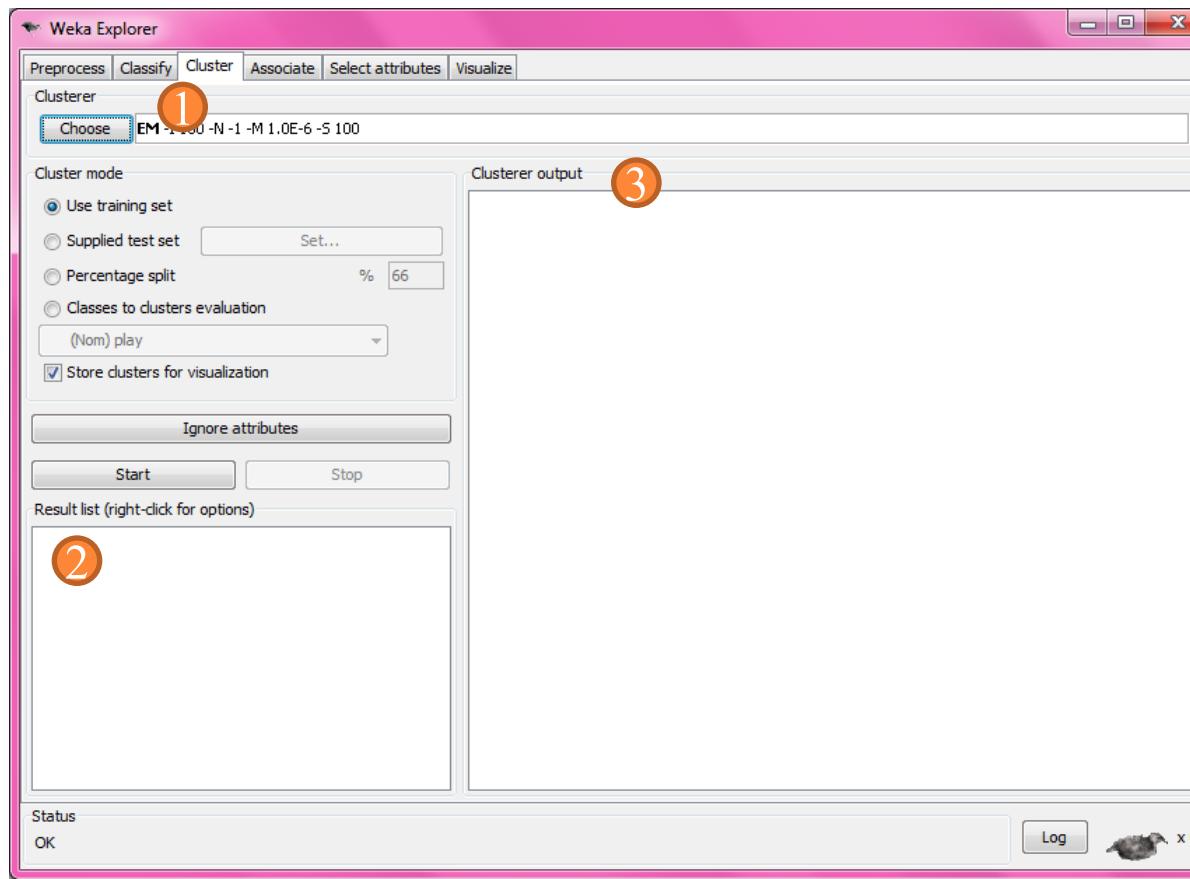
# CLUSTERING IN WEKA

- เปิด Weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ data\weather.arff



# CLUSTERING IN WEKA (CONT’)

- คลิกที่ tab Cluster



# CLUSTERER

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

weka  
clusterers  
CLOPE  
Cobweb  
DBScan  
EM  
FarthestFirst  
FilteredClusterer  
MakeDensityBasedClusterer  
OPTICS  
STAB  
SimpleKMeans  
XMeans

stance -R first-last" -I 500 -S 10

Clusterer output

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs False

distanceFunction Choose EuclideanDistance -R first-last

don'tReplaceMissingValues False

maxIterations 500

numClusters 2

preserveInstancesOrder False

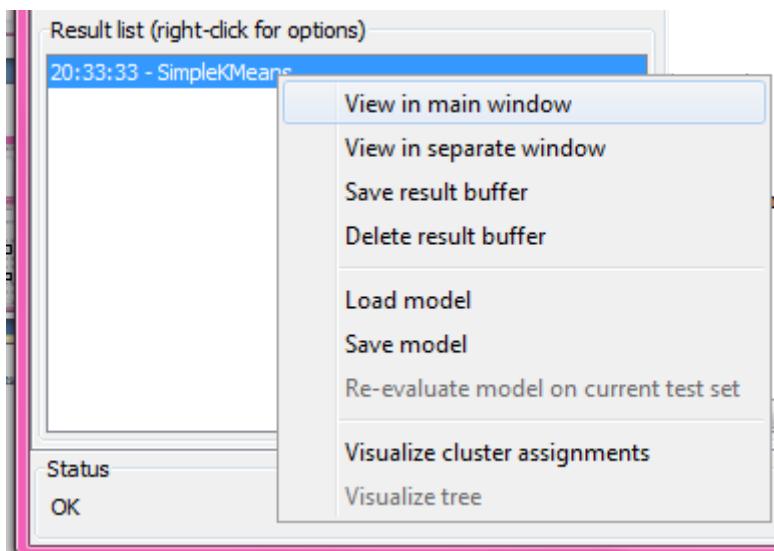
seed 10

Open... Save... OK Cancel

OK

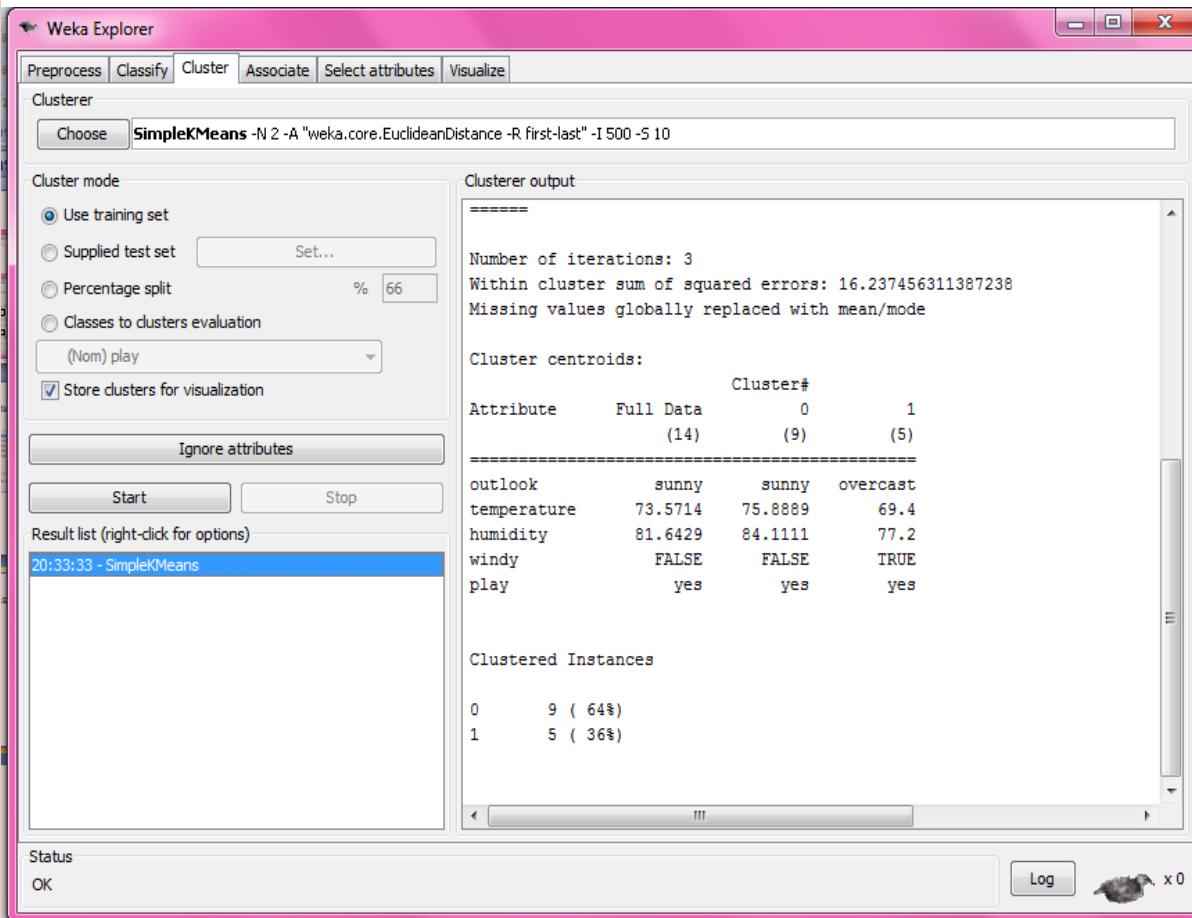
- ส่วนของการเลือกเทคนิคในการคลัสเตอร์
- ส่วนใหญ่จะใช้เทคนิค K-Means
  - DistanceFunction สำหรับเลือกวิธีการวัดระยะห่างระหว่างอินสแตนซ์
  - numClusters สำหรับเลือกจำนวนกลุ่มที่ต้องการ

## 2: RUSULT LIST



- แสดงผลการทำงานครั้งก่อน ๆ
  - แสดงเวลา
  - เทคนิคที่ใช้
  - คลิกขวาที่ผลจะแสดง option !พิมเติม
    - Save result buffer : บันทึกผลการทำงานที่แสดงในส่วน Clusterer output
    - Visualize cluster assignment : แสดงกราฟความสัมพันธ์ของแอตทริบิวต์ต่าง ๆ กับ คลัสเตอร์
      - บันทึกผลการแบ่งคลัสเตอร์

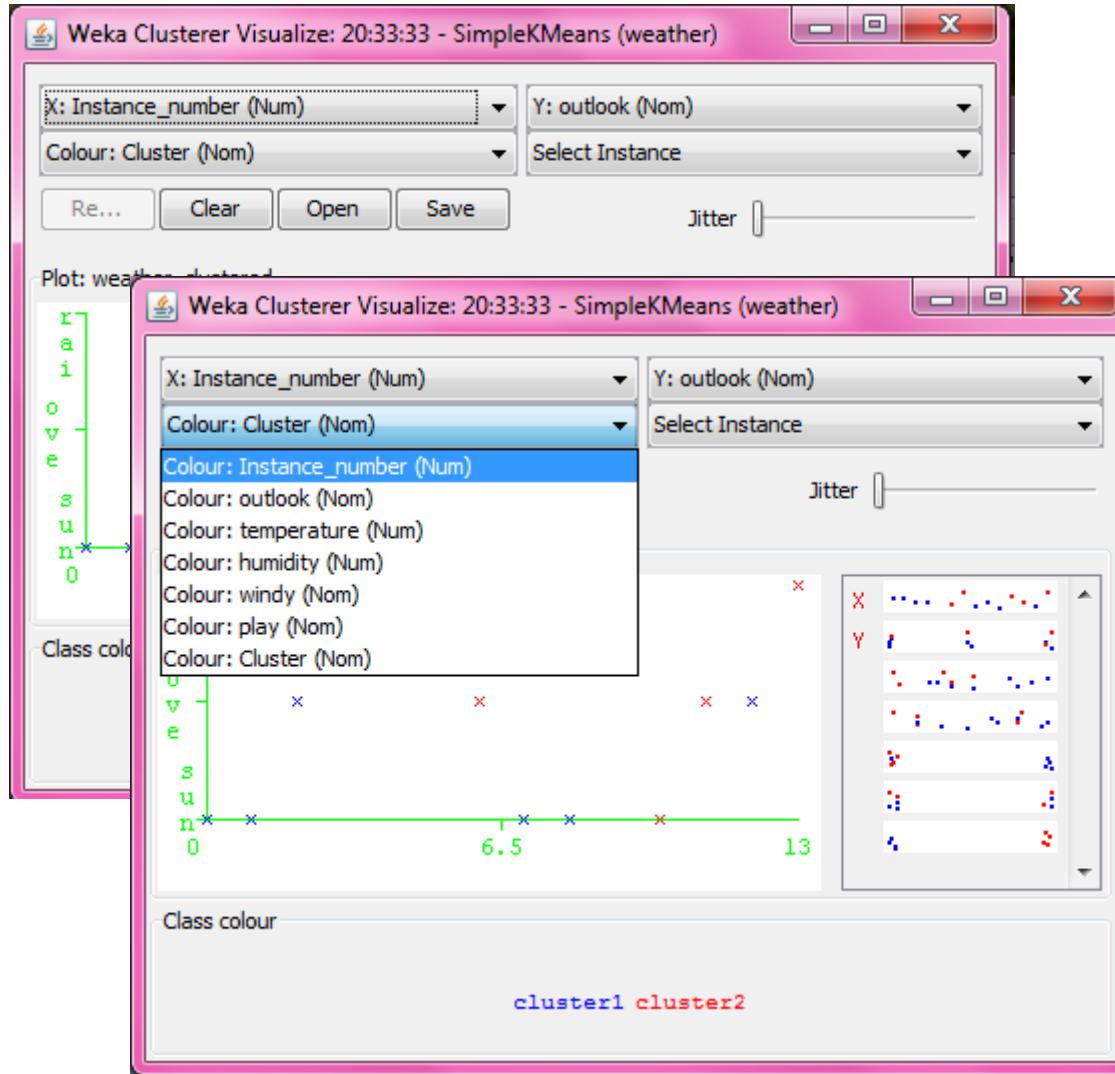
# 3 : CLUSTERER OUTPUT



แสดงผลการแบ่งกลุ่มข้อมูล  
(clustering)

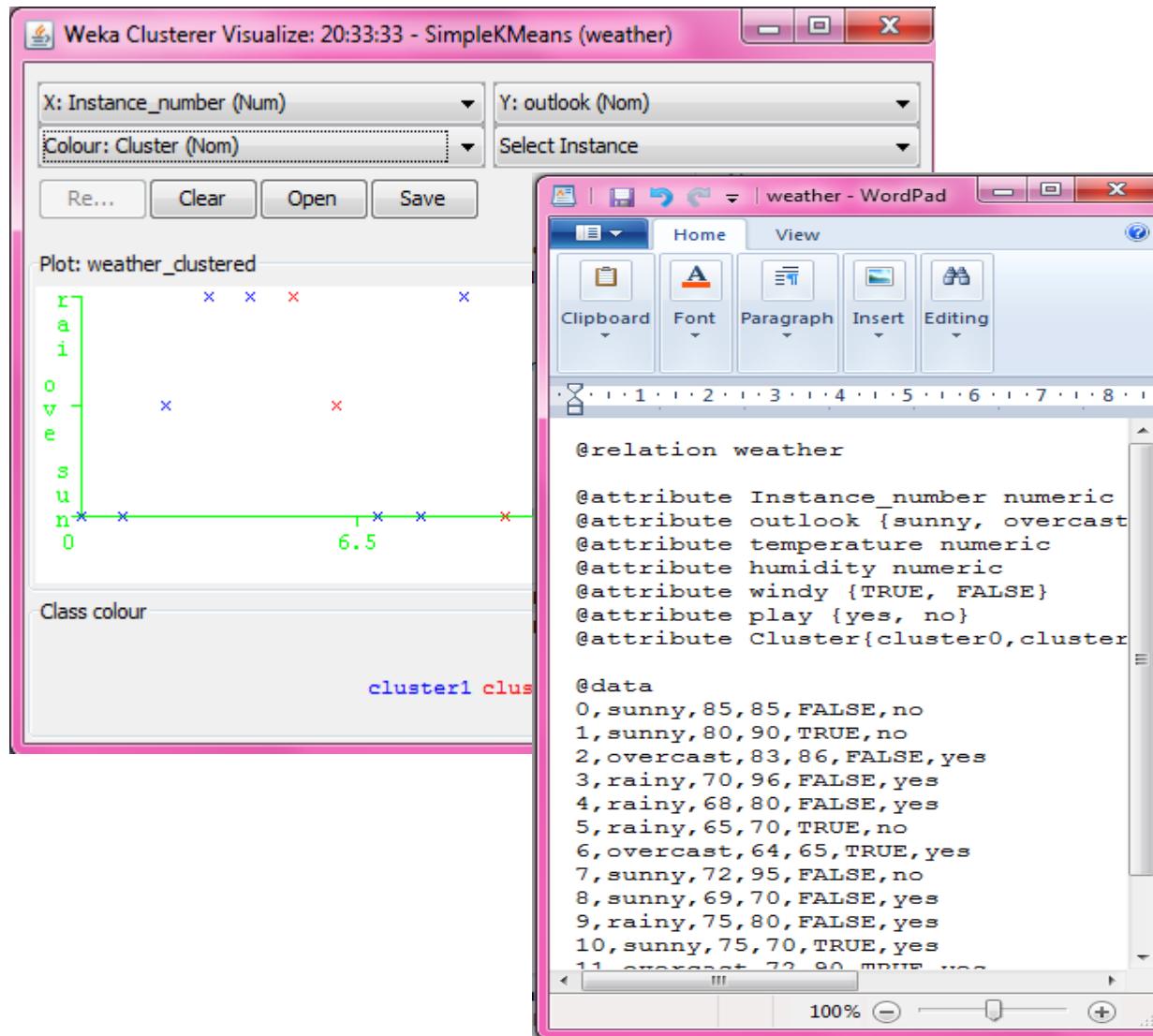
- Number of iterations : จำนวนรอบการทำงานของ K-Means
- Within cluster sum of squared Errors : ค่า distance รวมเมื่อเทียบกับสุดสูนย์กลางของแต่ละคลัสเตอร์
- Cluster centroids : แสดงจุดศูนย์กลางของแต่ละคลัสเตอร์
- Cluster Instances : จำนวนข้อมูลในแต่ละคลัสเตอร์

### 3 : CLUSTERER OUTPUT (CONT')



- แสดงความสัมพันธ์ระหว่างแอ็ตทริบิวต์แบบ 3 มิติ
  - แกน X
  - แกน Y
  - สี
- เลือกแกน X เป็นคลัสเตอร์
- เลือกแกน Y เป็น Instance number
- เลือกสีเป็นแอ็ตทริบิวต์ที่ต้องการดูความสัมพันธ์

### 3 : CLUSTERER OUTPUT (CONT')



- กดปุ่ม save
- ในไฟล์ ARFF จะมีผลการแบ่งกลุ่ม (cluster)

# EXAMPLE 1 : CLUSTERING BANK DATA

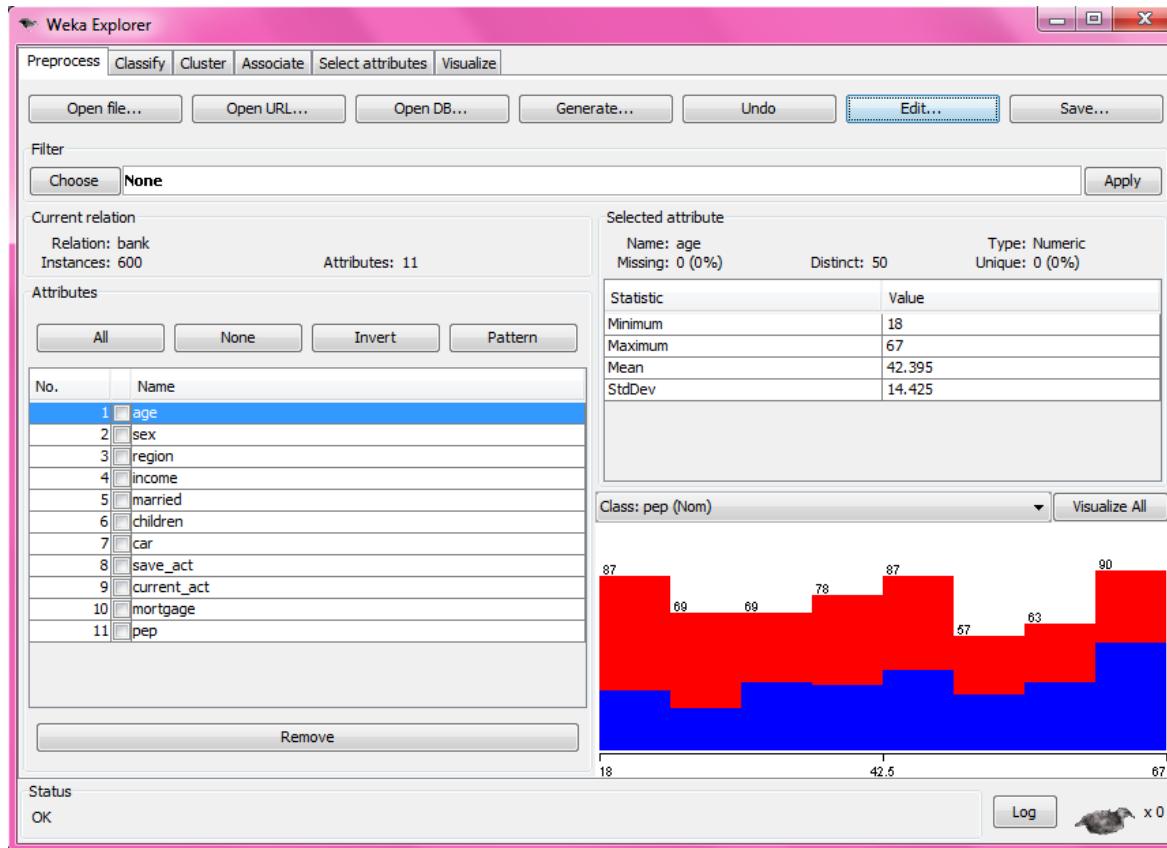
- ข้อมูลรายละเอียดลูกค้าของธนาคาร (bank)

No.	age Numeric	sex Nominal	region Nominal	income Numeric	married Nominal	children Nominal	car Nominal	save_act Nominal	current_act Nominal	mortgage Nominal	pep Nominal
1	48.0	FEMALE	INNER...	17546.0	NO	1	NO	NO	NO	NO	YES
2	40.0	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
3	51.0	FEMALE	INNER...	16575.4	YES	0	YES	YES	YES	NO	NO
4	23.0	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
5	57.0	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
6	57.0	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
7	22.0	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
8	58.0	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
9	37.0	FEMALE	SUBU...	25304.3	YES	2	YES	NO	NO	NO	NO
10	54.0	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO
11	66.0	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO
12	52.0	FEMALE	INNER...	26658.8	NO	0	YES	YES	YES	YES	NO
13	44.0	FEMALE	TOWN	15735.8	YES	1	NO	YES	YES	YES	YES
14	66.0	FEMALE	TOWN	55204.7	YES	1	YES	YES	YES	YES	YES

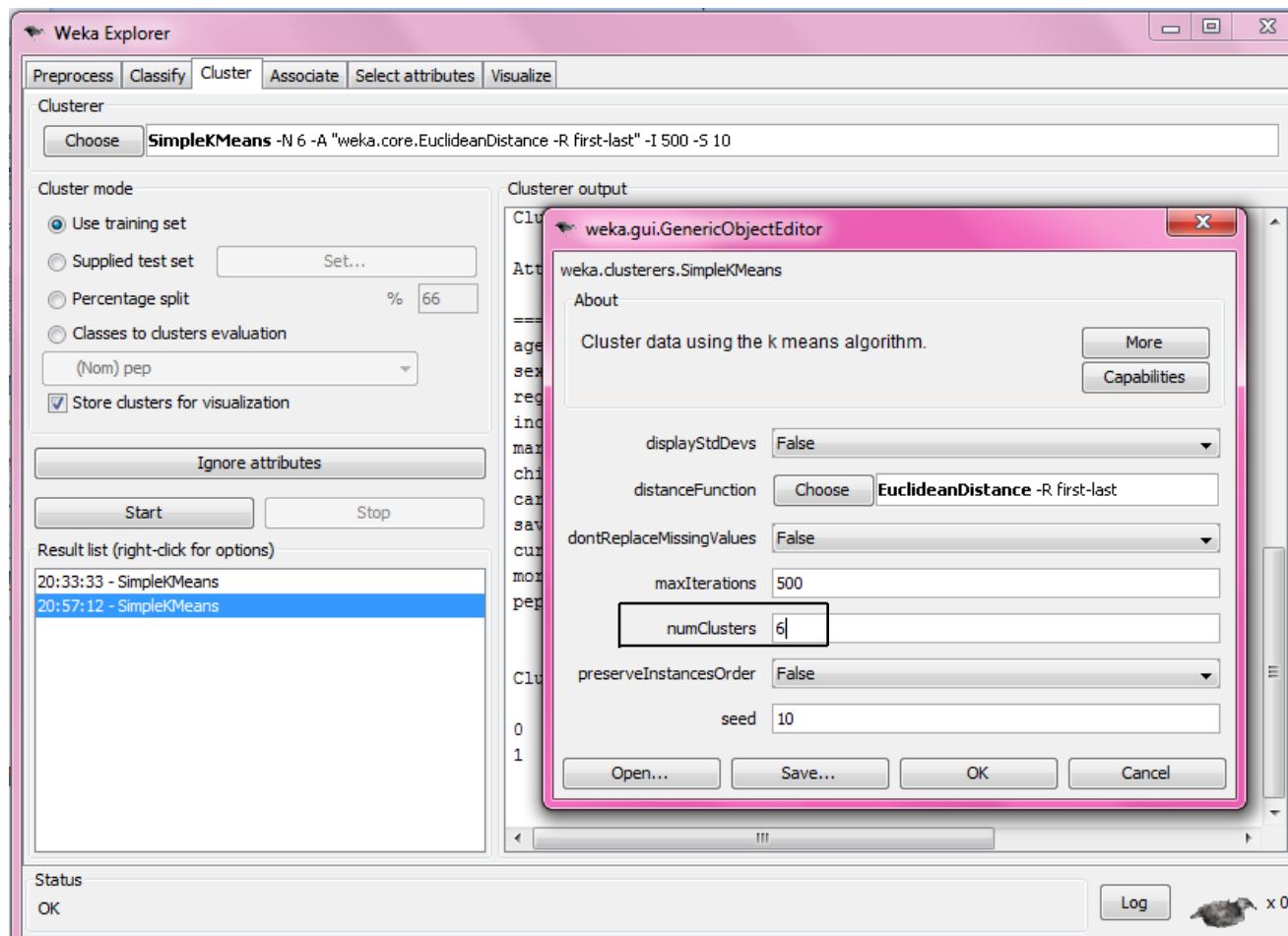


# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

- เปิด weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ dataset\bank.arff



# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)



- คลิกแท็บ Cluster
- เลือกเทคนิค SimpleKMeans
- เป็นค่าพารามิเตอร์ N เป็น 6
- คลิกปุ่ม Start

# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

## ○ ผลการแบ่งกลุ่ม (cluster)

The screenshot shows the Weka Explorer interface with the following details:

- Choose:** SimpleKMeans -N 6 -A "weka.core.EuclideanDistance" -R first-last -I 500 -S 10
- Cluster mode:** Use training set
- Clusterer output:**
  - Number of iterations: 6
  - Within cluster sum of squared errors: 1604.7416693522332
  - Missing values globally replaced with mean/mode
- Cluster centroids:**

Attribute	Full Data (600)	Cluster# 0 (77)	1 (76)	2 (77)	3 (147)
age	42.395	37.1299	44.2763	48.3117	39.1156
sex	FEMALE	FEMALE	FEMALE	FEMALE	FEMALE
region	INNER_CITY	INNER_CITY	RURAL	INNER_CITY	TOWN
income	27524.0312	23377.7604	27772.3746	27668.4396	24047.3865
married	YES	NO	YES	YES	YES
children	0	3	2	1	0
car	NO	NO	NO	NO	NO
save_act	YES	YES	YES	NO	YES
current_act	YES	YES	YES	YES	YES
mortgage	NO	NO	NO	NO	NO
pep	NO	NO	NO	YES	NO
- Result list:** 20:33:33 - SimpleKMeans, 20:57:12 - SimpleKMeans, 21:06:45 - SimpleKMeans (The last entry is highlighted in blue).
- Status:** OK

## EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

### ○ ผลของคลัสเตอร์ที่ 2

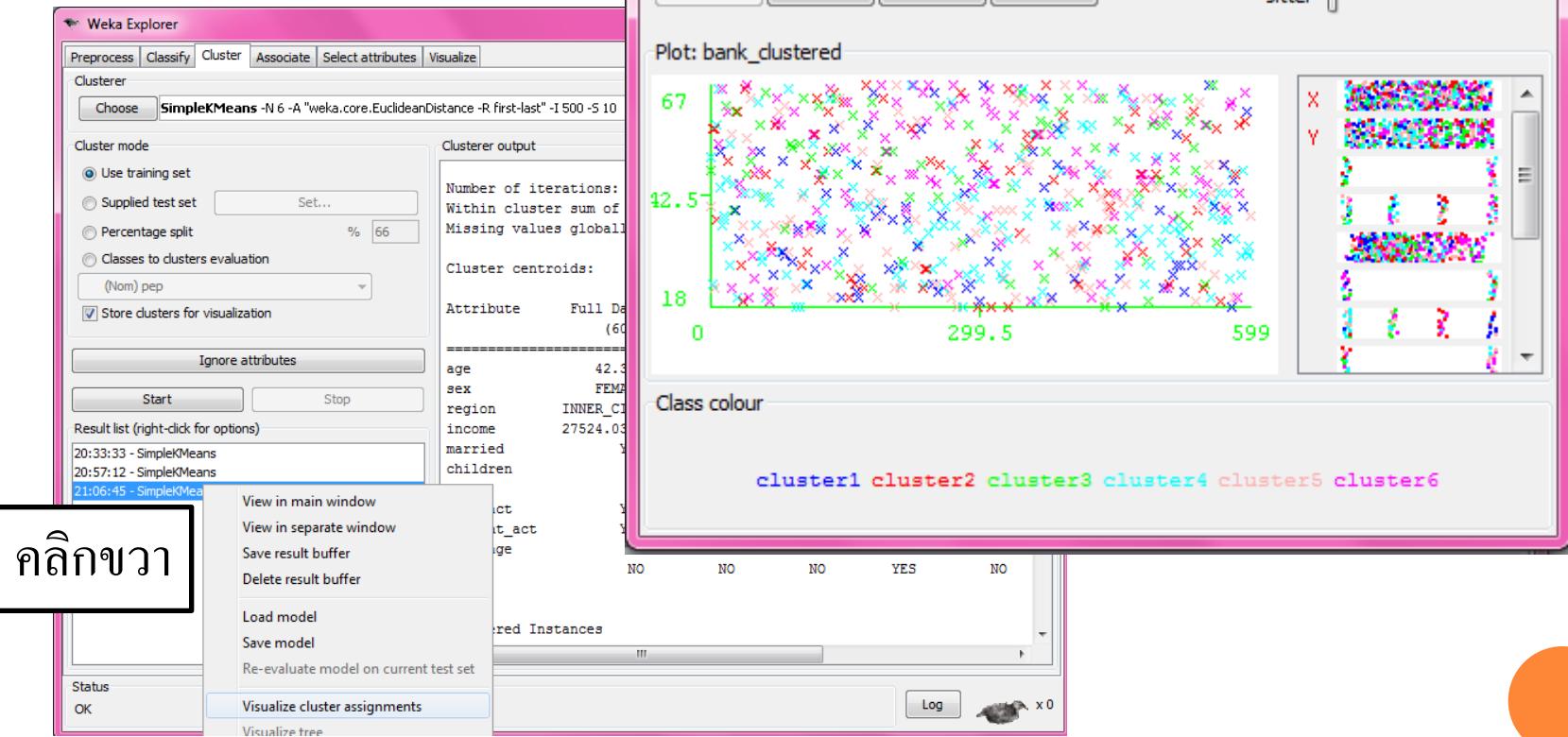
Age	Sex	Region	Income	Married	Children	Car	Save_act	Current_act	Mortgage	Pep
39.2569	FEMALE	INNER _CITY	25,148.6273	YES	0	YES	YES	YES	NO	NO

### ○ ลูกค้าส่วนใหญ่ในคลัสเตอร์นี้

- อายุเฉลี่ยประมาณ 39 ปี
- เพศหญิง
- อาศัยอยู่ในเมือง
- รายได้เฉลี่ยประมาณ 25,148 บาท
- แต่งงานแล้วแต่ยังไม่มีบุตร
- มีบัญชีเงินฝากกับธนาคารเรียบร้อยแล้ว

# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

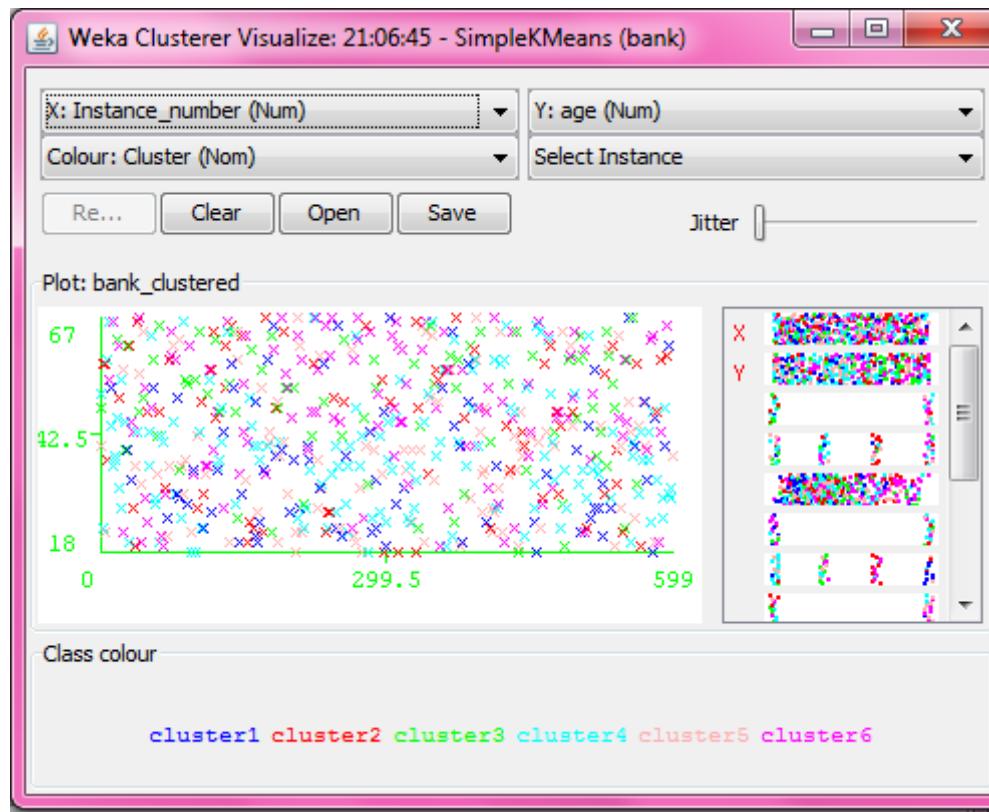
- คลิกขวาที่ผลใน Result list
  - เลือก Visualize cluster assignments



# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

## ○ แสดงกราฟแบบ 3 มิติ

- แกน X
- แกน Y
- สี



# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

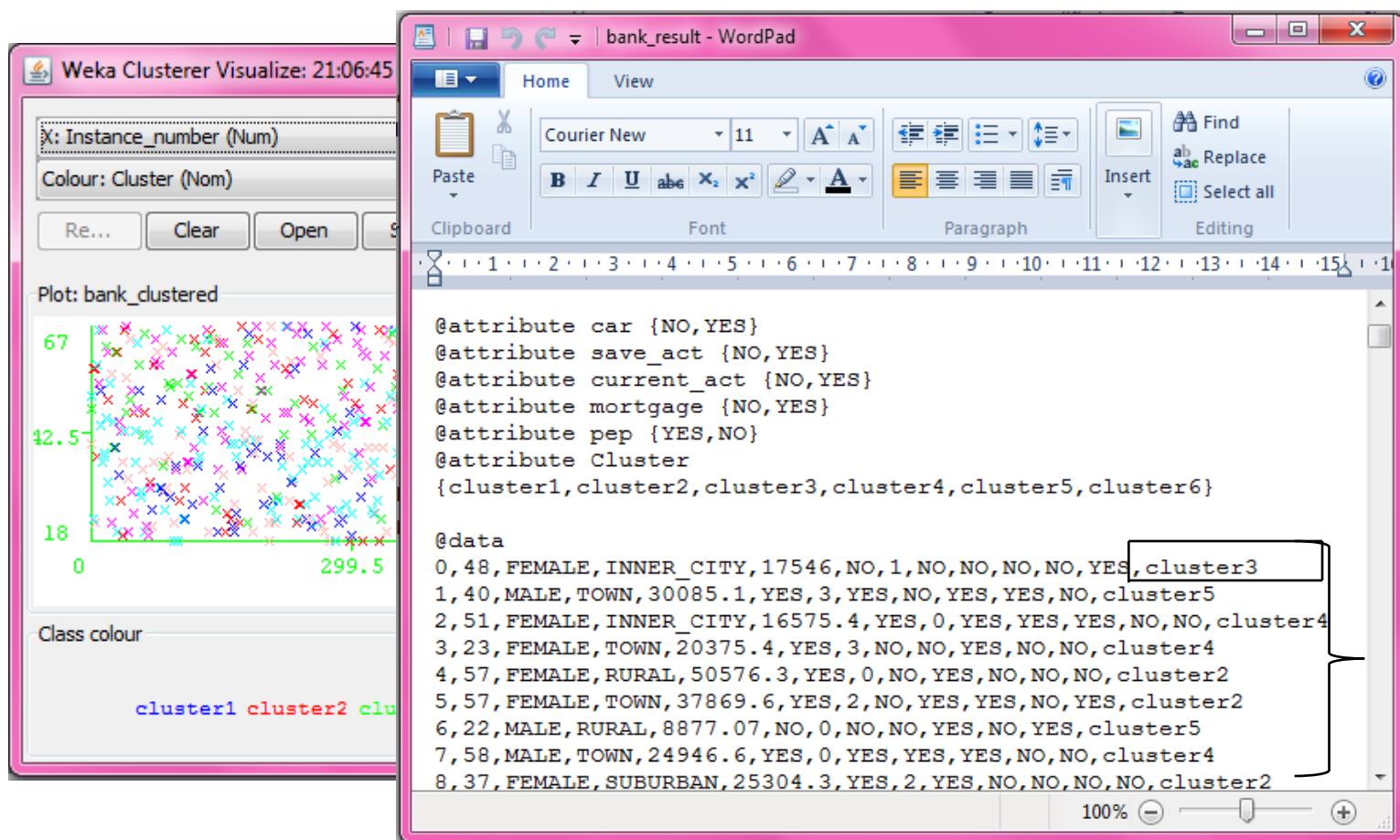
- เลือกให้แกน X เป็น Cluster และแกน Y เป็น Instance\_number
- เลือก Colour เป็น sex



- Hint!!!
  - คลัสเตอร์ 3 ลูกค้าส่วนใหญ่จะเป็นเพศหญิง (พนักงานมากกว่า)
  - Cluster 2 ลูกค้าส่วนใหญ่จะเป็นเพศชาย (พนักงานมากกว่า)

# EXAMPLE 1 : CLUSTERING BANK DATA(CONT’)

- กดปุ่ม Save ในหน้าต่าง Weka Clusterer Visualize
- Save ในชื่อ bank\_result.arff



มีค่าคลัสเตอร์  
(cluster)  
เพิ่มขึ้นมา

# LAB 6-1 : CUSTOMER BEHAVIORS

## ○ Business Understanding

- ตัวแทนจำหน่ายรถยนต์ BMW ได้ทำการติดตามคุณภาพติกรรมของลูกค้าที่เข้ามาเยี่ยมชมรถภายในโชว์รูมว่าลูกค้าคือรถยนต์รุ่นไหนและทำการสั่งซื้อรถหรือไม่

## ○ Data Understanding

- ตัวแทนจำหน่ายรถยนต์รายนี้ได้ทำการติดตามคุณภาพติกรรมของลูกค้าทั้งหมดจำนวน 100 คน
- ลูกค้าแต่ละรายจะมีจำนวนแอตทริบิวต์ทั้งหมด 8 แอตทริบิวต์
- ข้อมูลเหล่านี้ได้ถูกรวบรวมเก็บไว้ในไฟล์ bmw\_browsers.arff ในแผ่น CD โฟลเดอร์ dataset/bmw



# LAB 6-1 : CUSTOMER BEHAVIORS (CONT’)

## ○ Data Understanding

- ข้อมูลที่เก็บของแต่ละคนมีแอ็ตทริบิวต์ดังนี้

ลำดับ	แอ็ตทริบิวต์	ประเภท
1		Numeric
2	Dealership	.....
3	Showroom	Numeric
4	Computer Search	.....
5	M5	Numeric
6	3Series	Numeric
7	Z4	.....
8	Financing	
9	Puchase	Numeric

# LAB 6-1 : CUSTOMER BEHAVIORS (CONT’)

- Data Preparation

- .....

- Modeling

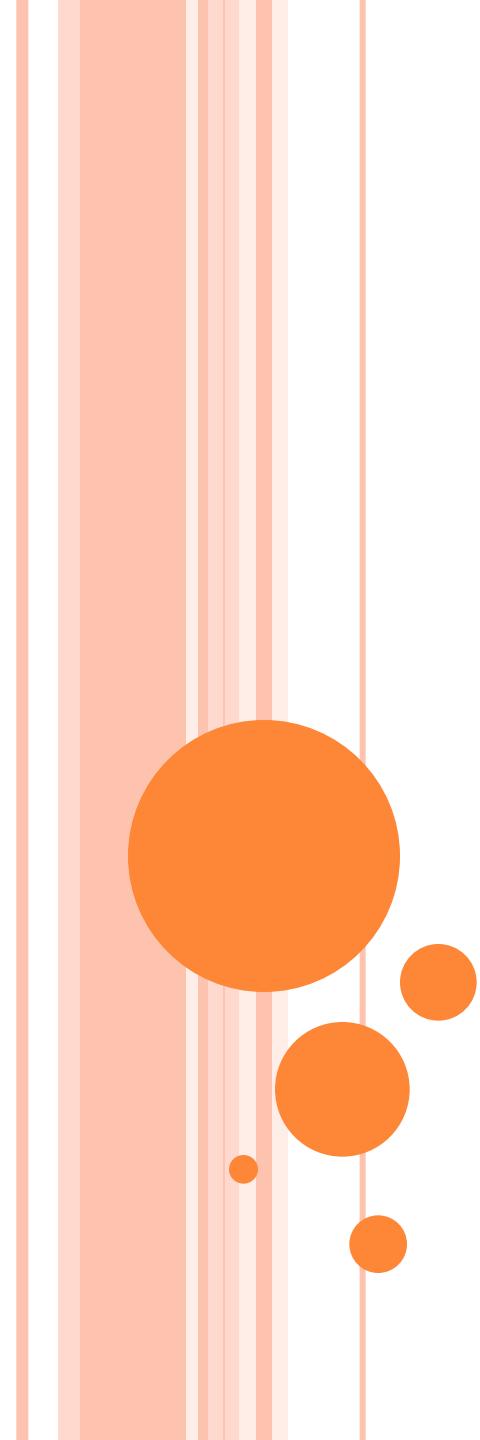
- ใช้เทคนิคการแบ่งกลุ่มด้วย K-means Clustering
  - แบ่งกลุ่มออกเป็น 5 กลุ่ม



End Part VI

Any Question ?

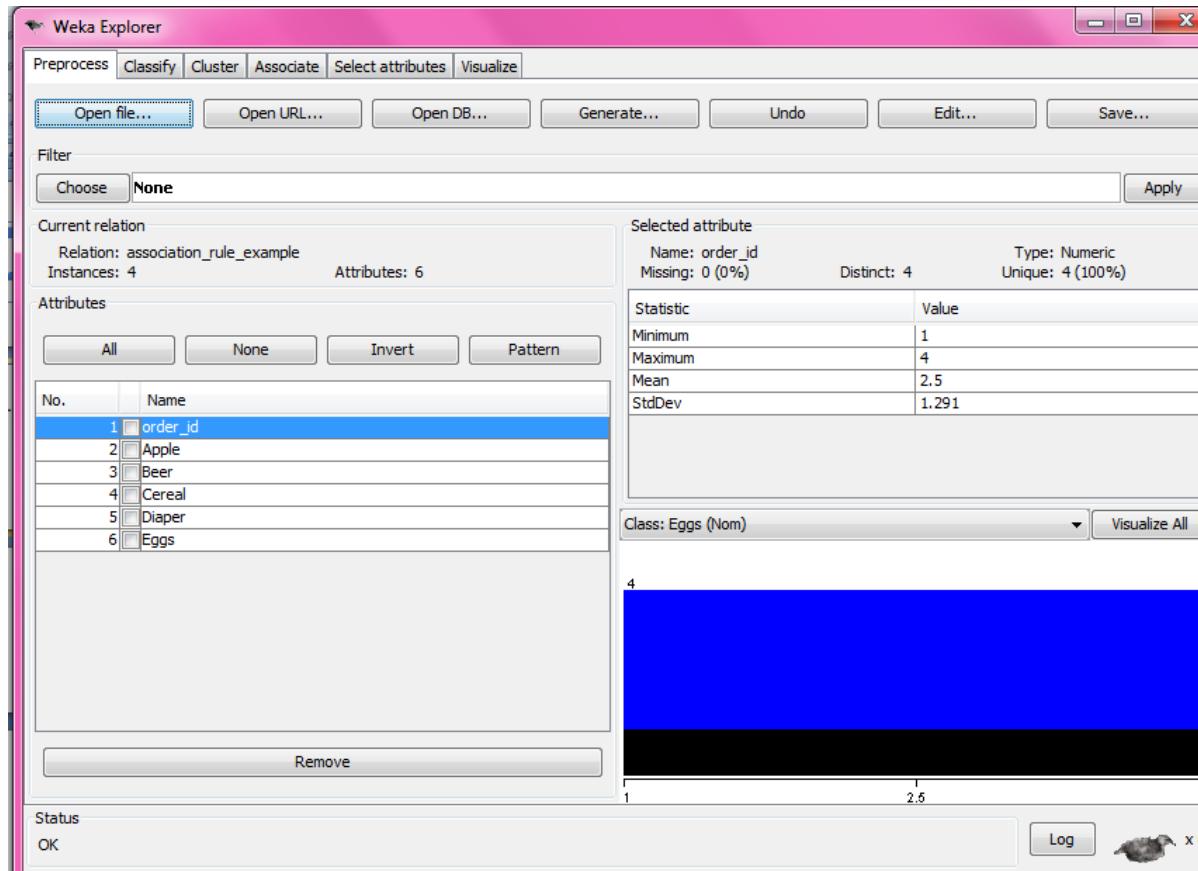




# **AN INTRODUCTION TO DATA MINING WITH WEKA Association Rules**

# ASSOCIATION RULES IN WEKA

- เปิด weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ dataset\Example1.arff



# ASSOCIATION RULES IN WEKA

○ កត់ប្រើ Edit

QID	Items
1	Apple, Ceral, Diapers
2	Beer, Cereal, Eggs
3	Apple, Beer, Cereal, Eggs
4	Beer, Eggs



Weka Viewer window showing the same data in a tabular format:

No.	order_id Numeric	Apple Nominal	Beer Nominal	Cereal Nominal	Diaper Nominal	Eggs Nominal
1	1.0	y		y	y	
2	2.0		y	y		y
3	3.0	y	y	y		y
4	4.0	y				y

Buttons at the bottom: Undo, OK, Cancel.

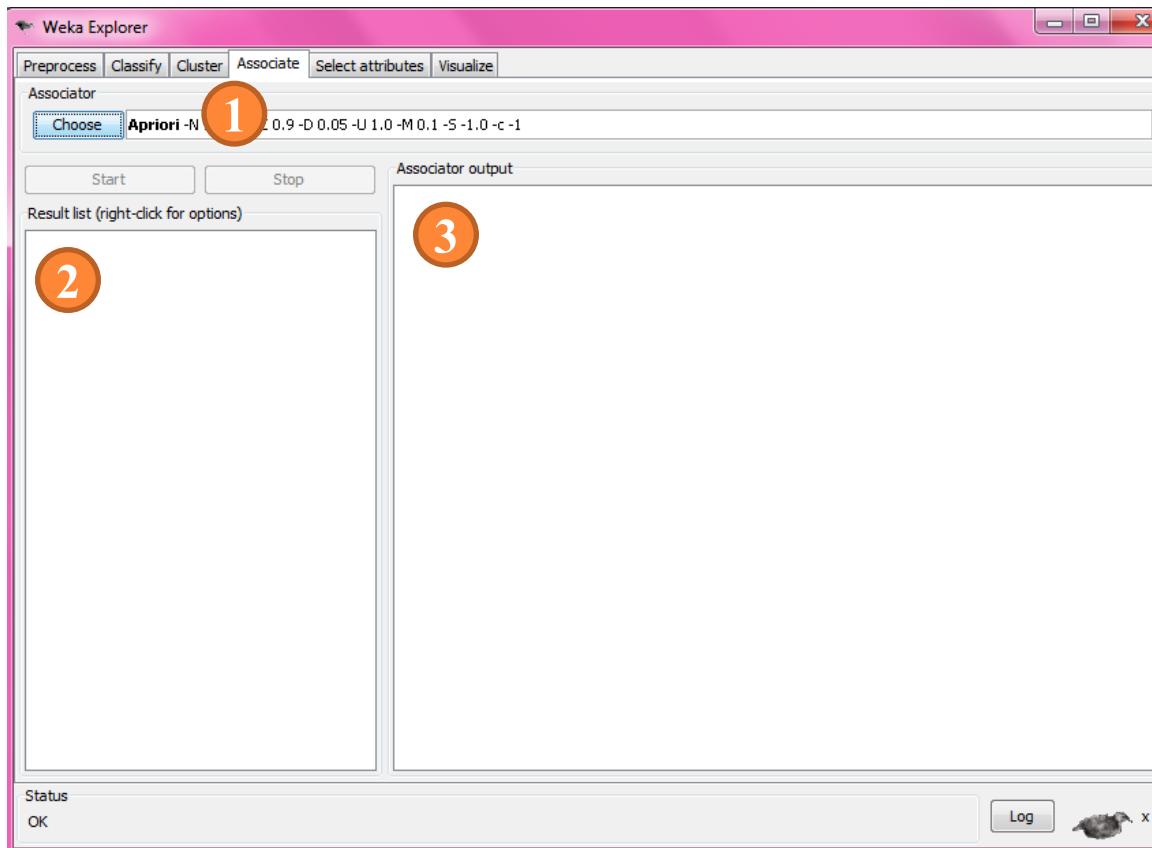
```
@relation association_rule_example

@attribute order_id numeric
@attribute Apple {y,?}
@attribute Beer {y,?}
@attribute Cereal {y,?}
@attribute Diaper {y,?}
@attribute Eggs {y,?}

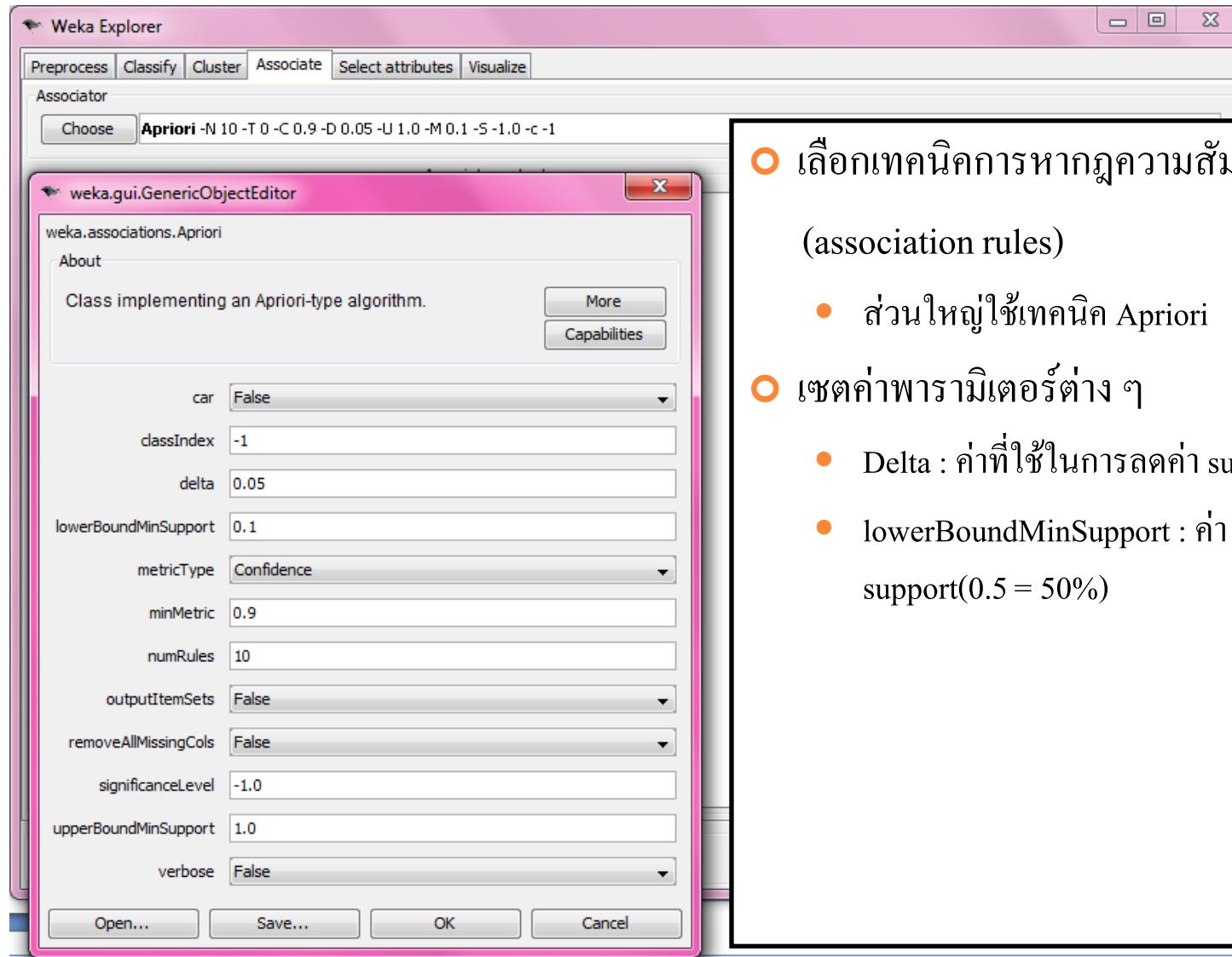
@data
1,y,?,y,y,?
2,?,y,y,?,y
3,y,y,y,?,y
4,?,y,?,?,y
```

# ASSOCIATION RULES IN WEKA(CONT')

- คลิกที่ tab Associate



# 1 : ASSOCIATOR



## เลือกเทคนิคการหากฎความสัมพันธ์

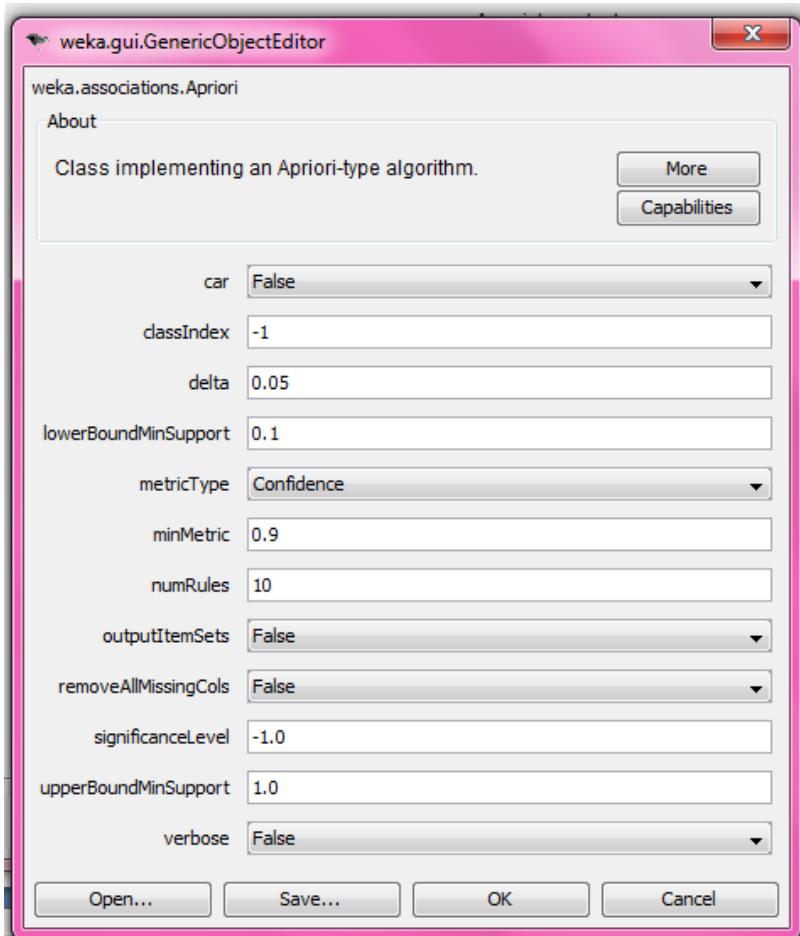
(association rules)

- ส่วนใหญ่ใช้เทคนิค Apriori

## เซตค่าพารามิเตอร์ต่าง ๆ

- Delta : ค่าที่ใช้ในการลดค่า support
- lowerBoundMinSupport : ค่า minimum support( $0.5 = 50\%$ )

# 1 : ASSOCIATOR (CONT')



## ○ เช็ตค่าพารามิเตอร์ต่าง ๆ

- MetricType:score ที่ใช้ในการเรียน (rank) กฎความสัมพันธ์
- minMetric : ค่า Score ที่เลือกกันใน Metric Type
  - กฎที่สนใจจะต้องมีค่ามากกว่าที่กำหนด
- numRules:จำนวนกฎความสัมพันธ์ที่ต้องการ
- outputItemSets:แสดงความถี่ของ สินค้าที่มีการซื้อพร้อมกัน

## 2 : RESULT LIST

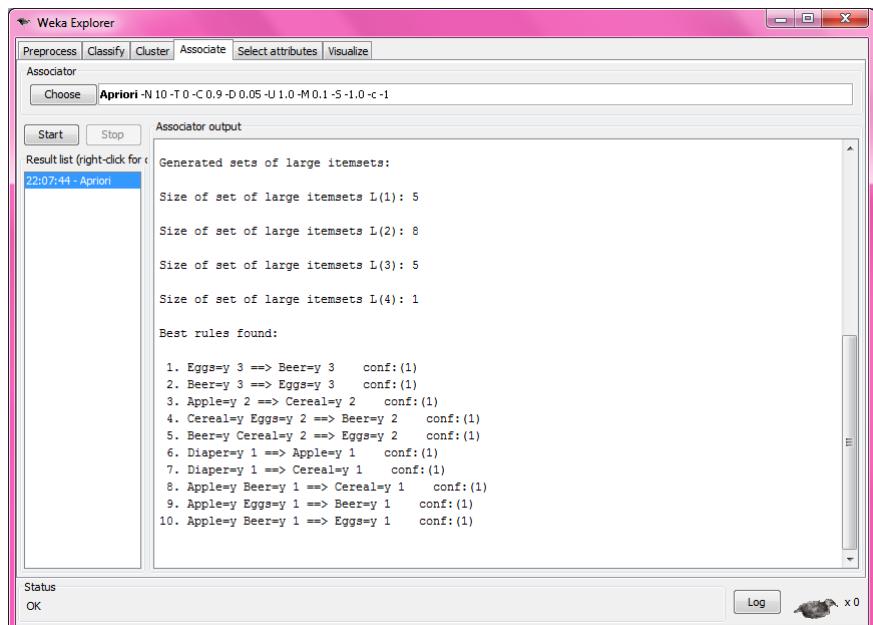


### ○ แสดงผลการทำงานครั้งก่อน ๆ

- แสดงเวลา
- เทคนิคที่ใช้
- คลิกขวาที่ผลจะแสดง option เพิ่มเติม
  - Save result buffer : ใช้บันทึกผลการทำงาน



# 3 : ASSOCIATOR OUTPUT



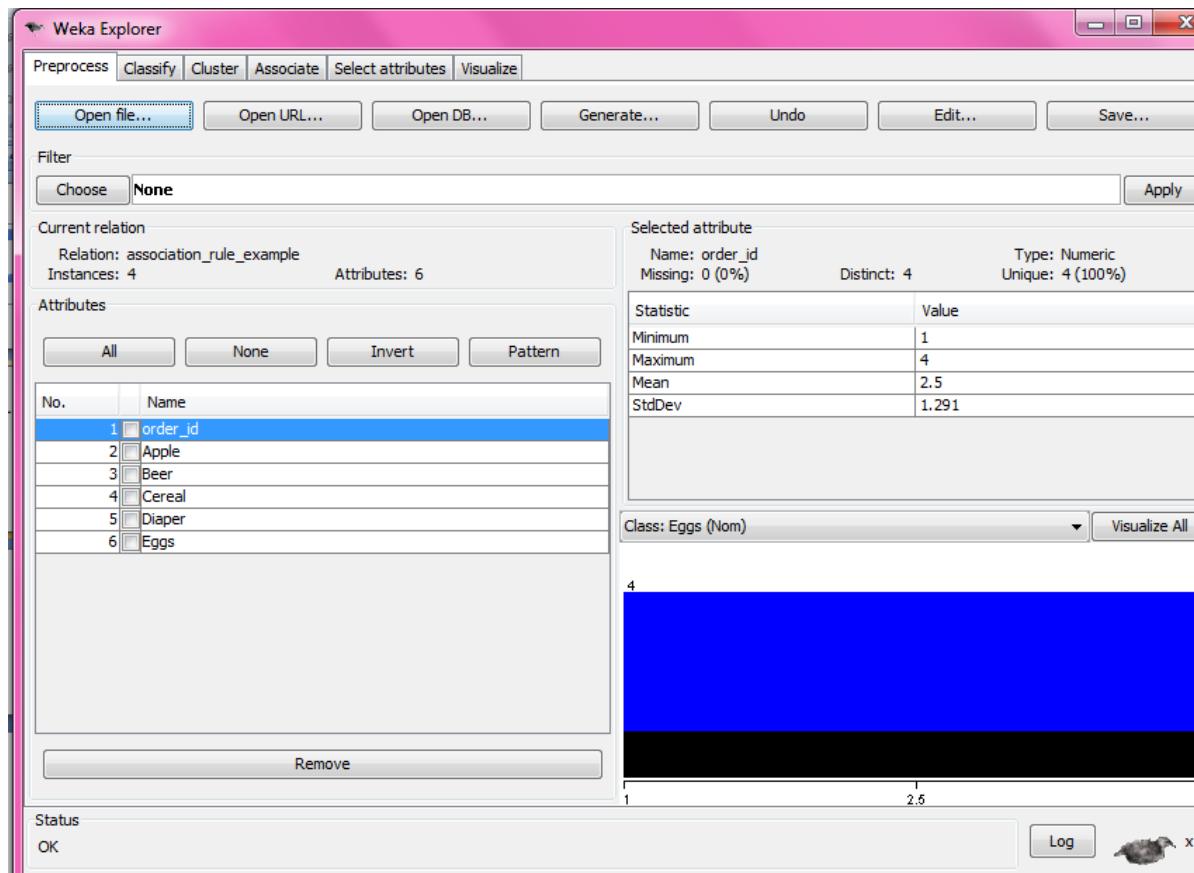
## ○ แสดงผลการหากฎความสัมพันธ์

- $\text{Eggs=y} \Rightarrow \text{Beer=y}$  conf:1  
“ทุกครั้งที่ลูกค้าซื้อไข่ไก่แล้วจะซื้อบีฟร์ด้วย”
- $\text{Apple Beer=y} \Rightarrow \text{Egg=y}$  conf:1  
“ทุกครั้งที่ลูกค้าซื้อแอปเปิลและเบียร์แล้ว จะซื้อไข่ไก่ด้วย”

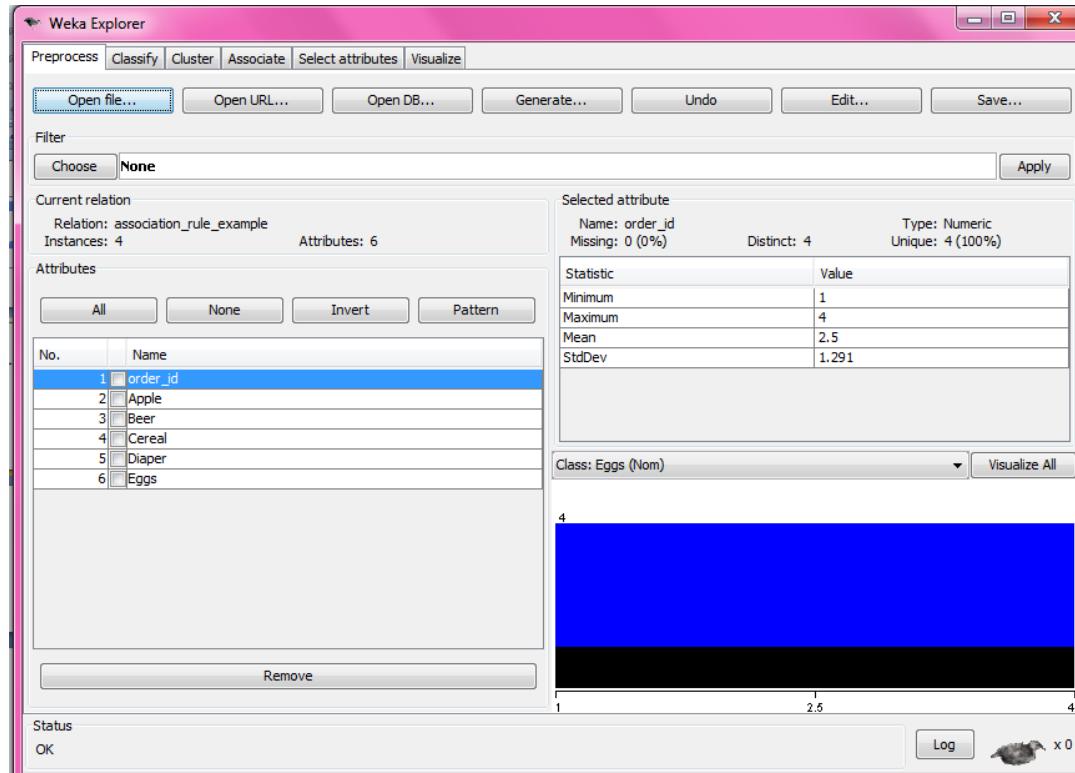
QID	Items
1	Apple, Ceral, Diapers
2	Beer, Cereal, Eggs
3	Apple, Beer, Cereal, Eggs
4	Beer, Eggs

# EXAMPLE : ASSOCIATION RULES

- เปิด weka > เลือก Explorer > กดปุ่ม Open file... > เลือกไฟล์ dataset\Example1.arff

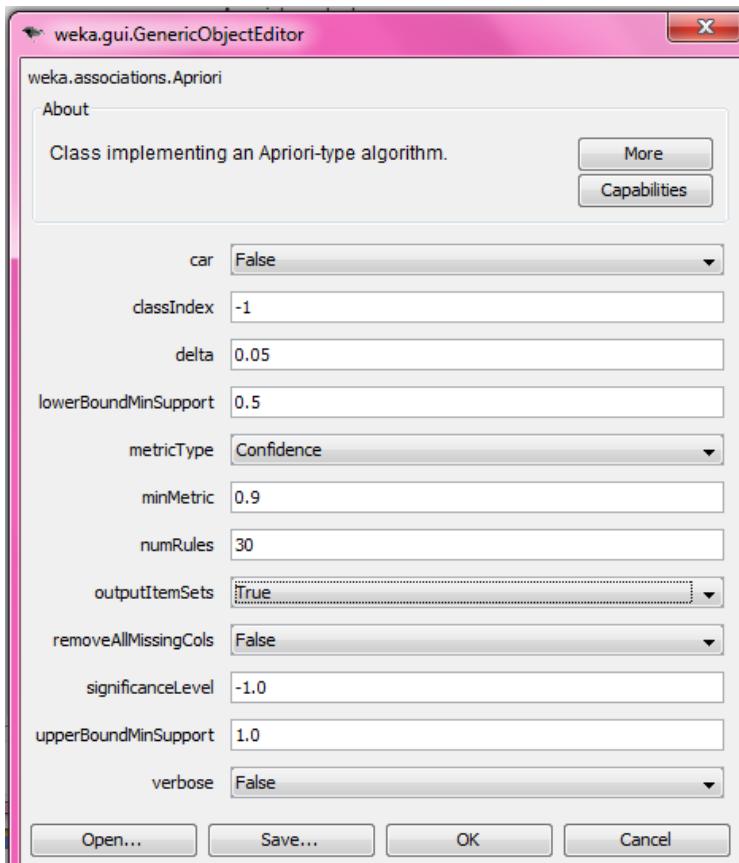


# EXAMPLE : ASSOCIATION RULES



- Remove แอ็ตทริบิวต์ที่ไม่สำคัญออก
  - คลิกที่ order\_id
  - คลิกที่ Remove

# EXAMPLE : ASSOCIATION RULES



- เปลี่ยนค่า lowerBoundMinSupport เป็น 0.5
- เปลี่ยนค่า minMetric เป็น 0.5
- เปลี่ยนค่า numRules เป็น 30
- เลือก OutputItemSets เป็น True
- กดปุ่ม OK
- กดปุ่ม Start

# EXAMPLE : ASSOCIATION RULES

The image displays two instances of the Weka Explorer interface, each showing the output of an Apriori algorithm run on a dataset. The left window shows the initial results, and the right window shows the final results after processing all cycles.

**Weka Explorer (Left Window):**

- Preprocess, Classify, Cluster, Associate (selected), Select attributes, Visualize tabs.
- Associate button: Choose Apriori -I -N 30 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1.
- Start, Stop buttons.
- Result list (right-click for context menu):
  - Minimum support: 0.5 (2 instances)
  - Minimum metric <confidence>: 0.5
  - Number of cycles performed: 10
- Generated sets of large itemsets:
  - Size of set of large itemsets L(1): 4
    - Large Itemsets L(1):
      - Apple=y 2
      - Beer=y 3
      - Cereal=y 3
      - Eggs=y 3
  - Size of set of large itemsets L(2): 4
    - Large Itemsets L(2):
      - Apple=y Cereal=y 2
      - Beer=y Cereal=y 2
      - Beer=y Eggs=y 3
      - Cereal=y Eggs=y 2
  - Size of set of large itemsets L(3): 1
- Status: OK

**Weka Explorer (Right Window):**

- Preprocess, Classify, Cluster, Associate (selected), Select attributes, Visualize tabs.
- Associate button: Choose Apriori -I -N 30 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1.
- Start, Stop buttons.
- Result list (right-click for context menu):
  - Cereal=y Eggs=y 2
- Size of set of large itemsets L(3): 1
- Large Itemsets L(3):
  - Beer=y Cereal=y 2
  - Eggs=y 2
- Best rules found:
  - Eggs=y 3 ==> Beer=y 3 conf:(1)
  - Beer=y 3 ==> Eggs=y 3 conf:(1)
  - Apple=y 2 ==> Cereal=y 2 conf:(1)
  - Cereal=y Eggs=y 2 ==> Beer=y 2 conf:(1)
  - Beer=y Cereal=y 2 ==> Eggs=y 2 conf:(1)
  - Cereal=y 3 ==> Apple=y 2 conf:(0.67)
  - Cereal=y 3 ==> Beer=y 2 conf:(0.67)
  - Beer=y 3 ==> Cereal=y 2 conf:(0.67)
  - Eggs=y 3 ==> Cereal=y 2 conf:(0.67)
  - Cereal=y 3 ==> Eggs=y 2 conf:(0.67)
  - Beer=y Eggs=y 3 ==> Cereal=y 2 conf:(0.67)
  - Eggs=y 3 ==> Beer=y Cereal=y 2 conf:(0.67)
  - Cereal=y 3 ==> Beer=y Eggs=y 2 conf:(0.67)
  - Beer=y 3 ==> Cereal=y Eggs=y 2 conf:(0.67)
- Status: OK

# LAB 7-1 : MARKET BASKET

## ○ Business Understanding

- ชูปเปอร์มาร์เก็ตแห่งหนึ่งต้องการทำระบบ CRM กับลูกค้าที่เข้ามาซื้อสินค้าโดยต้องการหาว่ามีสินค้าชนิดใดบ้างที่ลูกค้ามักจะซื้อพร้อมกันบ่อย ๆ เพื่อนำไปจัดโปรโมชั่น

## ○ Data Understanding

- ชูปเปอร์มาร์เก็ตแห่งนี้ได้ทำการเก็บประวัติการซื้อสินค้าของลูกค้าจำนวน 1,000 คน
- โดยข้อมูลของลูกค้าแต่ละรายจะแบ่งเป็น 2 ส่วนใหญ่ ๆ คือ
  - ข้อมูลรายละเอียดเกี่ยวกับลูกค้าแต่ละรายมีจำนวน.....แอตทริบิวต์
  - ข้อมูลสินค้าที่ลูกค้าซื้อแต่ละครั้งมีจำนวน .....แอตทริบิวต์
- ข้อมูลเหล่านี้ได้ถูกรวบรวมเก็บไว้ในไฟล์ supermarket\_basket\_transactions\_2005.arff ในแผ่น CD ไฟล์เดอร์ dataset/supermarket

