

# การฝึกหัด

Introduction to Data Mining I

# Classification

1. Decision tree
2. Naive Bayes

# Decision tree (J48)

- J48 ในโปรแกรม Weka ใช้อัลกอริทึม C4.5 ในการสร้างโมเดล
- Tree model: binary tree
- C4.5 คิดค้นโดย Ross Quinlan (1993)
- C4.5 เลือก attribute จาก information gain

*information gain = entropy ก่อนแบ่ง - entropy หลังแบ่ง*

$$\begin{aligned} \text{entropy}(p_1, p_2, \dots, p_n) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 \dots - p_n \log_2 p_n \\ &= -\sum p_i \log_2 p_i \end{aligned}$$

# To use Decision tree

- Known relationship between attributes and class
- Experts consensus
- Most of possible cases (>80%?) are available
- Measured values can be grouped to name?
- No conflicts between cases
- False cases leads to wrong results

# ព័វិសាយទីផ្សារ

1. Sunburn
2. Weather
3. Fruit

# ចុះបញ្ជីអនុលោ : Sunburn

	សិរី	សំណង	ឆាន់អាណក	ទាល់ខ័ណ្ឌ	ពិវិជ្ជម៉ោ
	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
3	brown	short	average	yes	none
4	blonde	short	average	no	sunburned
5	red	average	heavy	no	sunburned
6	brown	tall	heavy	no	none
7	brown	average	heavy	no	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none

# Sunburn.arff

```
@relation Sunburn

@attribute Hair {blonde,brown,red}
@attribute Height {average,tall,short}
@attribute Weight {light,average,heavy}
@attribute Lotion {no,yes}
@attribute Result {sunburned,none}

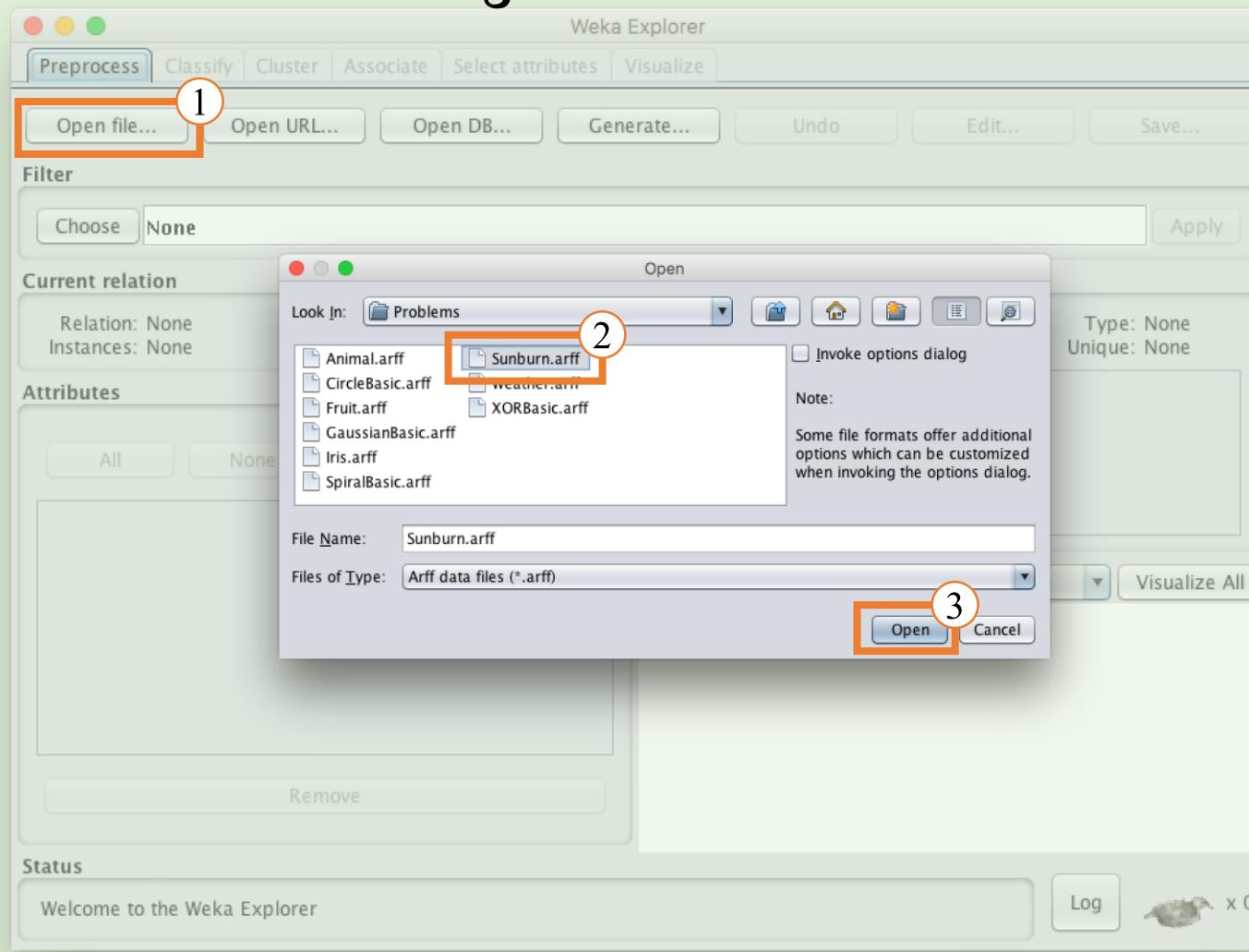
@data
blonde,average,light,no,sunburned
blonde,tall,average,yes,none
brown,short,average,yes,none
blonde,short,average,no,sunburned
red,average,heavy,no,sunburned
brown,tall,heavy,no,none
brown,average,heavy,no,none
blonde,short,light,yes,none
red,short,light,yes,sunburned
blonde,short,heavy,yes,none
red,tall,average,no,sunburned
brown,tall,light,yes,none
```

# Hair? Weight? Height?

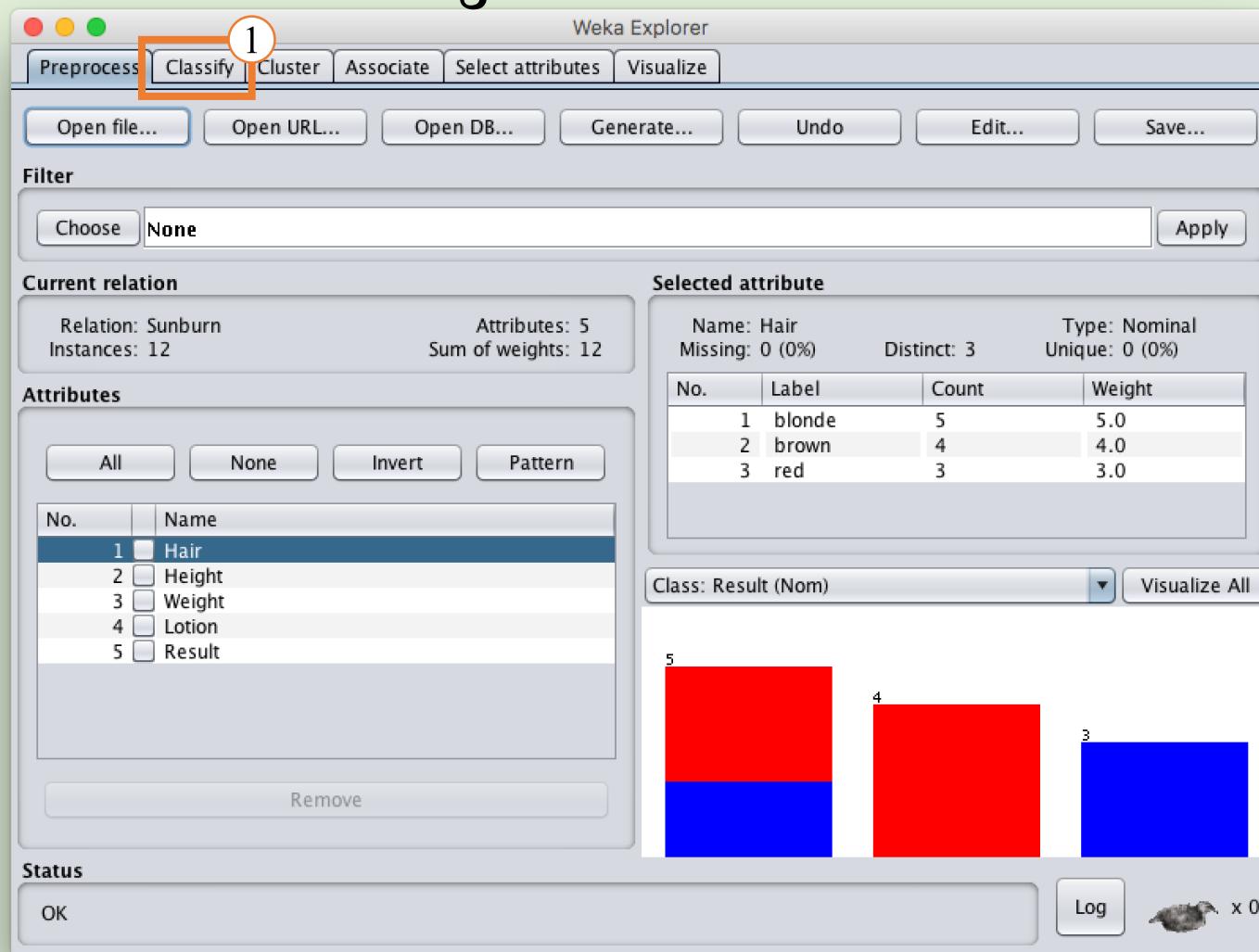
- Attributes directly or indirectly relate with sunburns?
- Number of samples cover most possible cases?
- Any conflict between cases?
- Experts agree on the cases' results?
- What if input cases are unknown to experts?



# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



# Attribute analysis

Selected attribute

Name: Hair

Type: Nominal

Missing: 0 (0%)

Distinct: 3

Unique : 0 (100%)

Label {blonde, brown, red}

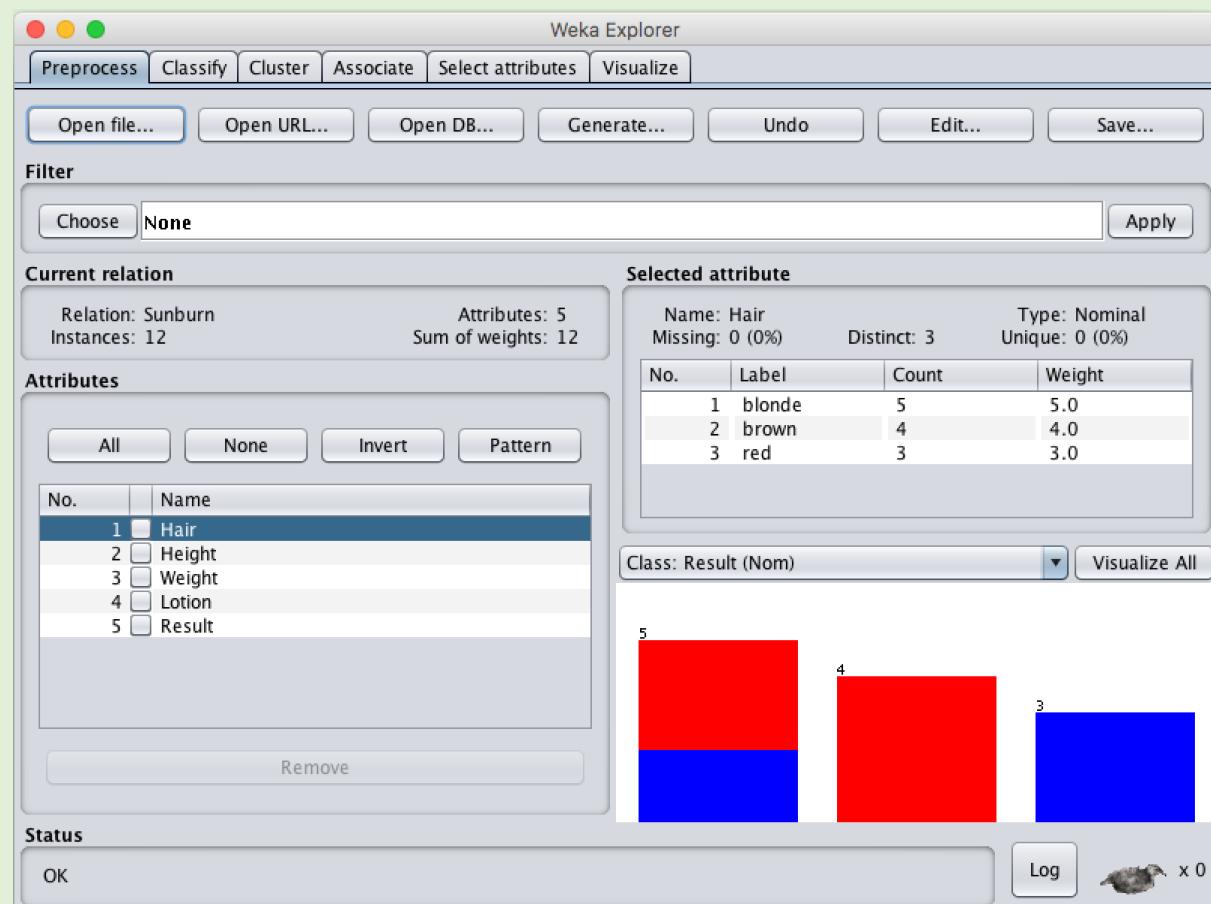
Count : {5,4,3}

Weight : {5.0,4.0,3.0}

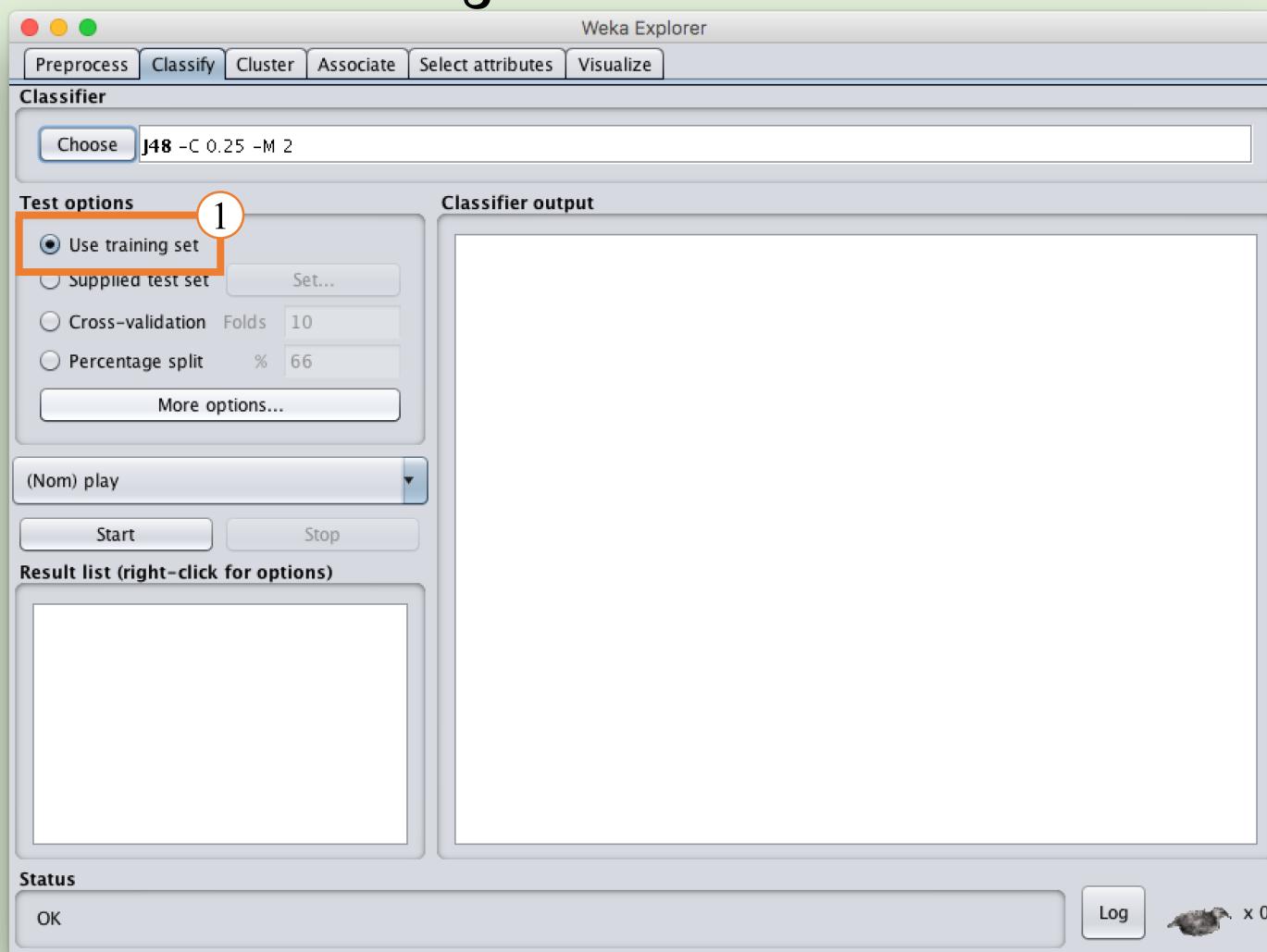
Class: Lable (Nom)

Attribute : (5, 4, 3)

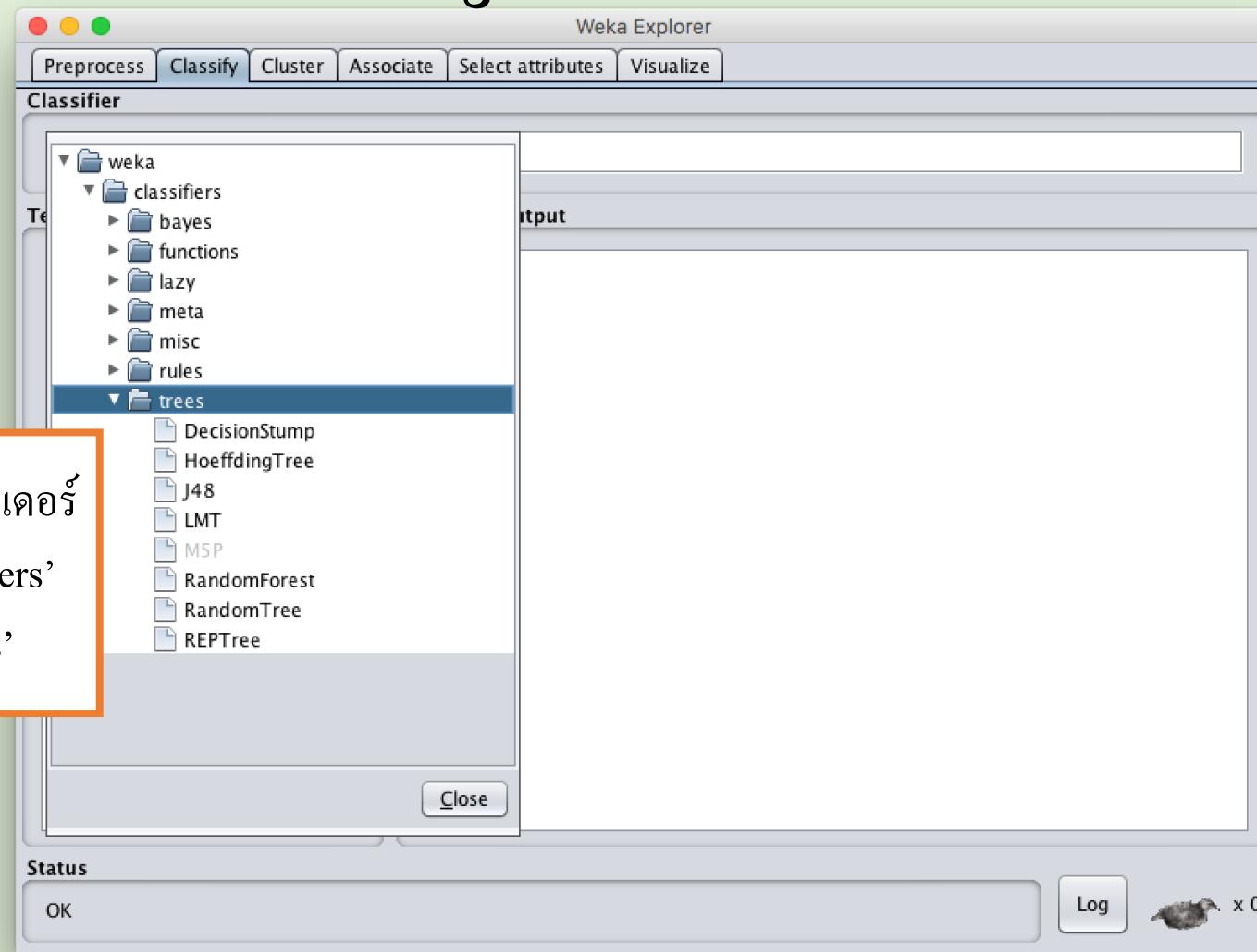
Histogram : Value – Class color



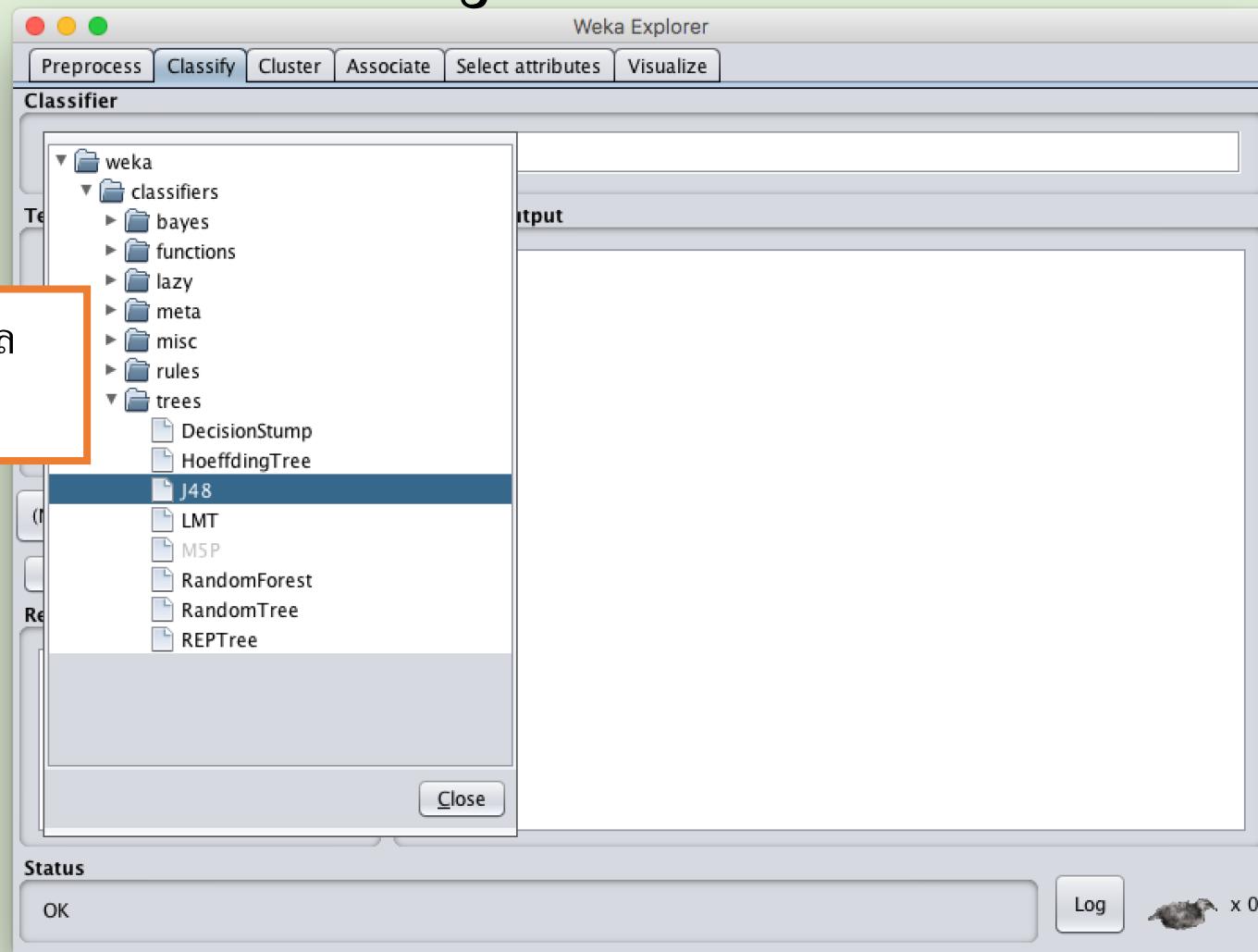
# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



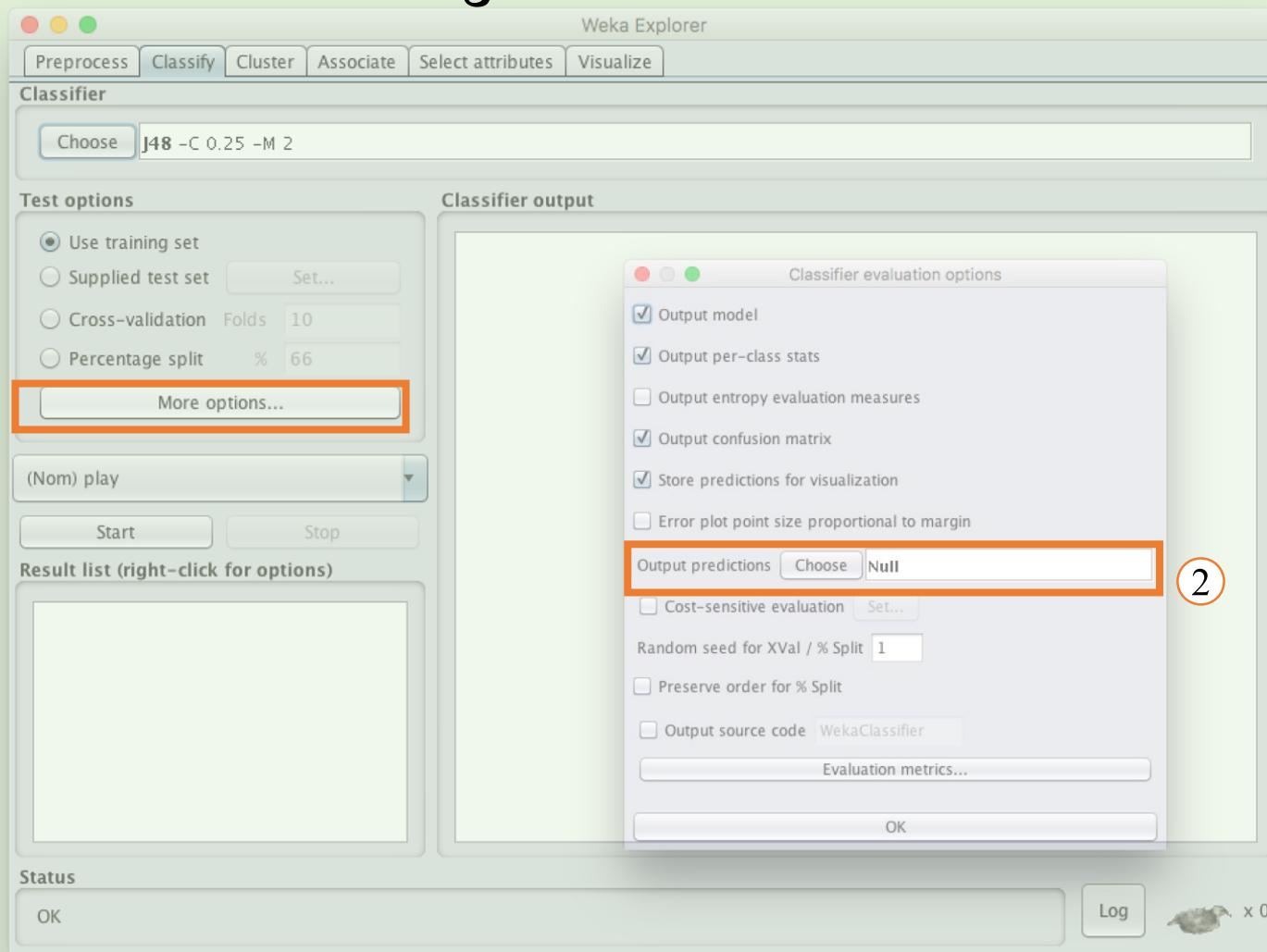
# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



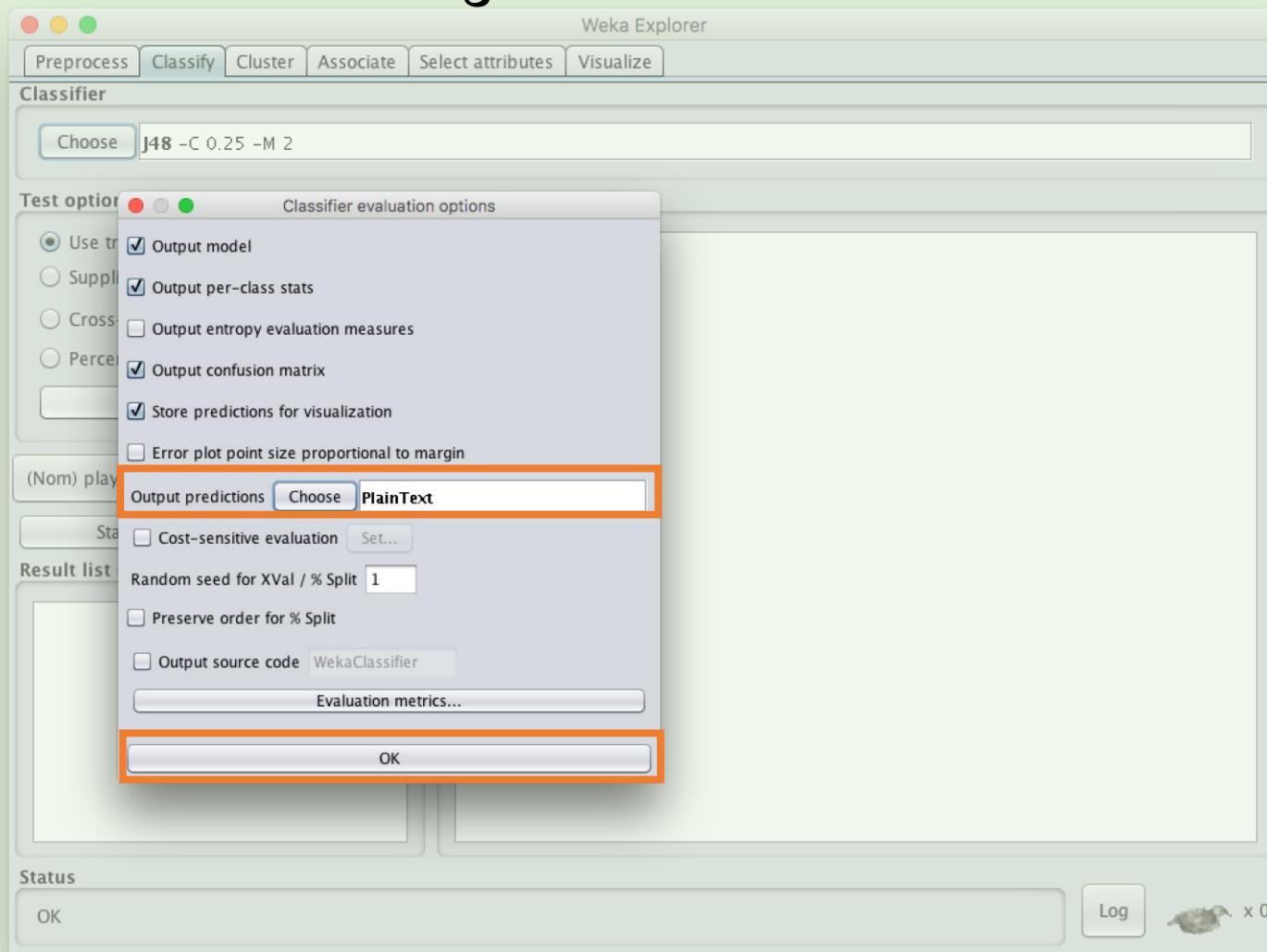
# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



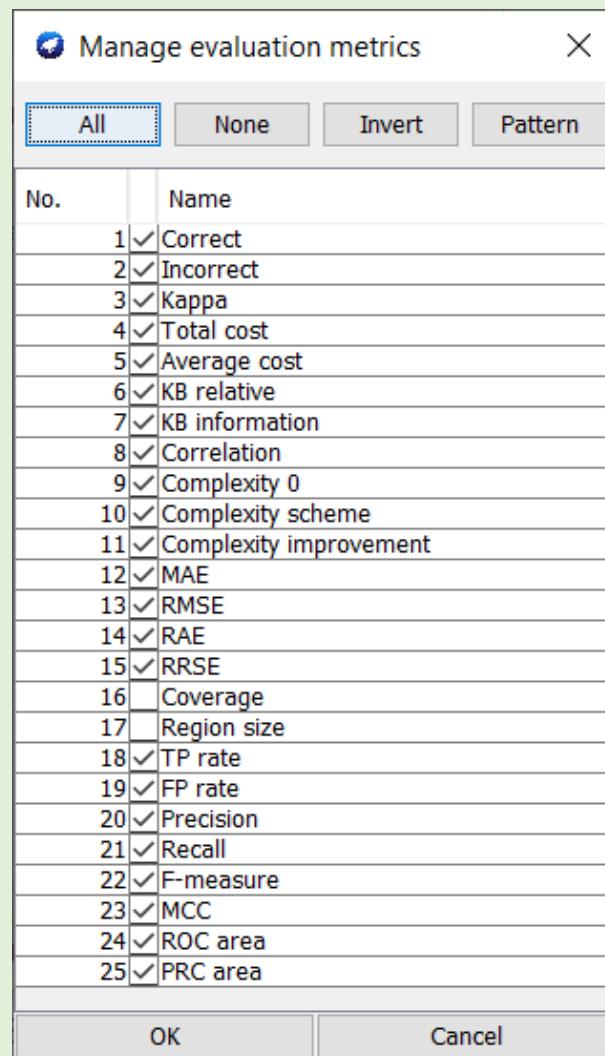
# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



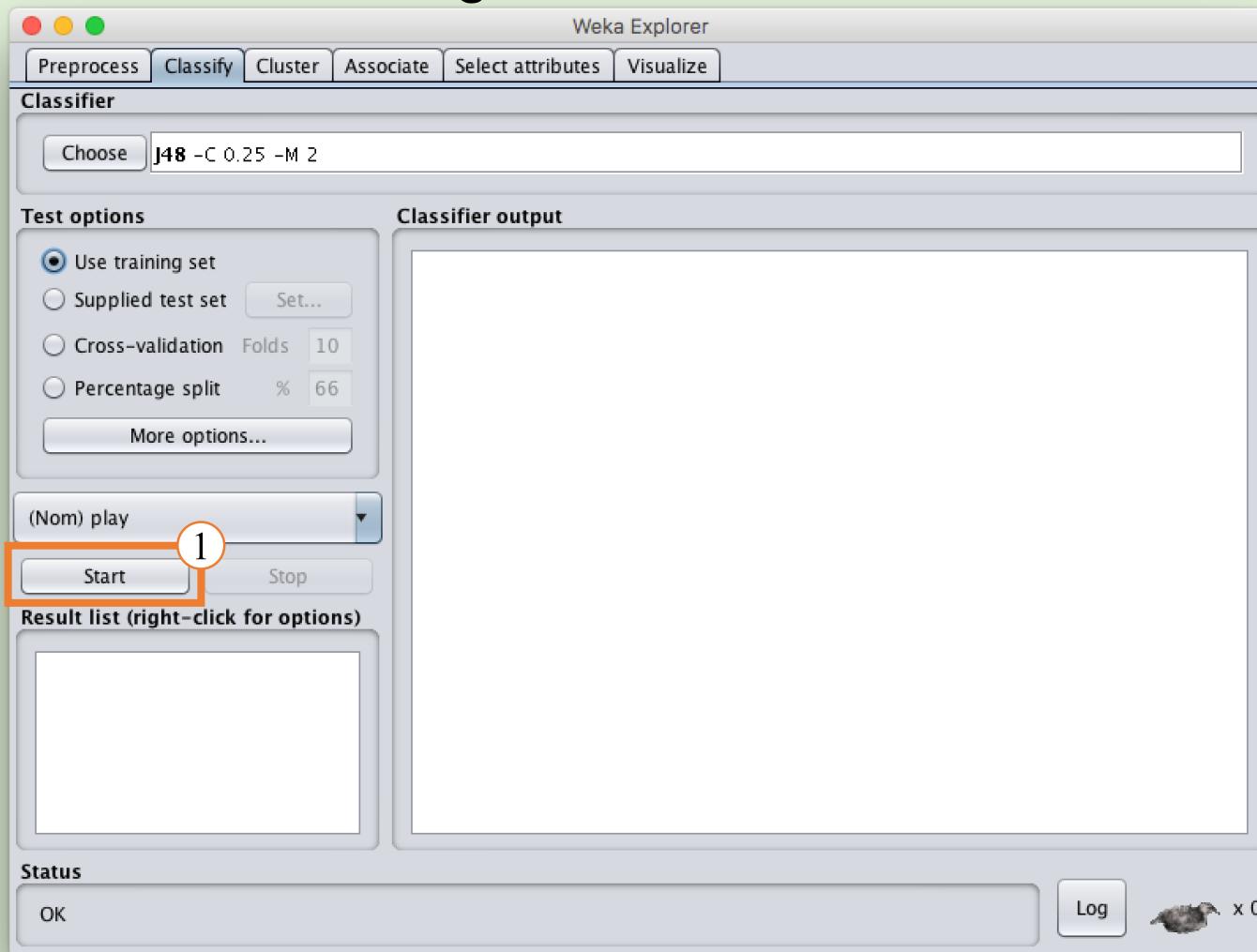
# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



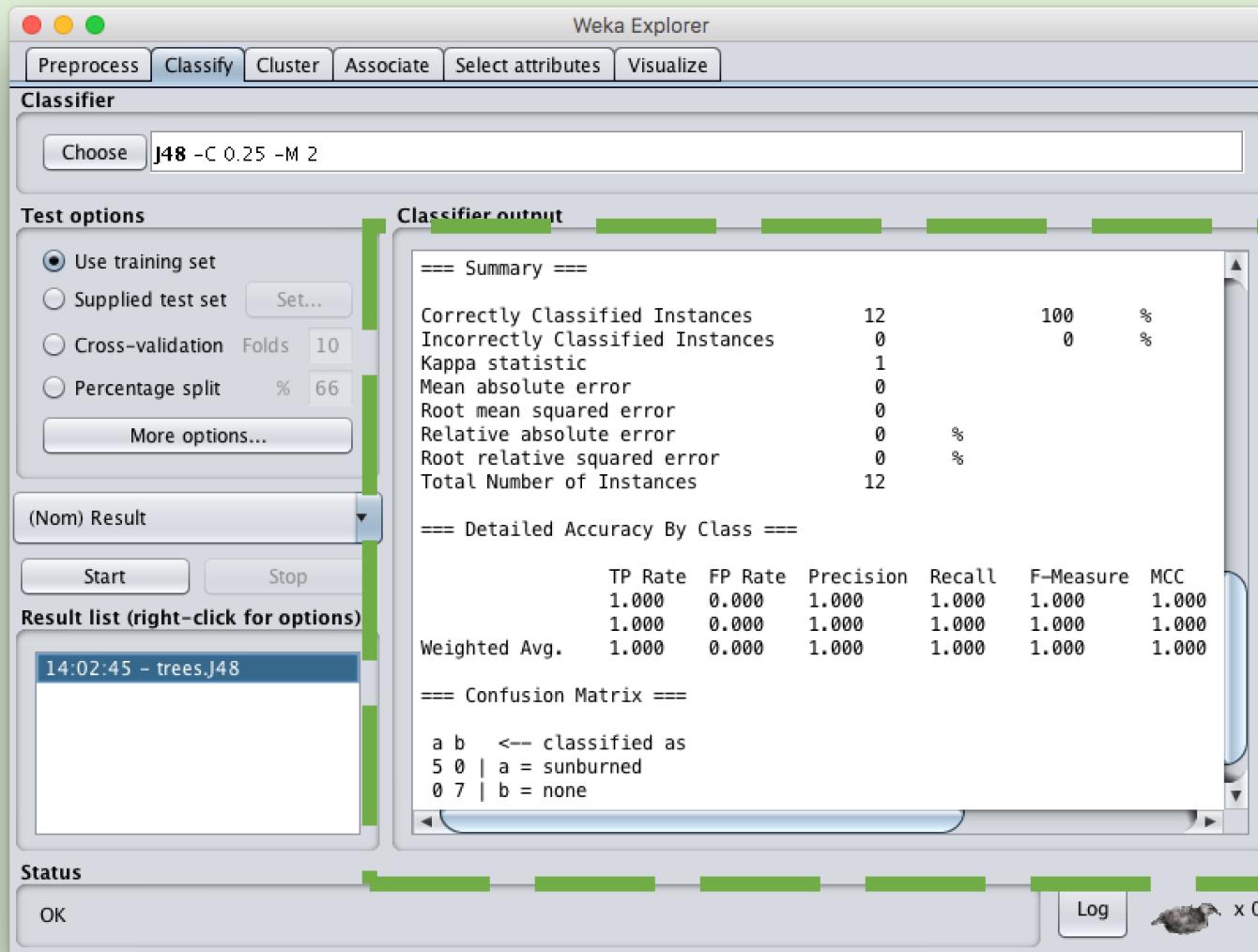
# Manage Evaluation Metrics



# ขั้นตอนการสร้าง Decision tree สำหรับปัญหา Sunburn



# ผล Classification ของปัญหา Sunburn ด้วย Decision tree



# การวิเคราะห์ประสิทธิภาพของ Classifier

- TP : Ture Positive
- FP : False Positive
- Precision :
- Recall :
- F-Measure :
- MCC :
  
- Confusion Matrix :

# การวิเคราะห์ผลของปัญหา Sunburn ด้วย Decision tree

สรุปเป็น Confusion Matrix ได้ดังนี้

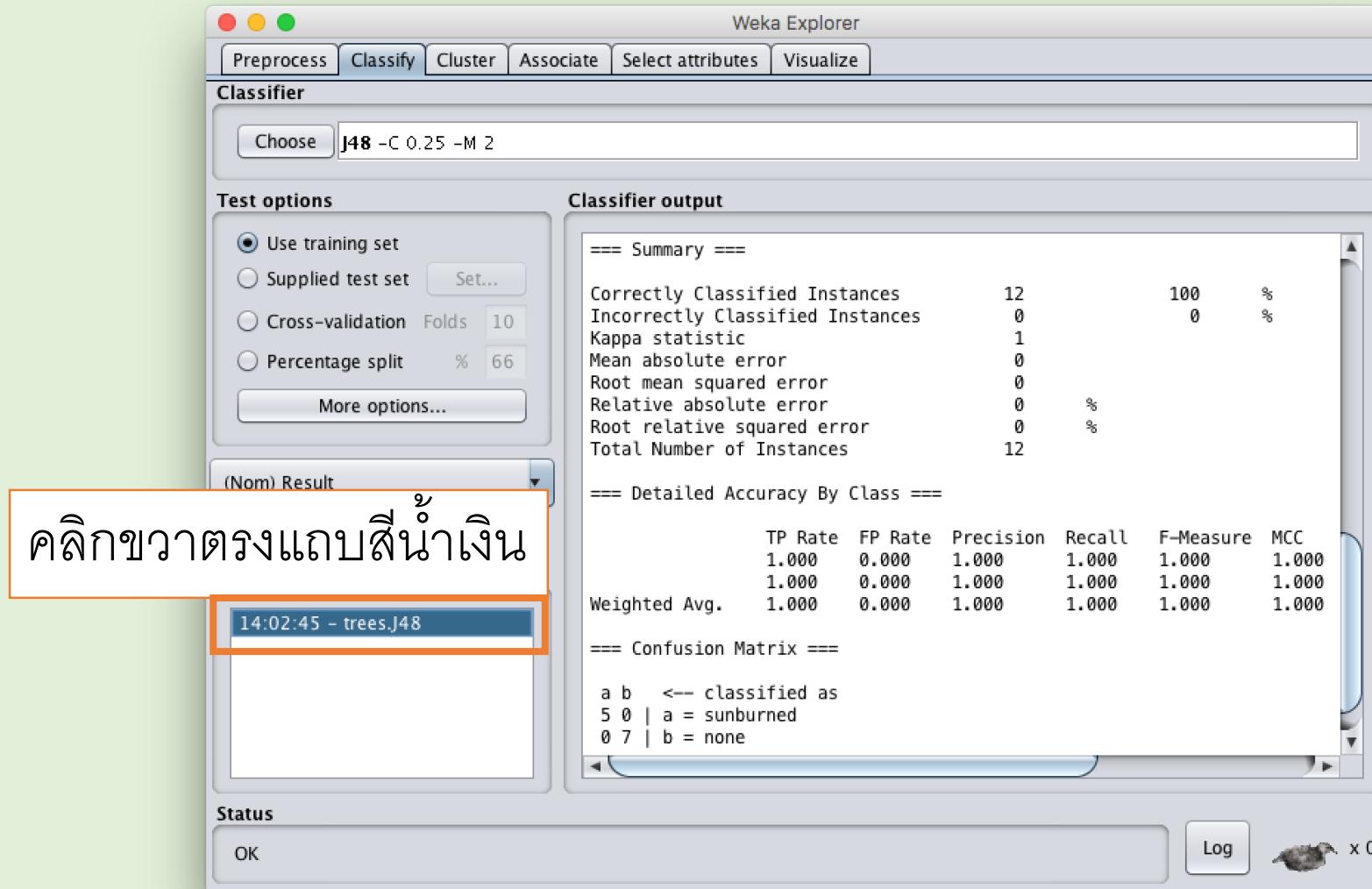
ไม่เดลท่านาย

1.ค่าของข้อมูลจริง

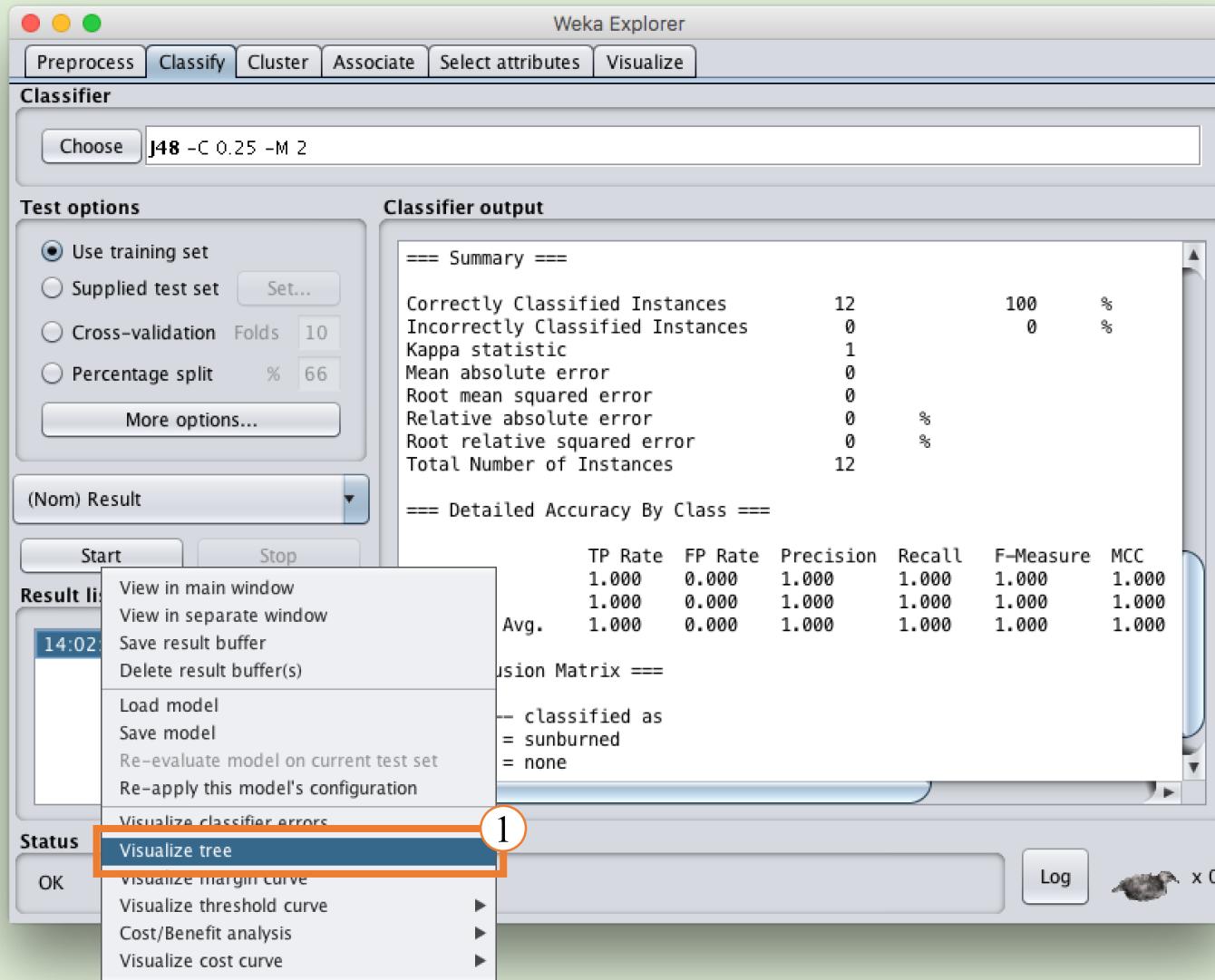
2.classified as =>	sunburned	none
sunburned	5	0
none	0	7



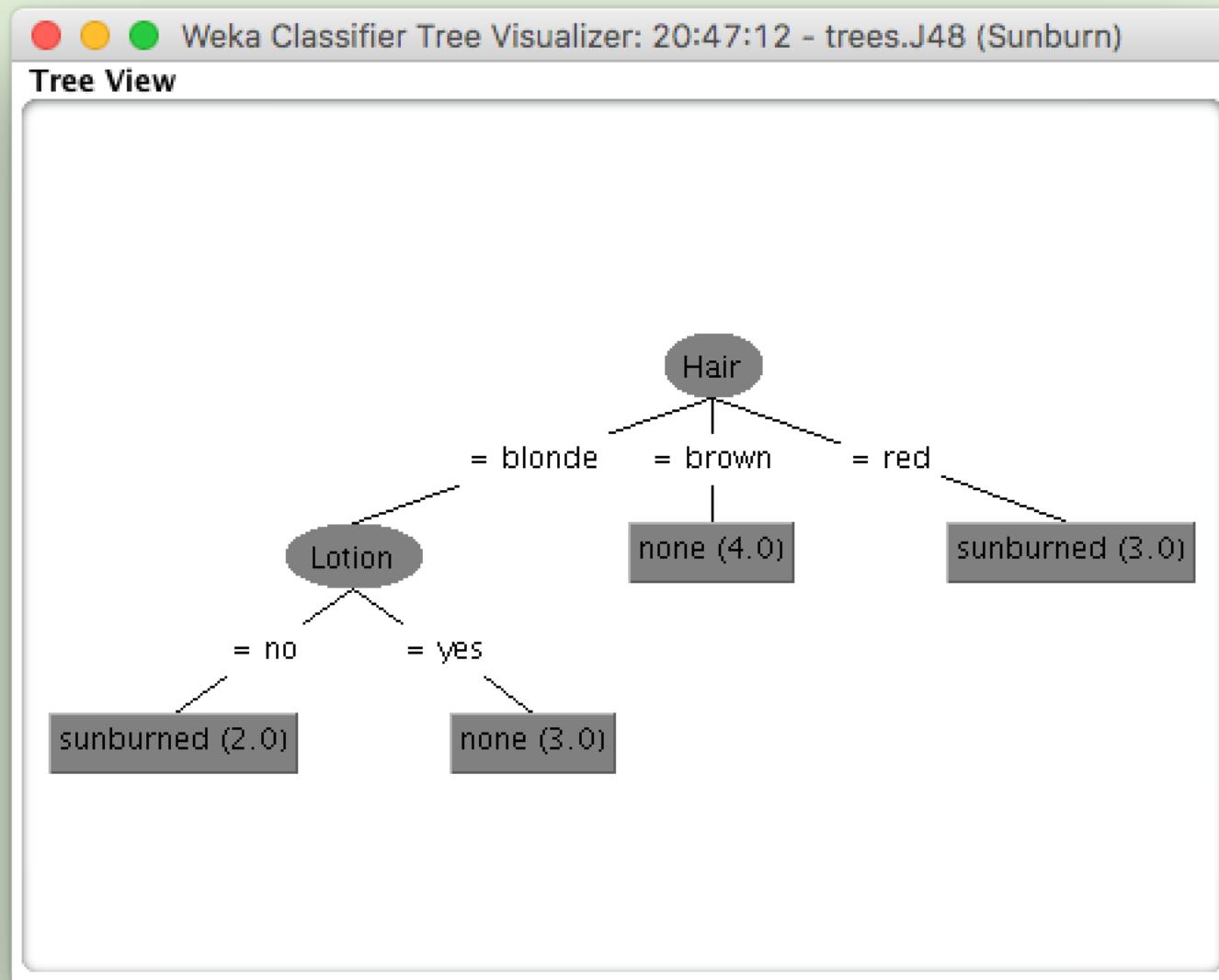
# การ Visualize โมเดล Decision tree



# การ Visualize ไม้เดล Decision tree



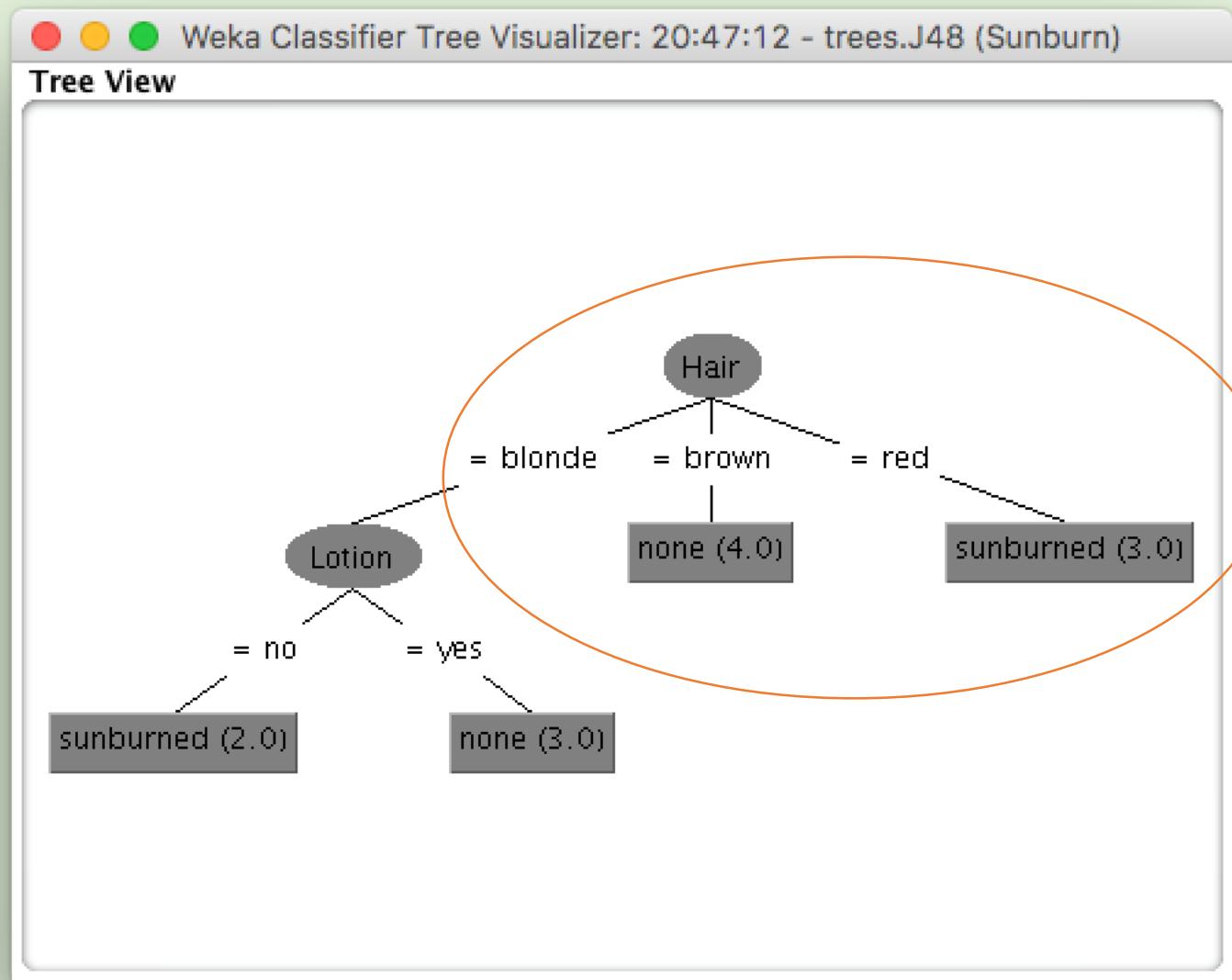
# ໂຄນເດລ Decision tree ຂອງປິ່ງຫາ Sunburn



# การวิเคราะห์ผลของปัญหา Sunburn ด้วย Decision tree

- ไม่เดลีมีความแม่นยำในอยู่ที่ 100%
- จากไม่เดล Decision tree
  - Hair เป็น attribute ที่มีผลในการแบ่งกลุ่มข้อมูลออกจากกันมากที่สุด
  - สังเกตได้ว่า
    - ถ้า Hair เป็น brown สามารถตอบได้ว่า พิวของคนนั้นจะไม่ไหม้
    - ถ้า Hair เป็น red สามารถตอบได้ว่าพิวของคนนั้นจะไหม้

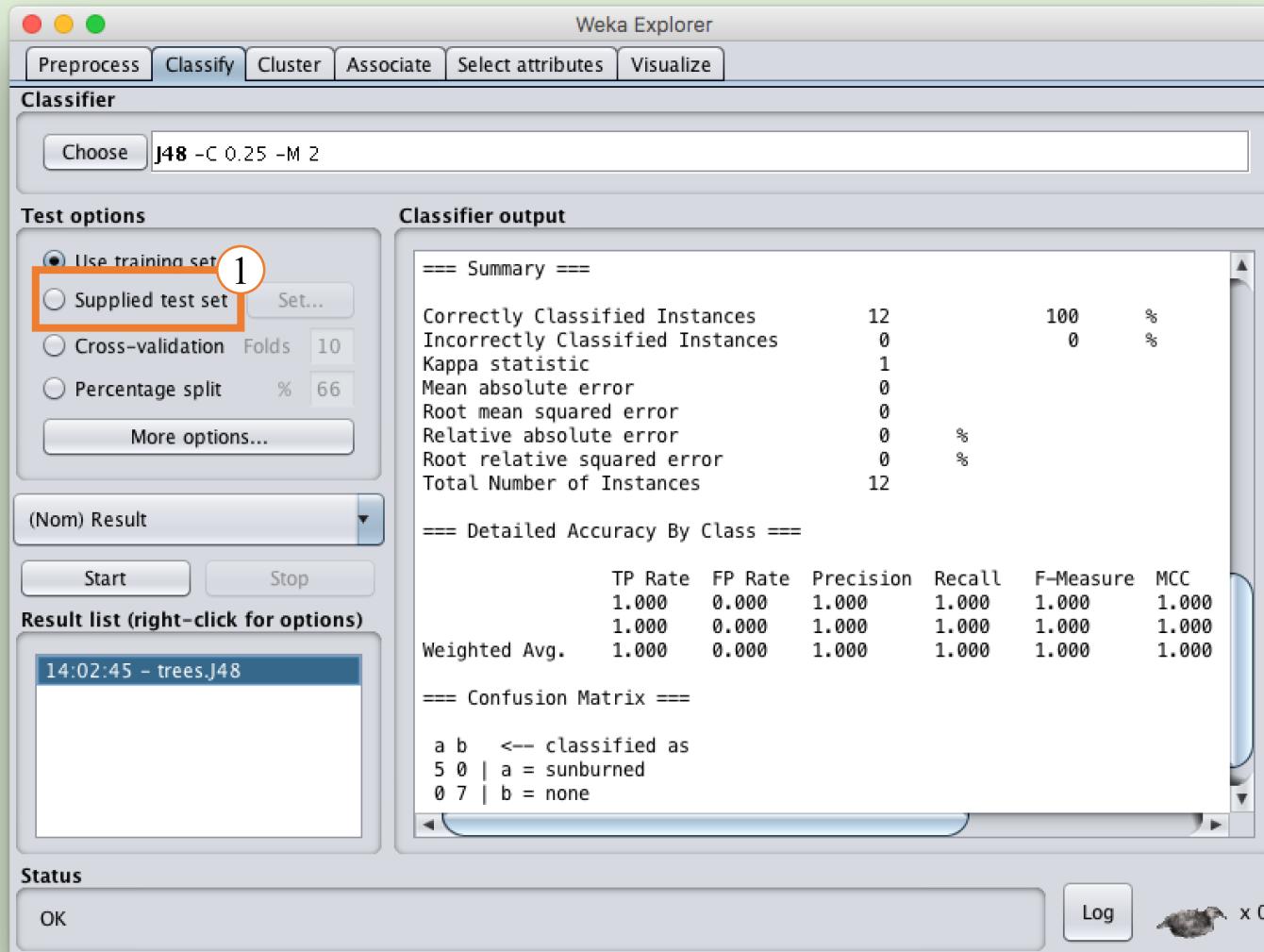
# ໂຄນເດລ Decision tree ຂອງປິ່ງຫາ Sunburn



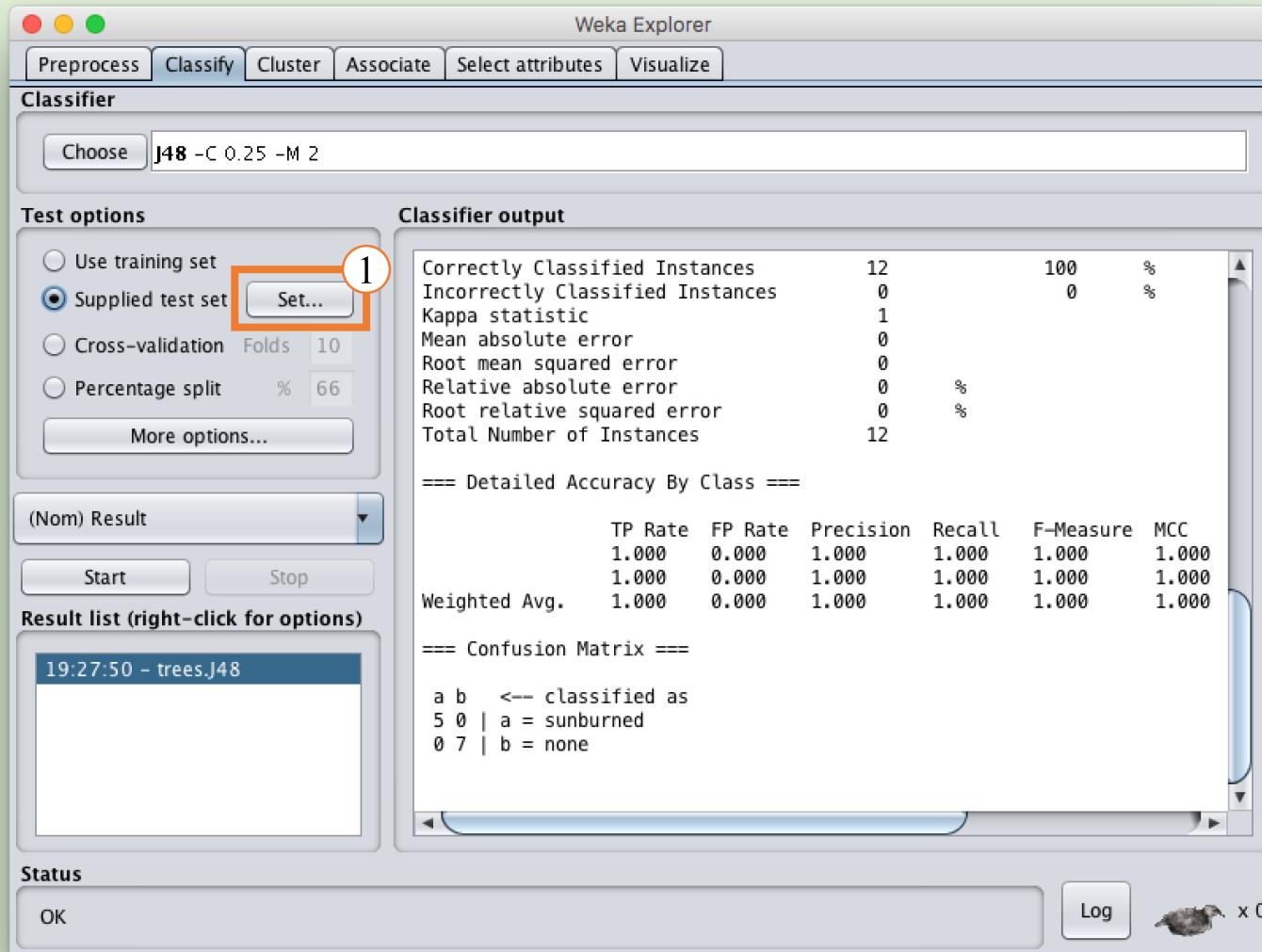
# การทำงานกับข้อมูลที่ไม่ทราบคำต่อไป

Hair	Height	Weight	Lotion	Result
red	average	light	no	?
brown	tall	light	no	?
blonde	short	heavy	no	?

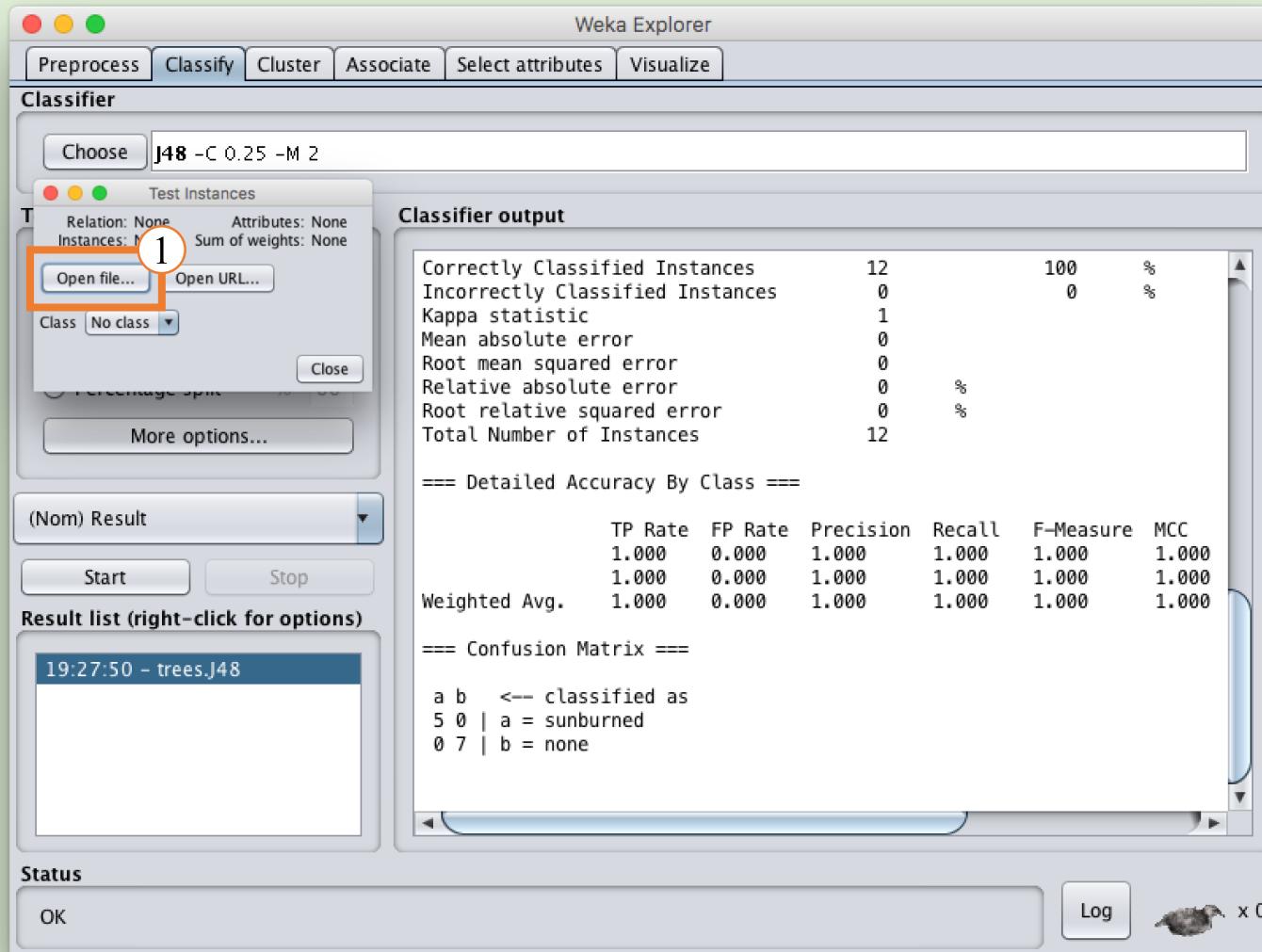
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



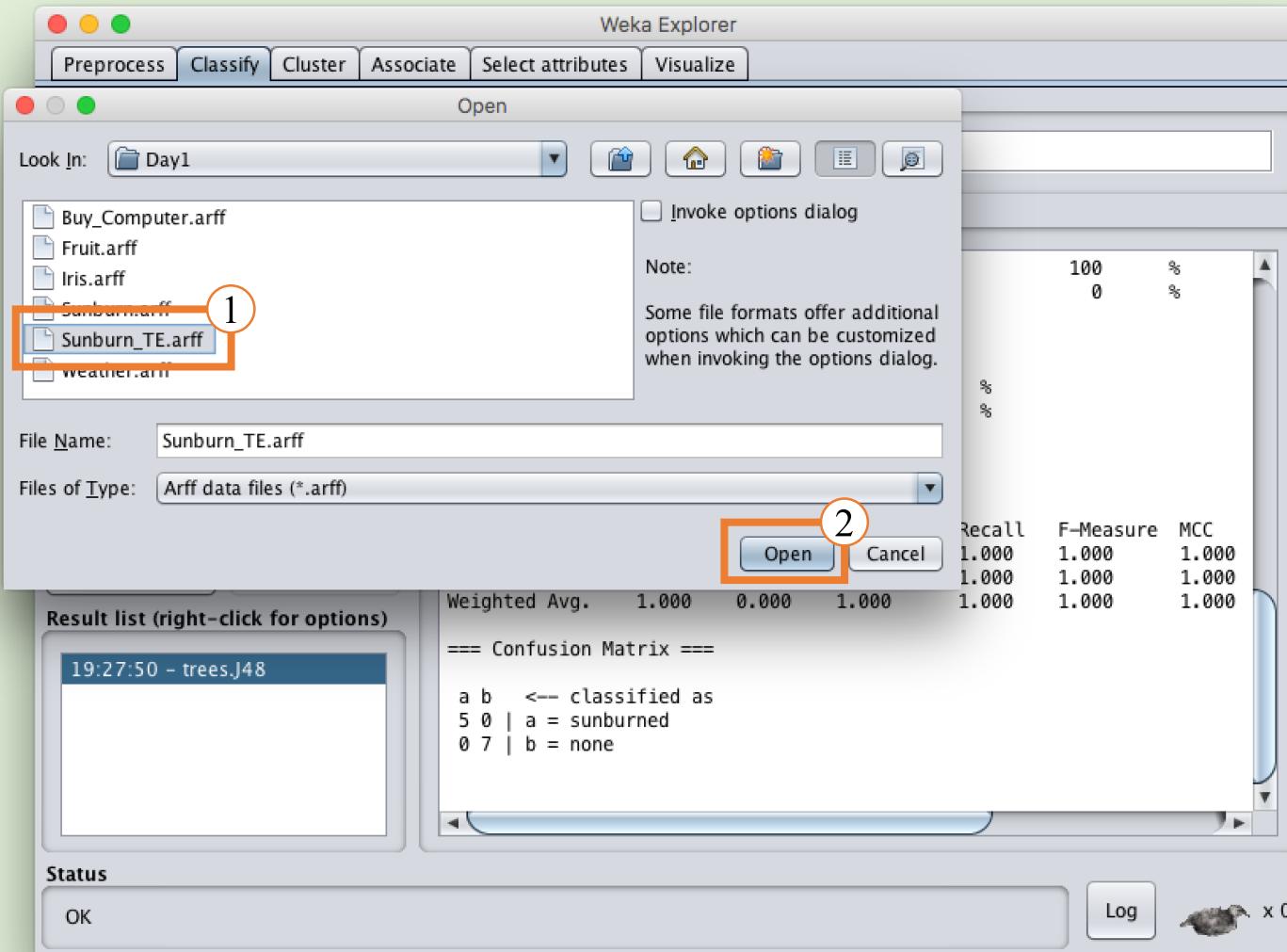
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



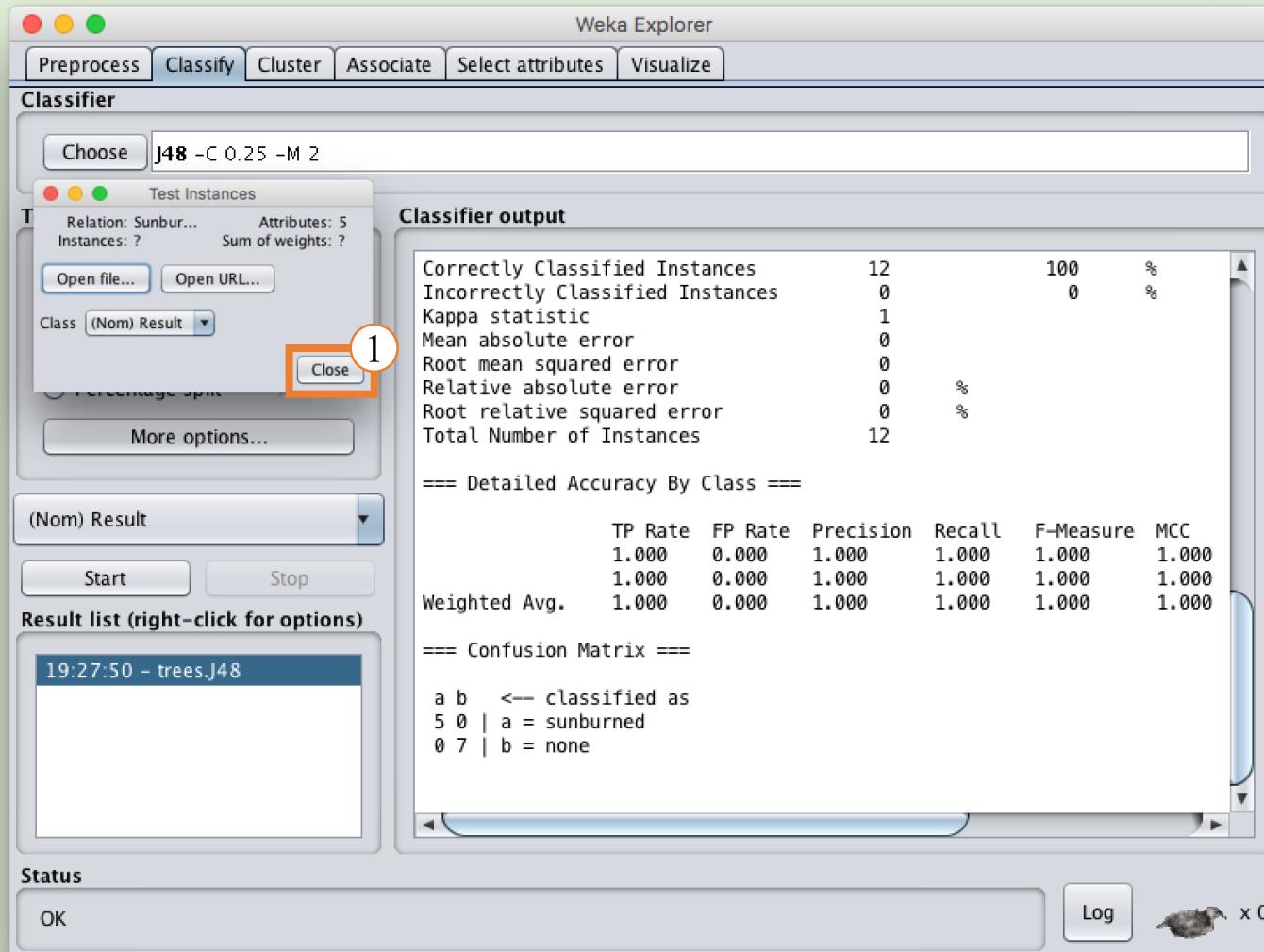
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



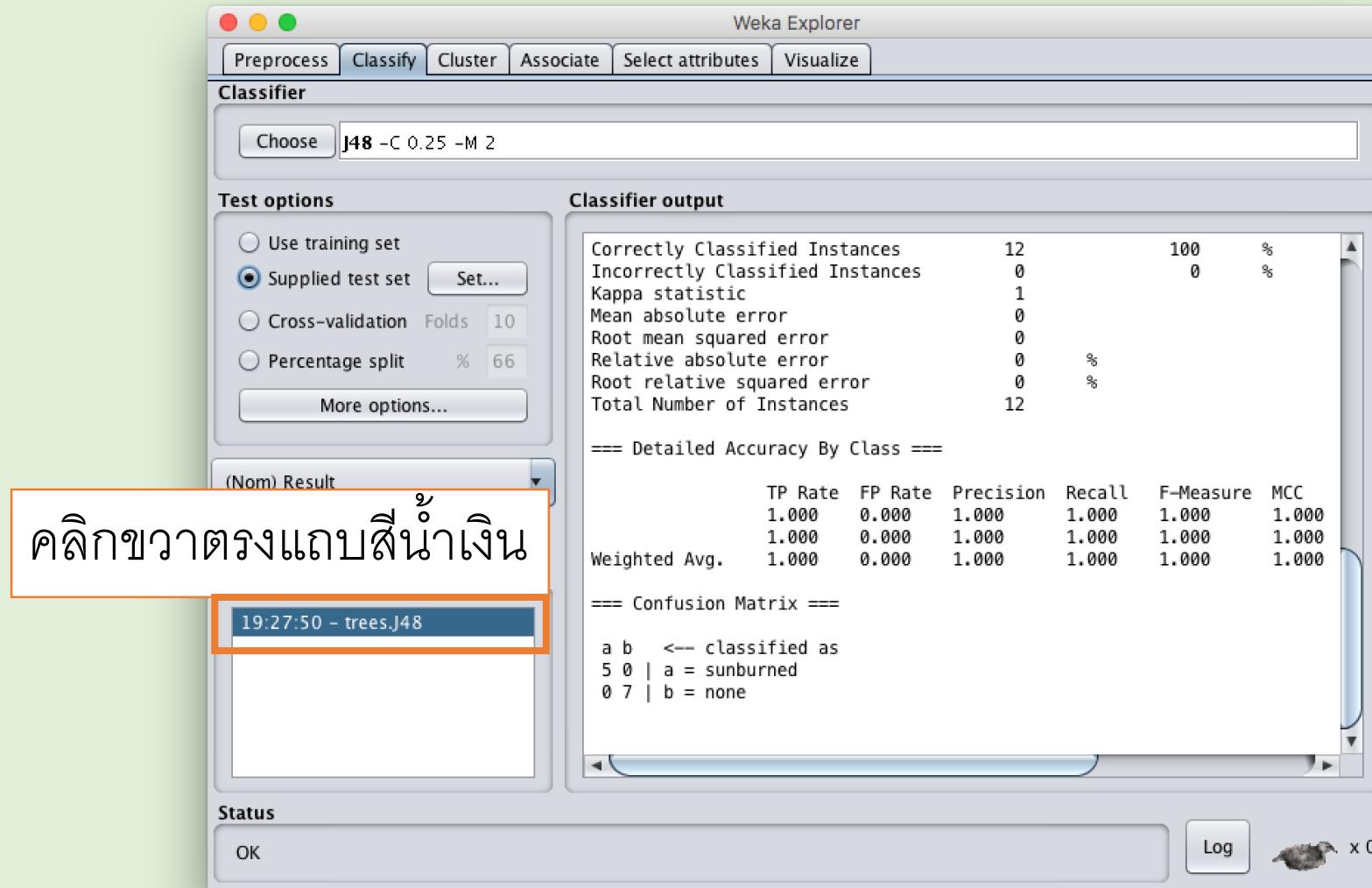
# การคัดกรณ์ข้อมูลที่ไม่ทราบคำต่อหน้า



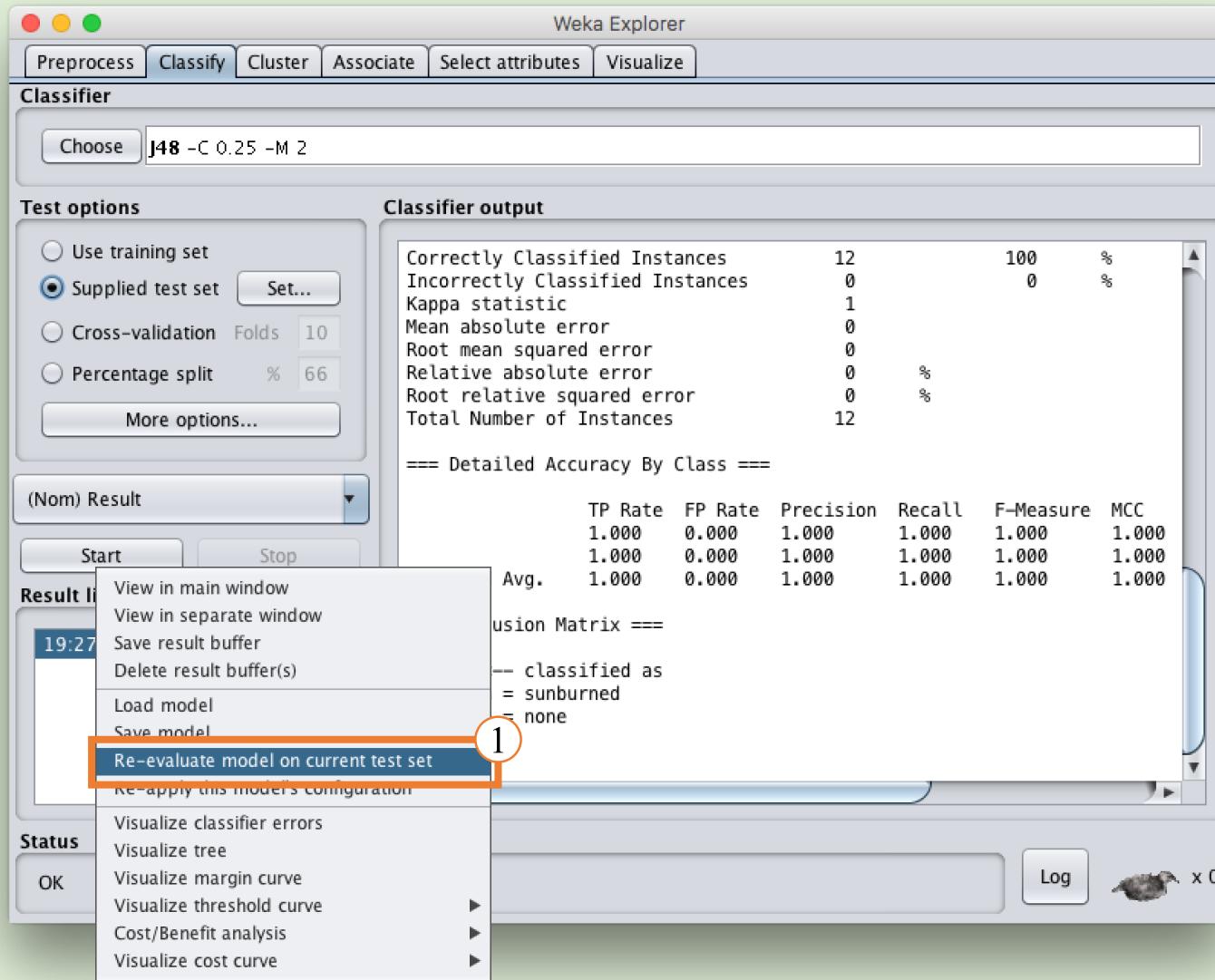
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



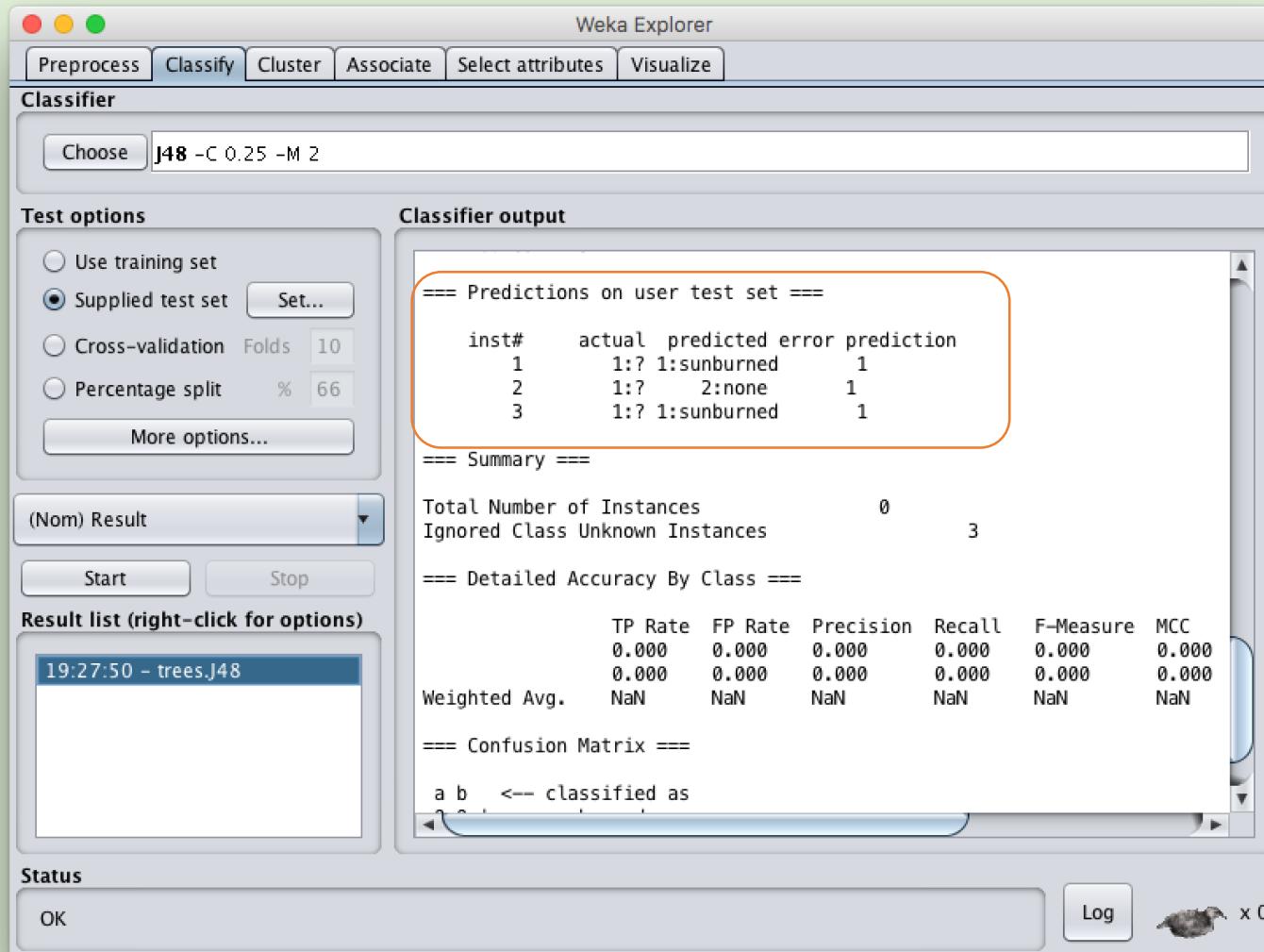
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



# การคาดการณ์ข้อมูลที่ไม่ทราบคำตอบ



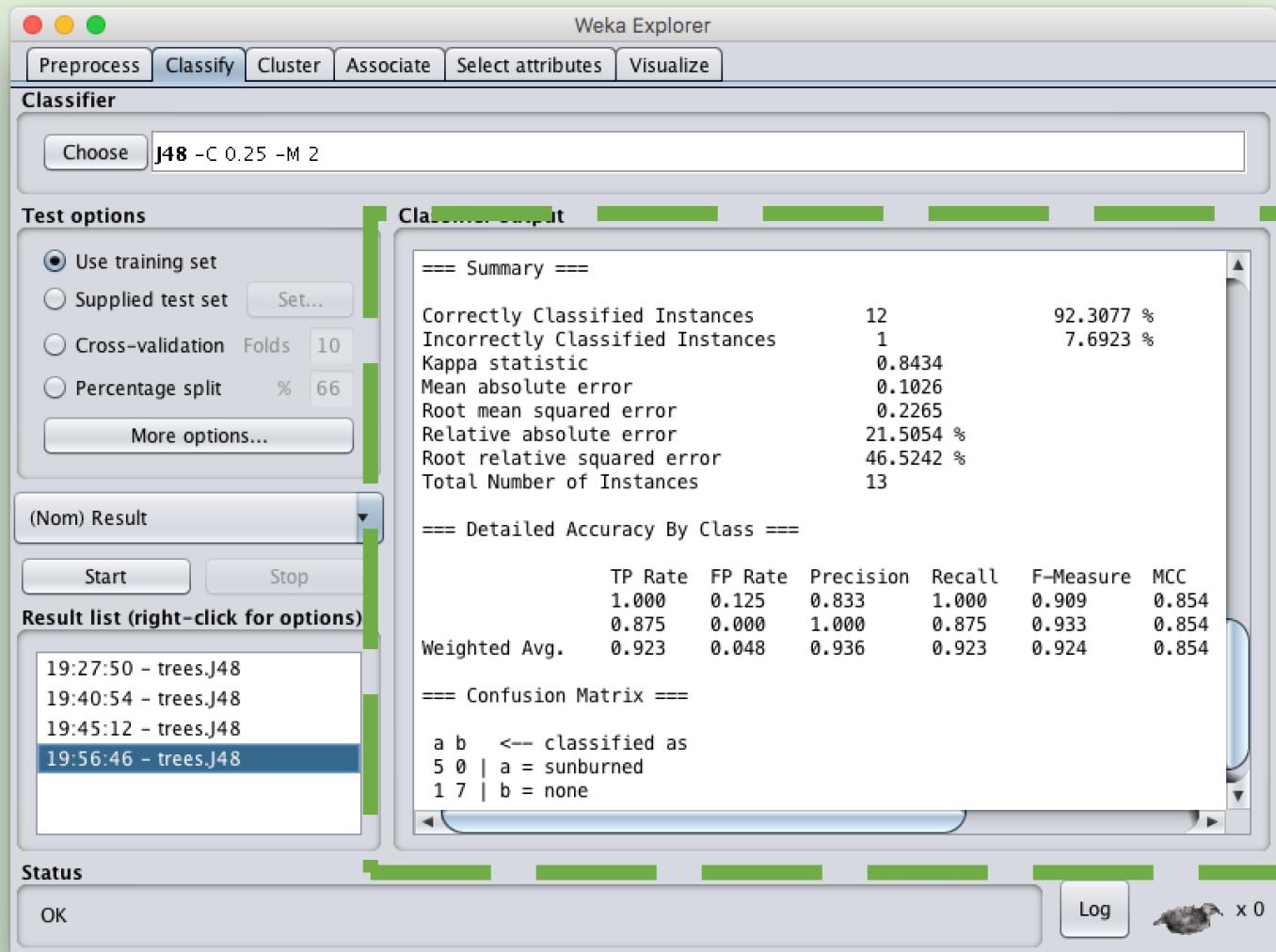
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



# What-if : เพิ่มข้อมูลใหม่ 1 ตัว ซึ่งมีผลไม่ตรงกันกับตัวอย่างเดิม

	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none
13	blonde	average	light	no	none

# ผล Classification ของ Decision tree



# Confusion Matrix

ไม่เดลทำนาย

classified as =>

	sunburned	none
sunburned	5	0
none	1	7

ค่าของข้อมูลจริง

# ผล Classification ของ Decision tree

== Predictions on training set ==

inst#	actual	predicted	error	prediction
1	1:sunburned	1:sunburned		0.667
2	2:none	2:none		1
3	2:none	2:none		1
4	1:sunburned	1:sunburned		0.667
5	1:sunburned	1:sunburned		1
6	2:none	2:none		1
7	2:none	2:none		1
8	2:none	2:none		1
9	1:sunburned	1:sunburned		1
10	2:none	2:none		1
11	1:sunburned	1:sunburned		1
12	2:none	2:none		1
13	2:none	1:sunburned	+	0.667

ไม่เดลคาดการณ์ instance ที่เพิ่มเข้าไปพิจ

# เพิ่มตัวอย่างข้อมูลใหม่ 2 ตัว

## ซึ่งมีผลไม่ตรงกันกับตัวอย่างเดิม

	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none
13	blonde	average	light	no	none
14	blonde	average	light	no	none

# Confusion Matrix

ไม่เดลทำนาย

classified as =>

	sunburned	none
sunburned	3	2
none	0	9

ค่าของข้อมูลจริง

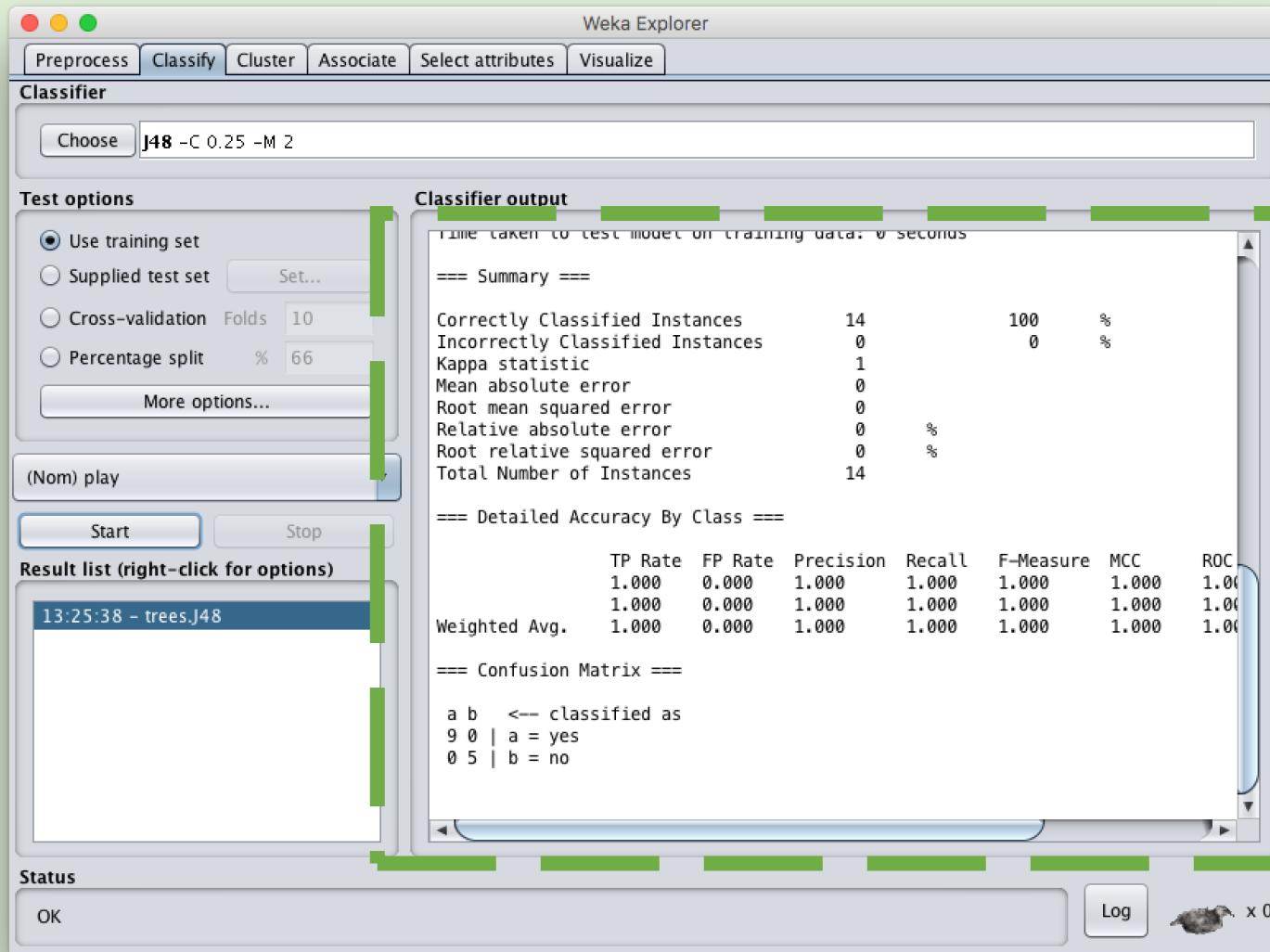
# តាមរយៈរបៀបទូទៅទី 2 : Weather

	outlook	temperature	humidity	windy	play
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

	Attributes				Class
	outlook	temperature	humidity	windy	play
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

สภาพอากาศ		ระดับอุณหภูมิ	ระดับความชื้น	มีลม หรือไม่	ออกป่า <sup>เล่นหรือไม่</sup>
	outlook	temperature	humidity	windy	play
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

# ผล Classification ของปัญหา Weather ด้วย Decision tree



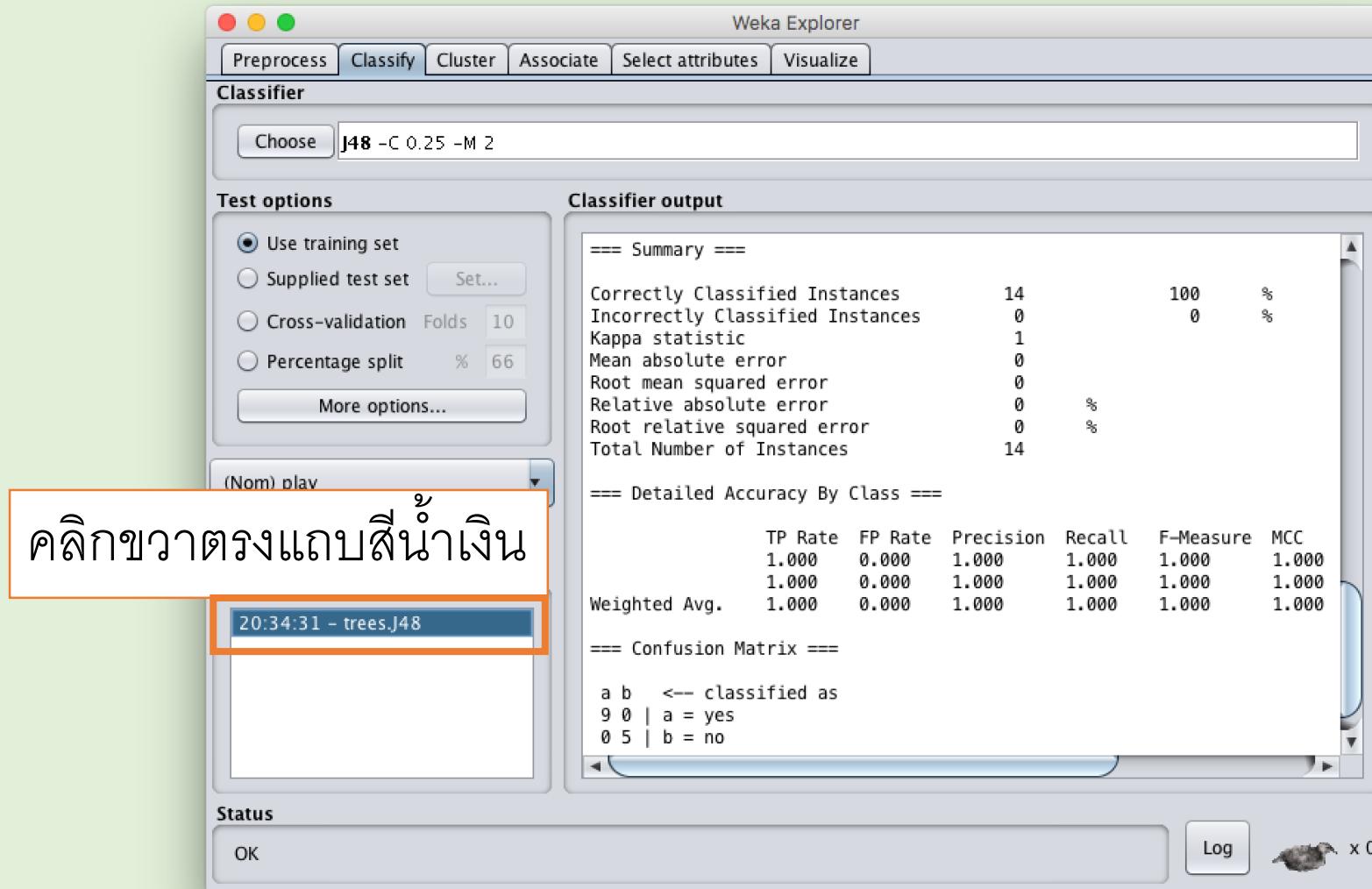
# Confusion Matrix

ค่าของข้อมูลจริง

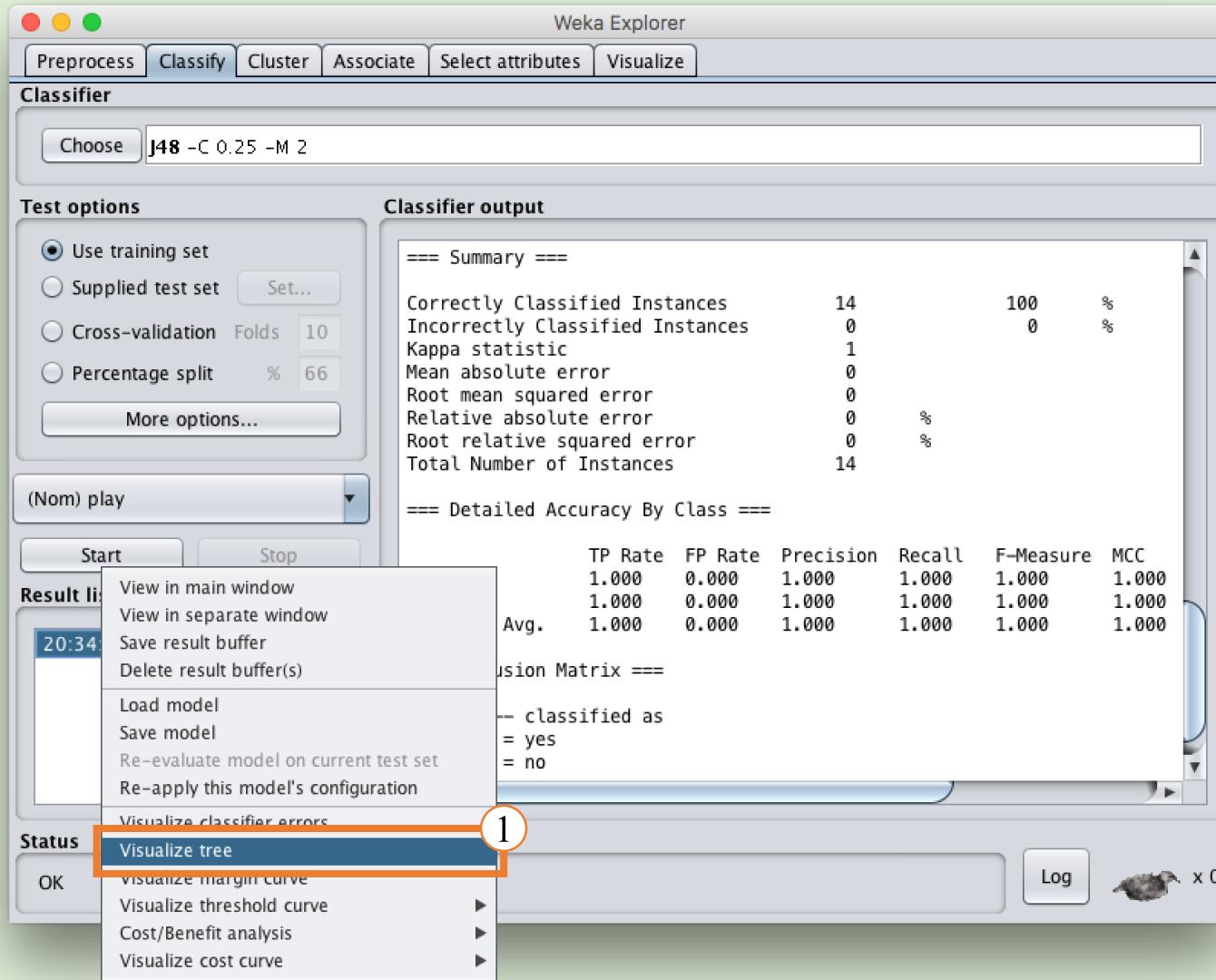
classified as =>

		ไม่เดลท์นาย	
		Yes	No
classified as =>	Yes	9	0
	No	0	5

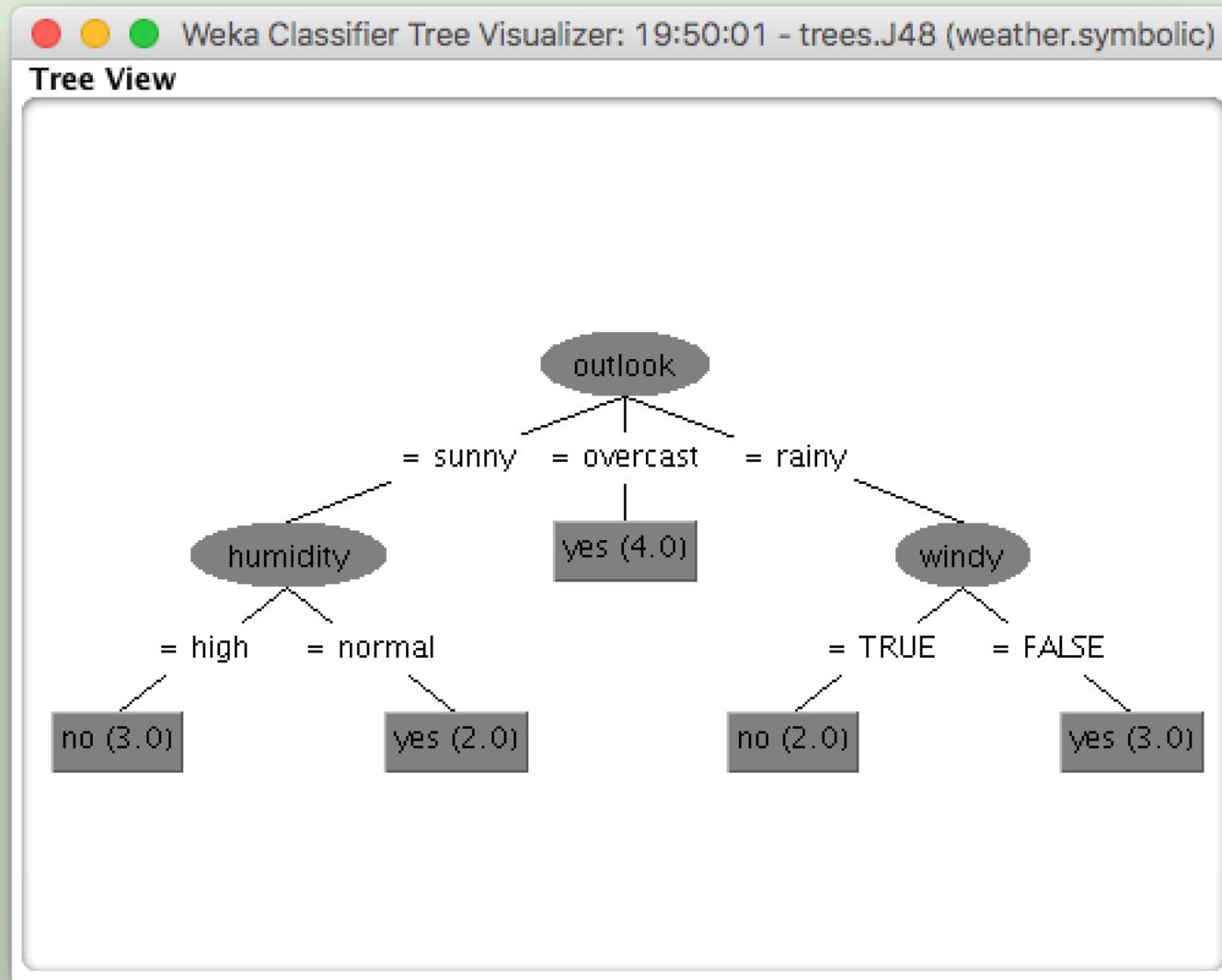
# การ Visualize โมเดล Decision tree



# การ Visualize ไม้เดล Decision tree



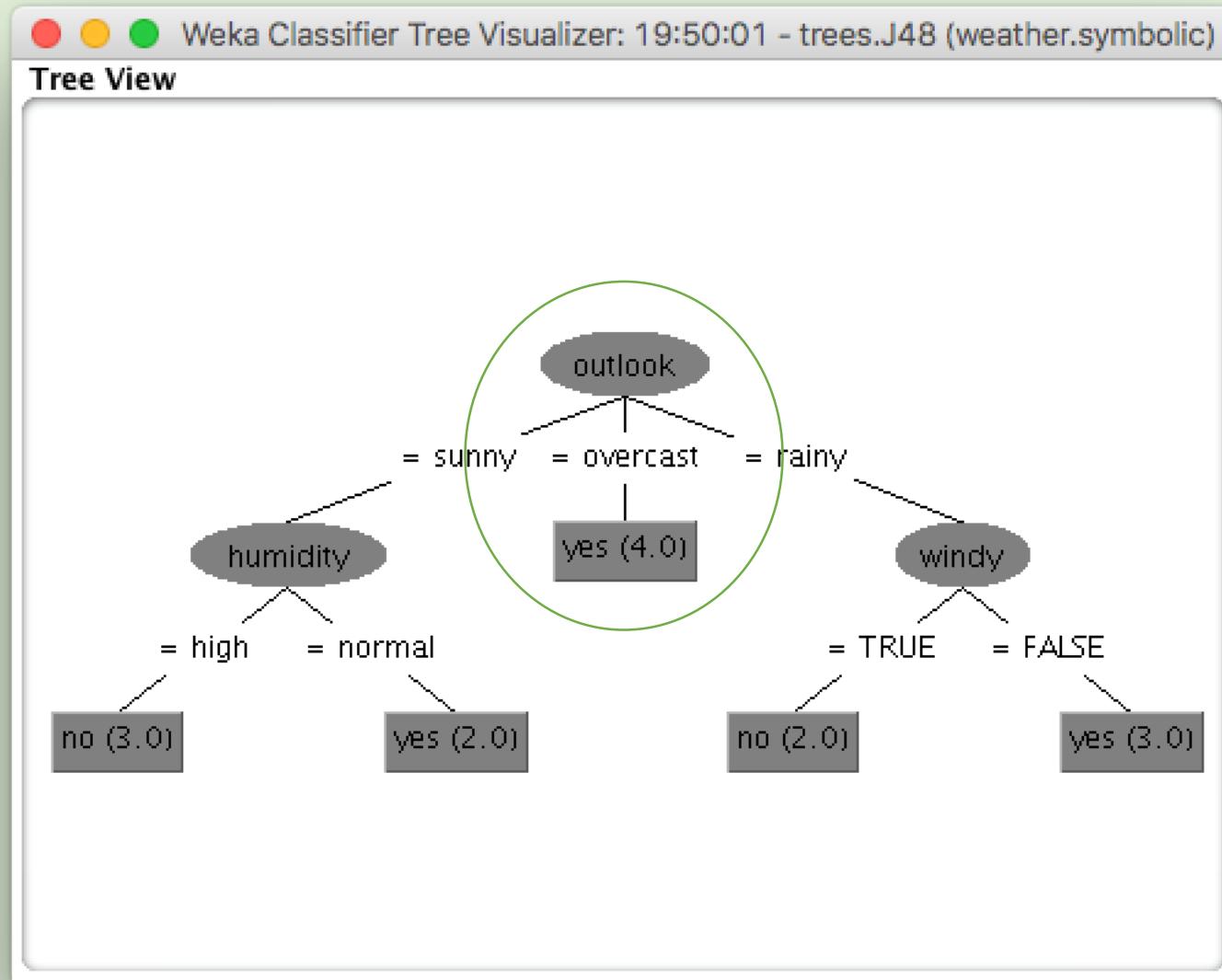
# ໂຄນເດລ Decision tree ຂອງປິ່ງຫາ Weather



# การวิเคราะห์ผลของปัญหา Weather ด้วย Decision tree

- ไม่เดลミความแม่นยำ 100%
- จากไม่เดล Decision tree
  - outlook เป็น attributes ที่มีอิทธิพลในการแบ่งกลุ่มข้อมูลออกจากกันมากที่สุด
  - สังเกตได้ว่าถ้า outlook เป็น sunny สามารถตอบได้ว่า Play จะเป็น yes ทั้งที่

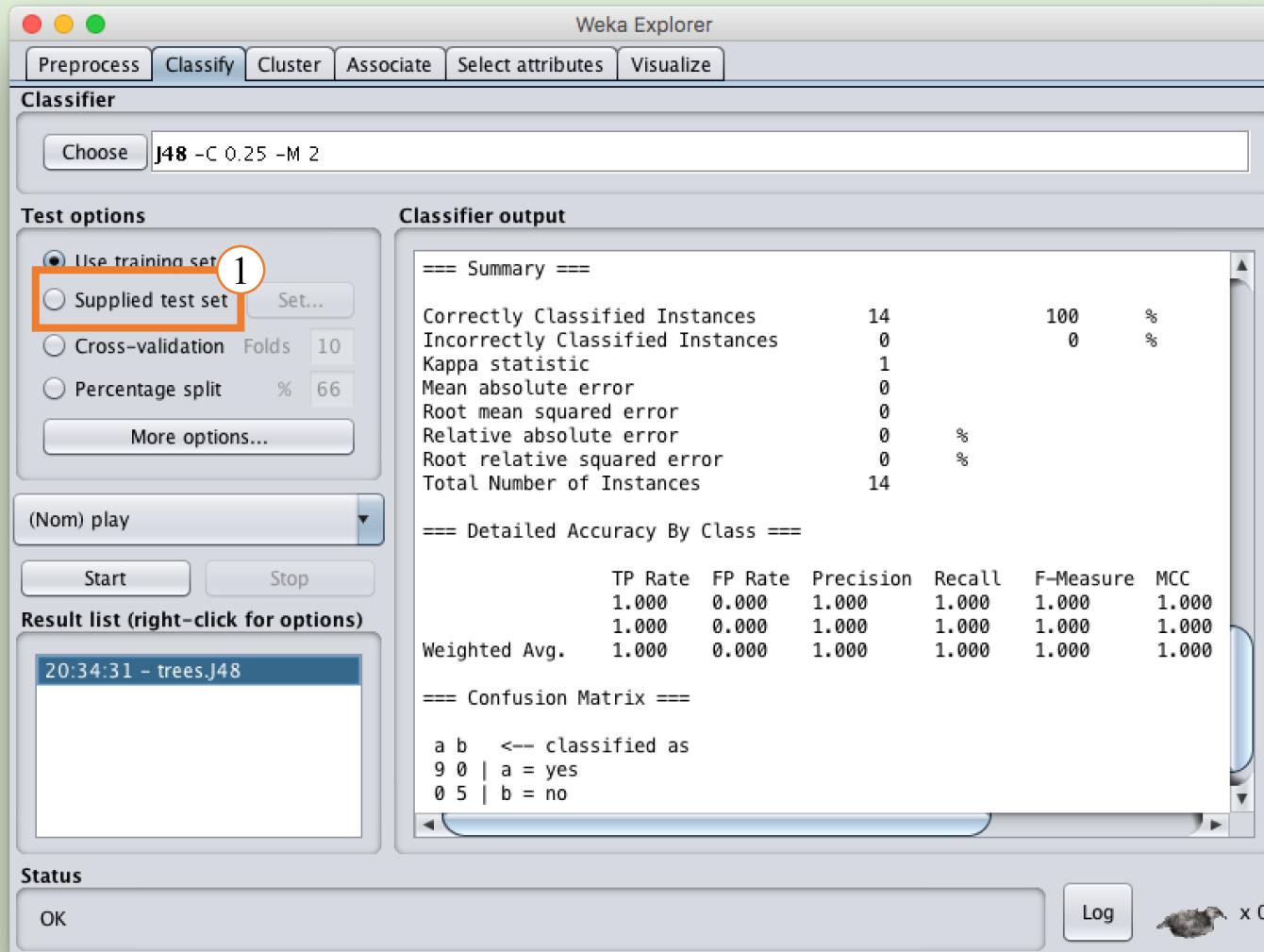
# ໂຄນເດລ Decision tree ຂອງປິ່ງຫາ Weather



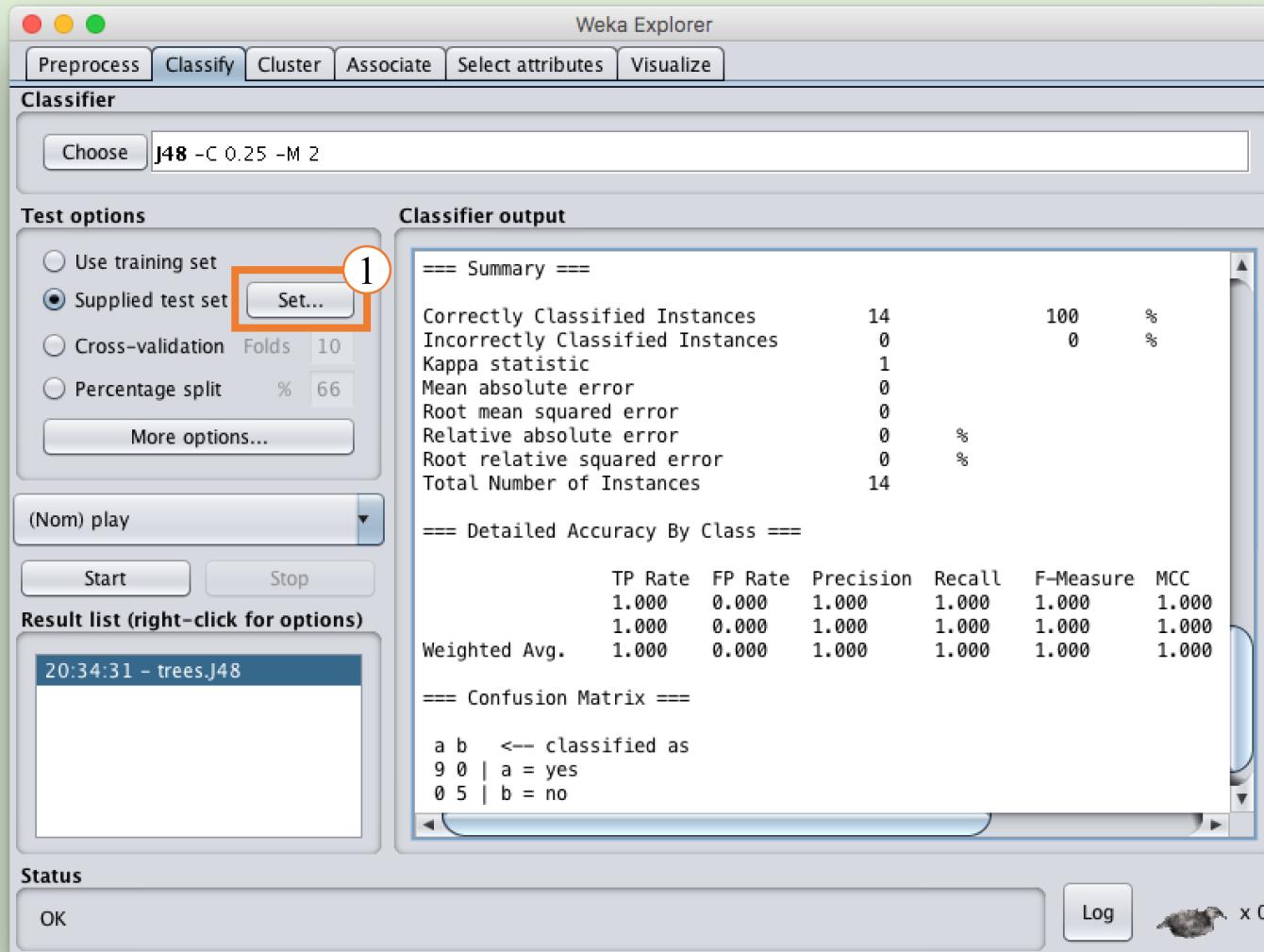
# การคาดการณ์ข้อมูลที่ไม่ทราบคำต่อไป

outlook	temperature	humidity	windy	play
sunny	hot	high	TRUE	?
overcast	mild	normal	FALSE	?
rainy	cool	high	FALSE	?

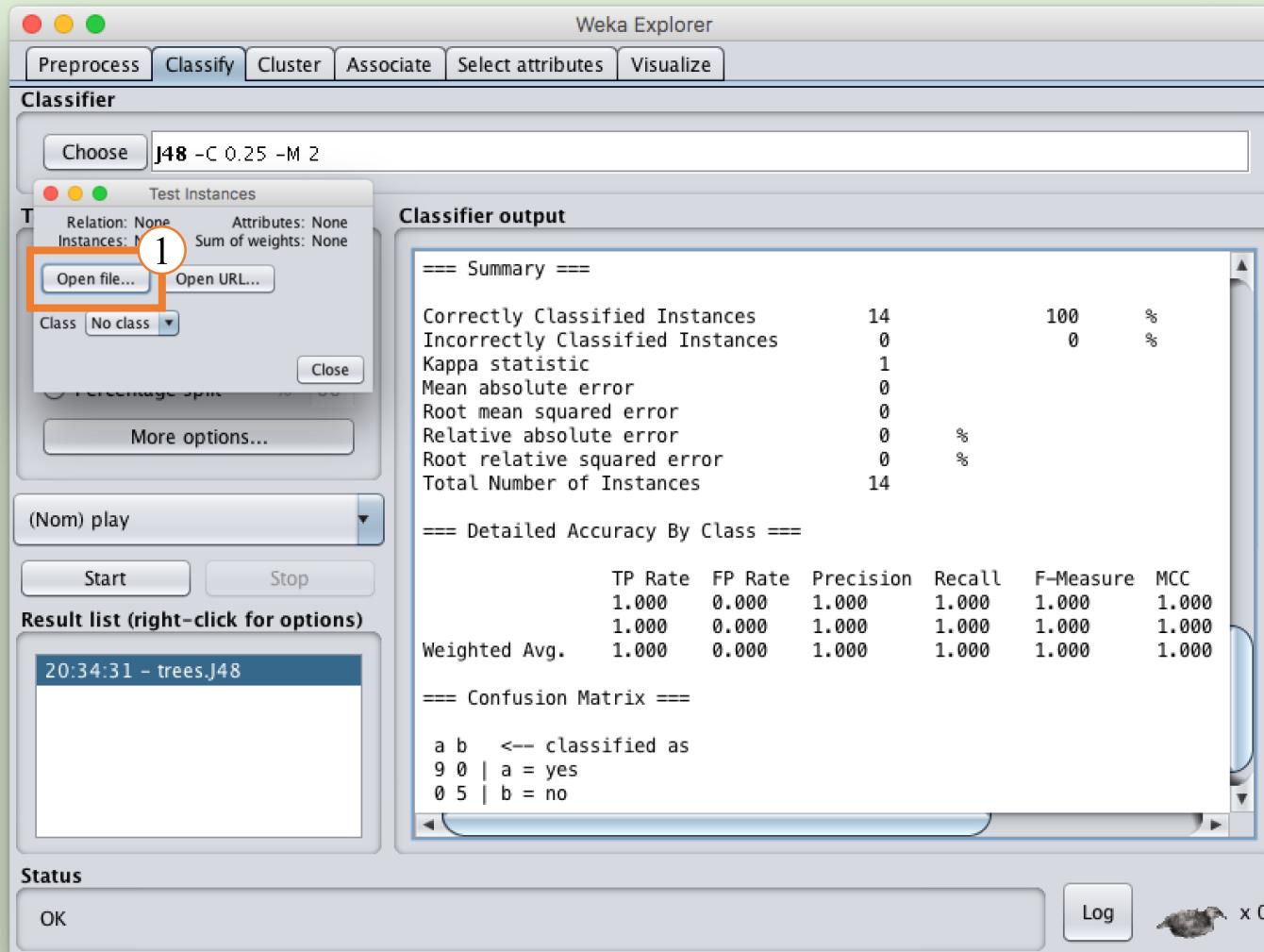
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



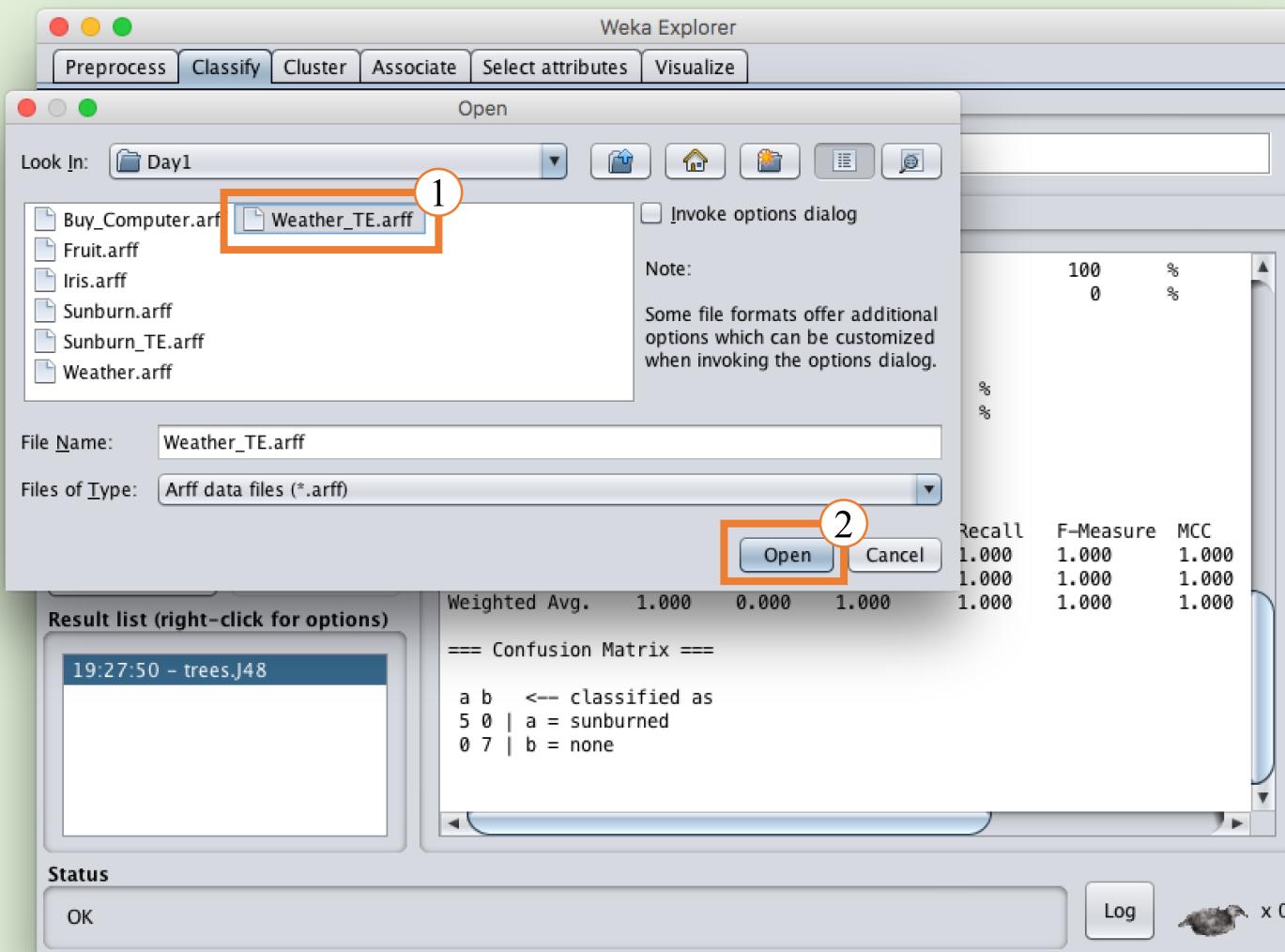
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



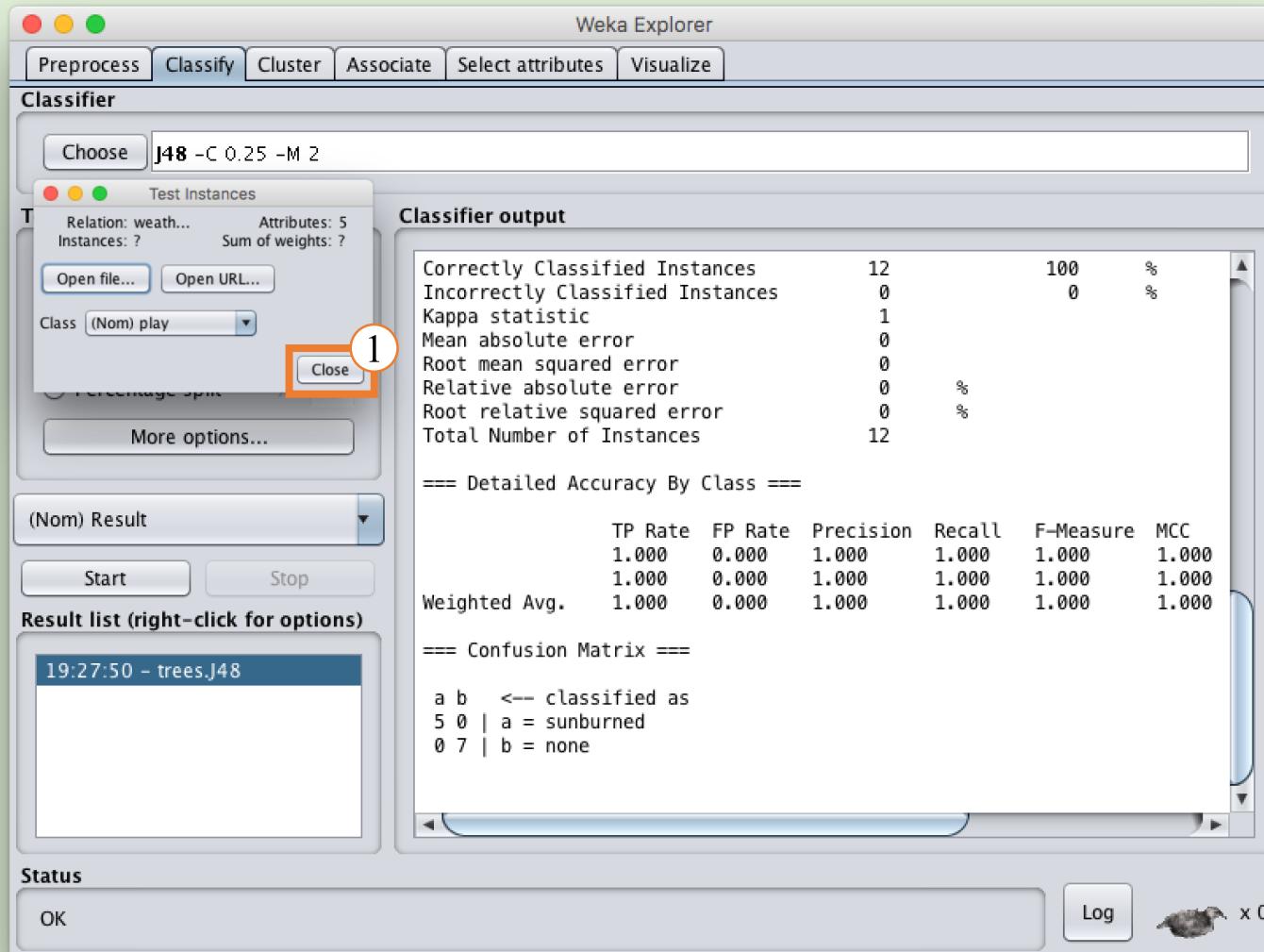
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



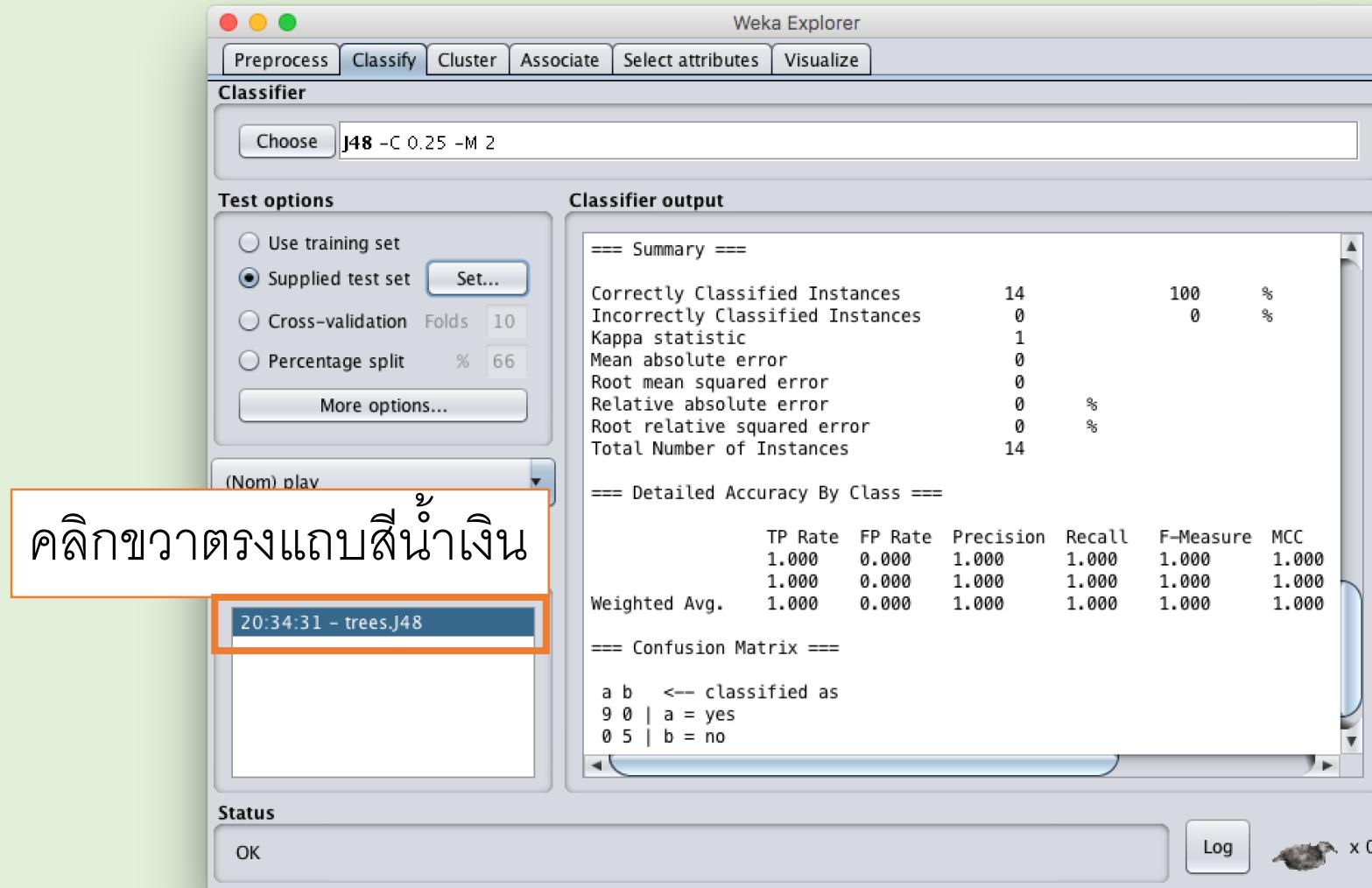
# การคัดกรณ์ข้อมูลที่ไม่ทราบคำศوب



# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب

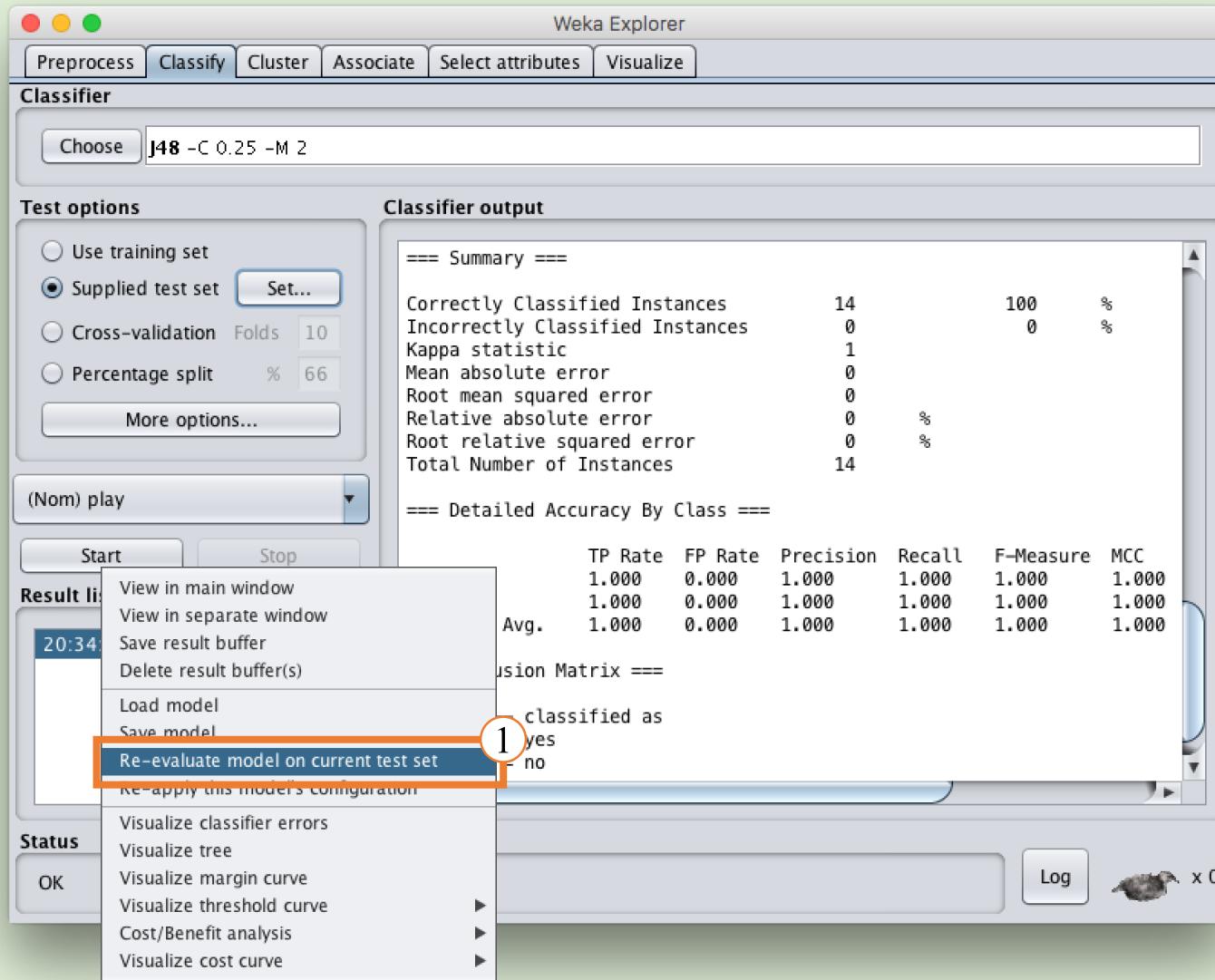


# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب

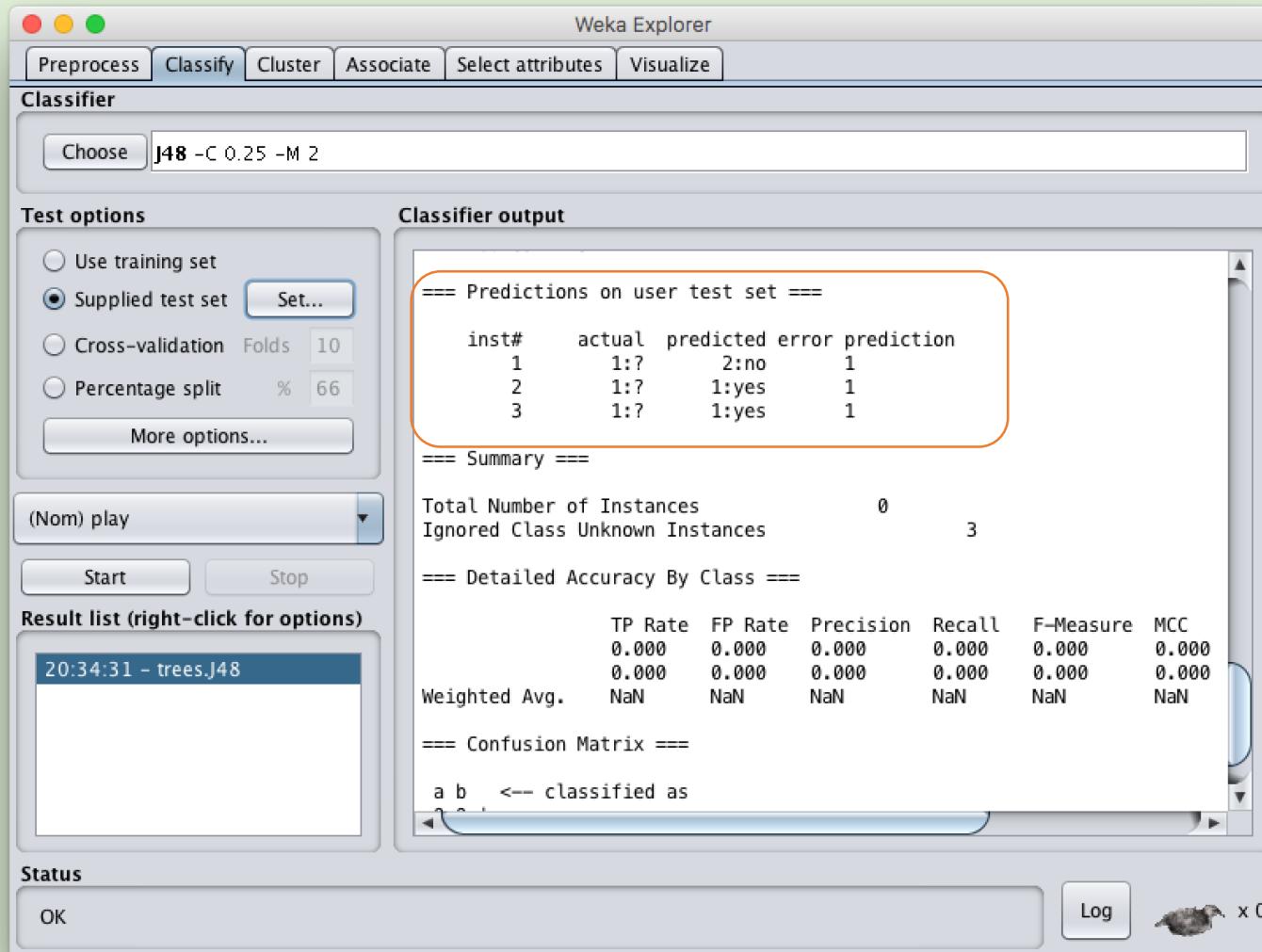


คลิกขวาตรงແບ່ນໍາເງິນ

# การคาดการณ์ข้อมูลที่ไม่ทราบคำตอบ



# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب

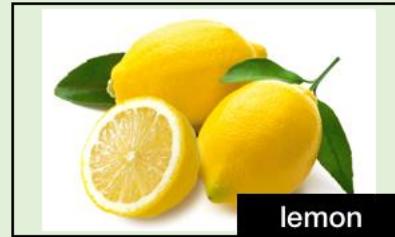


# ตัวอย่างปัญหาที่ 3 : Fruit

มีผลไม้อยู่ 4 ชนิด



mandarin



lemon



apple



orange

ตัวอย่างข้อมูลที่ใช้เคราะห์ชนิดผลไม้

	mass	width	height	color_score	fruit_name
1	192	8.4	7.3	0.55	apple
2	180	8	6.8	0.59	apple
	...	...	...	...	...
58	152	6.5	8.5	0.72	lemon
59	118	6.1	8.1	0.7	lemon

	Attributes				Class
	mass	width	height	color_score	fruit_name
1	<b>192</b>	8.4	7.3	0.55	apple
2	<b>180</b>	8	6.8	0.59	apple
3	<b>176</b>	7.4	7.2	0.6	apple
4	<b>178</b>	7.1	7.8	0.92	apple
5	<b>172</b>	7.4	7	0.89	apple
6	<b>166</b>	6.9	7.3	0.93	apple
	...	...	...	...	...
58	<b>152</b>	6.5	8.5	0.72	lemon
59	<b>118</b>	6.1	8.1	0.7	lemon

ประเภท

มวล

ความกว้าง

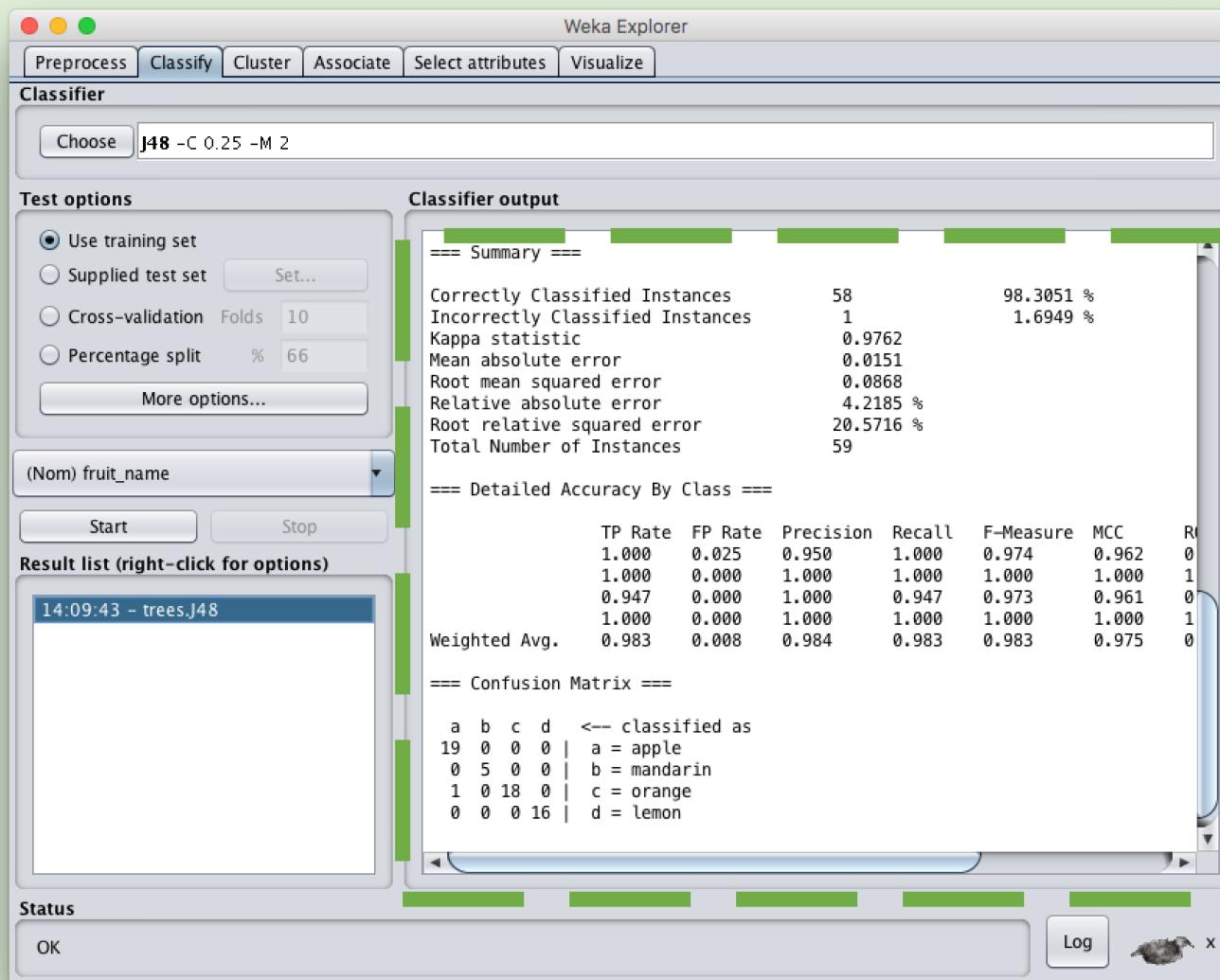
ความสูง

สี

ผลไม้

	mass	width	height	color_score	fruit_name
1	<b>192</b>	8.4	7.3	0.55	apple
2	<b>180</b>	8	6.8	0.59	apple
3	<b>176</b>	7.4	7.2	0.6	apple
4	<b>178</b>	7.1	7.8	0.92	apple
5	<b>172</b>	7.4	7	0.89	apple
6	<b>166</b>	6.9	7.3	0.93	apple
	...	...	...	...	...
58	<b>152</b>	6.5	8.5	0.72	lemon
59	<b>118</b>	6.1	8.1	0.7	lemon

# ผล Classification ของปัญหา Fruit ด้วย Decision tree



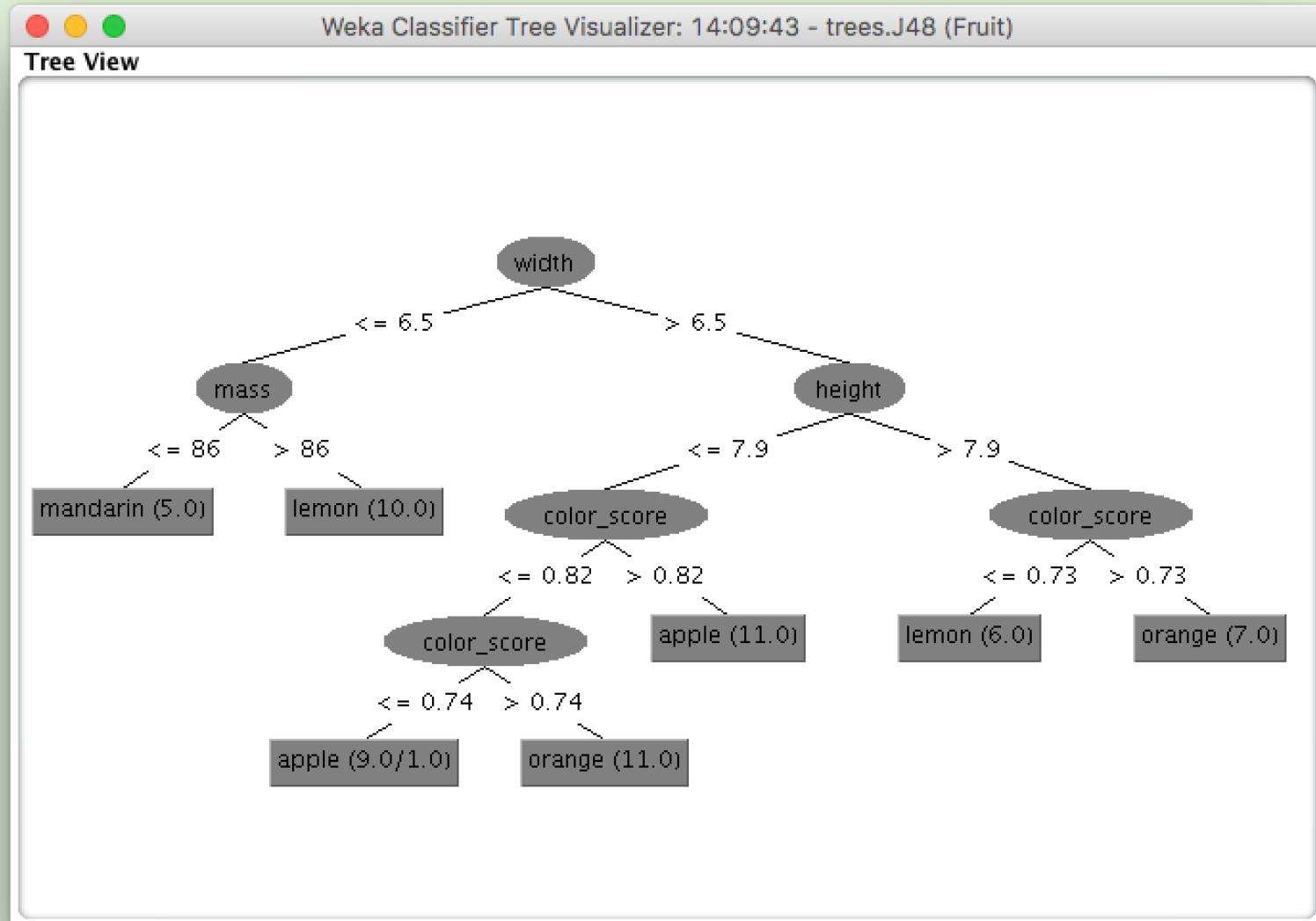
# การวิเคราะห์ผลของปัญหา Fruit ด้วย

## Decision tree

สรุปเป็น Confusion Matrix ได้ดังนี้

classified as =>	apple	mandarin	orange	lemon
apple	19	0	0	0
mandarin	0	5	0	0
orange	1	0	18	0
lemon	0	0	0	16

# ໂຄນເດລ Decision tree ຂອງປົມຫາ Fruit

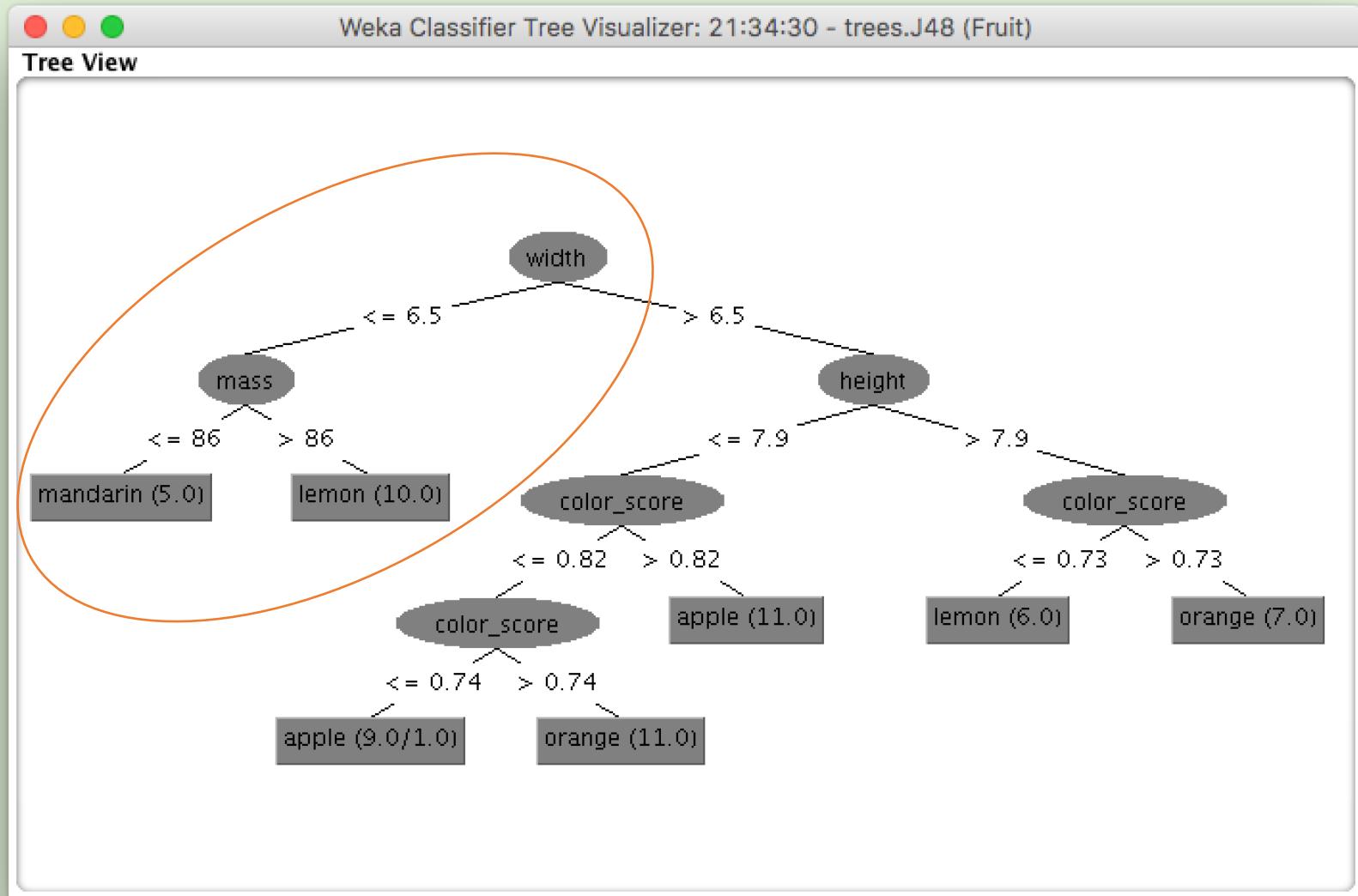


# การวิเคราะห์ผลของปัญหา Fruit ด้วย

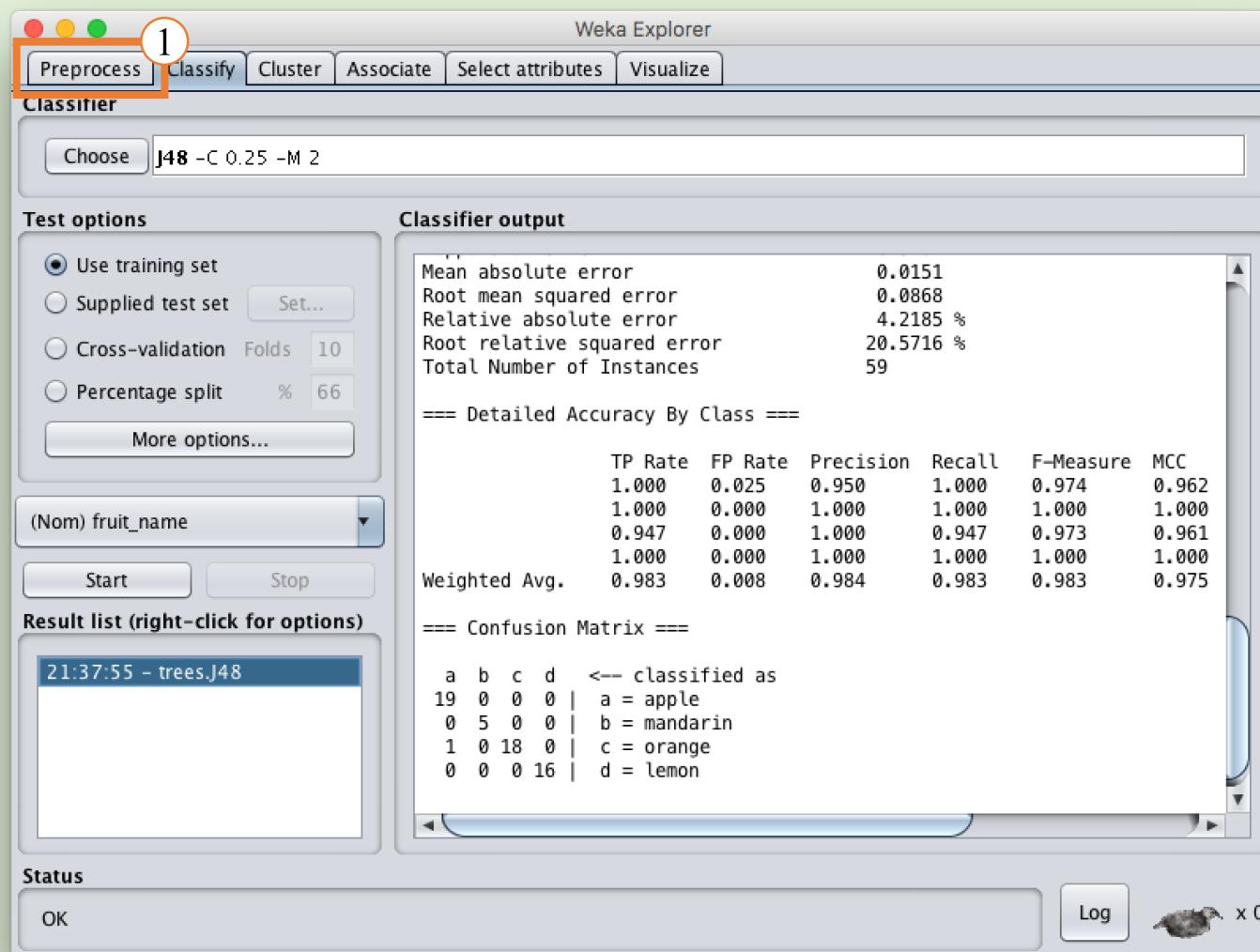
## Decision tree

- โมเดลมีความแม่นยำในอยู่ที่ 98.31%
- จากโมเดล Decision tree
  - ในการแบ่งกลุ่ม mandarin จะดูจาก attributes เพียง 2 ตัวคือ width และ mass
  - ส่วนการแบ่งกลุ่มผลไม้มีอินต์องจะดูจาก attributes หลายตัวประกอบกัน

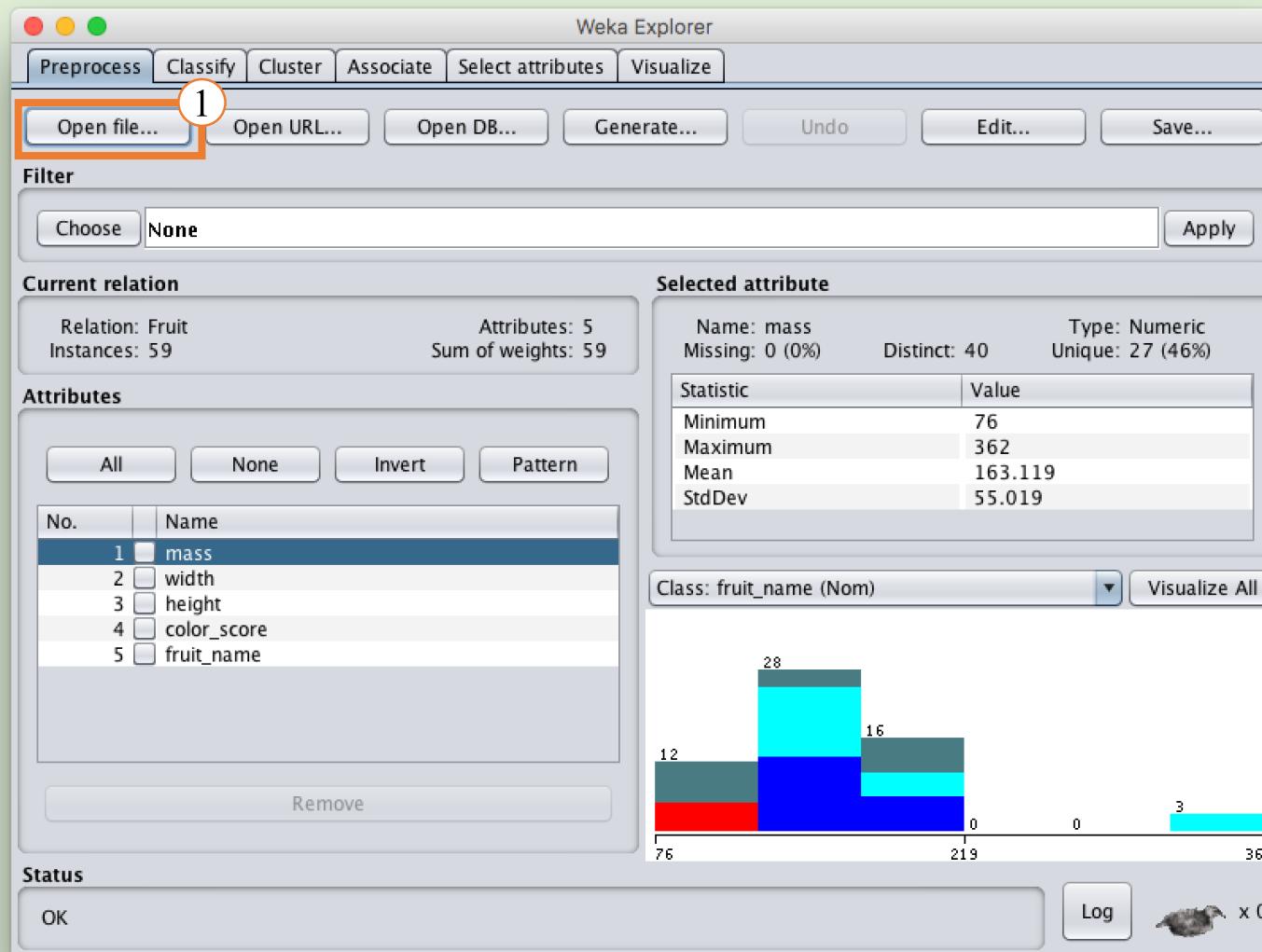
# ໂຄນເດລ Decision tree ຂອງປົມຫາ Fruit



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter 1 Choose None Apply

Current relation

Relation: Fruit Attributes: 5 Instances: 59 Sum of weights: 59

Selected attribute

Name: mass Type: Numeric  
Missing: 0 (0%) Distinct: 40 Unique: 27 (46%)

Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

Attributes

All None Invert Pattern

No.	Name
1	mass
2	width
3	height
4	color_score
5	fruit_name

Remove

Class: fruit\_name (Nom) Visualize All

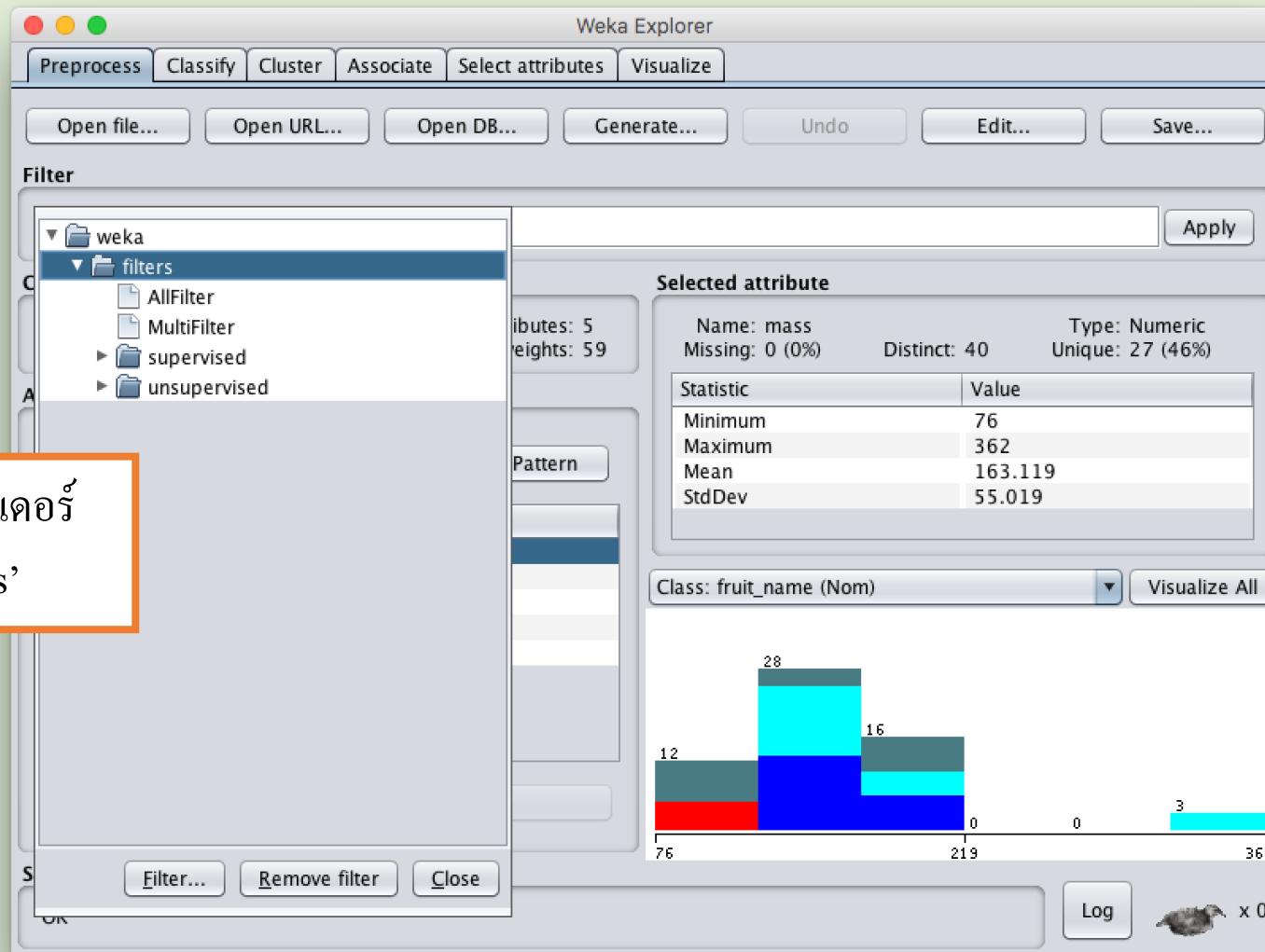
76 219 0 362

Status

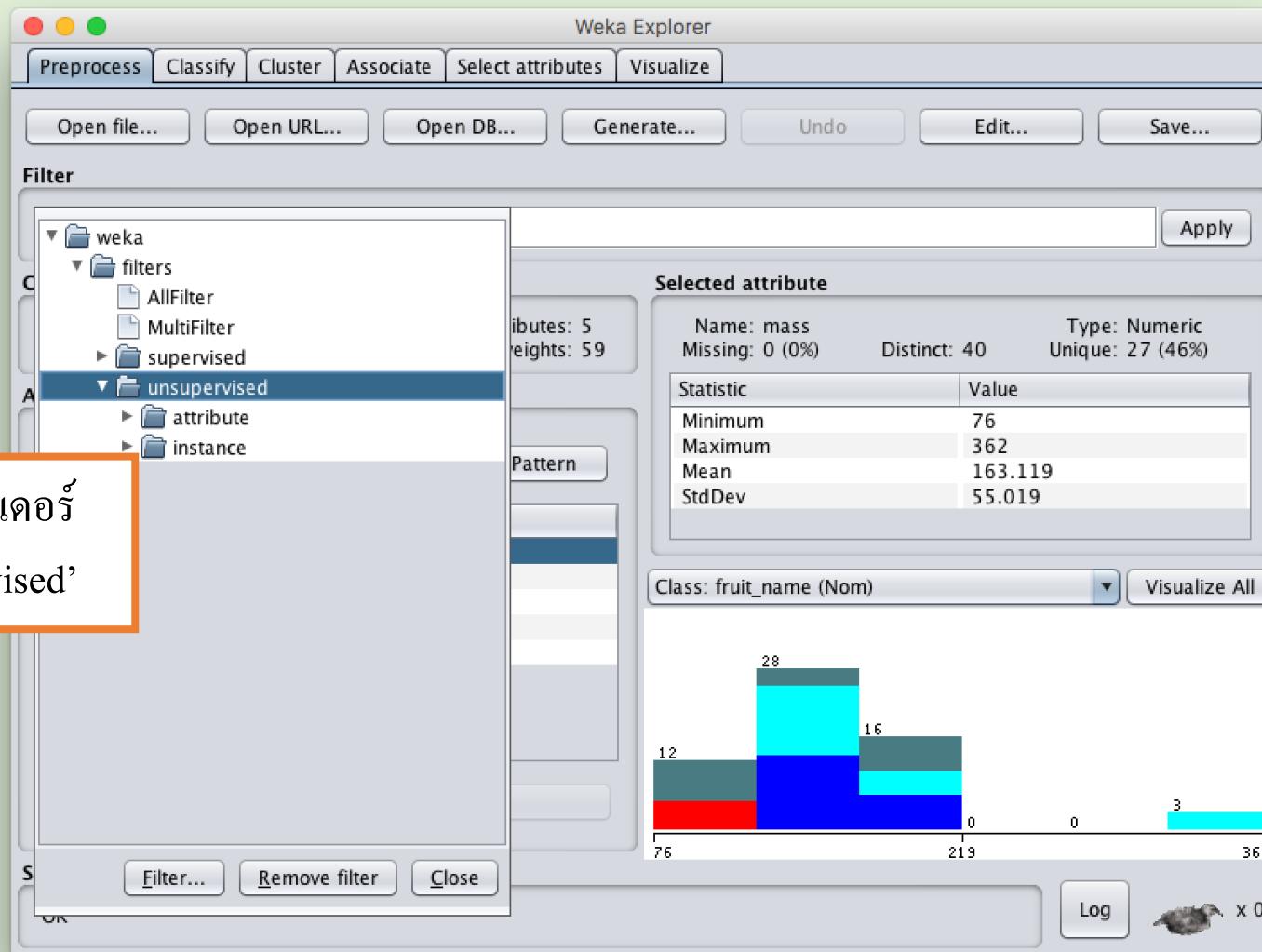
OK Log

x 0

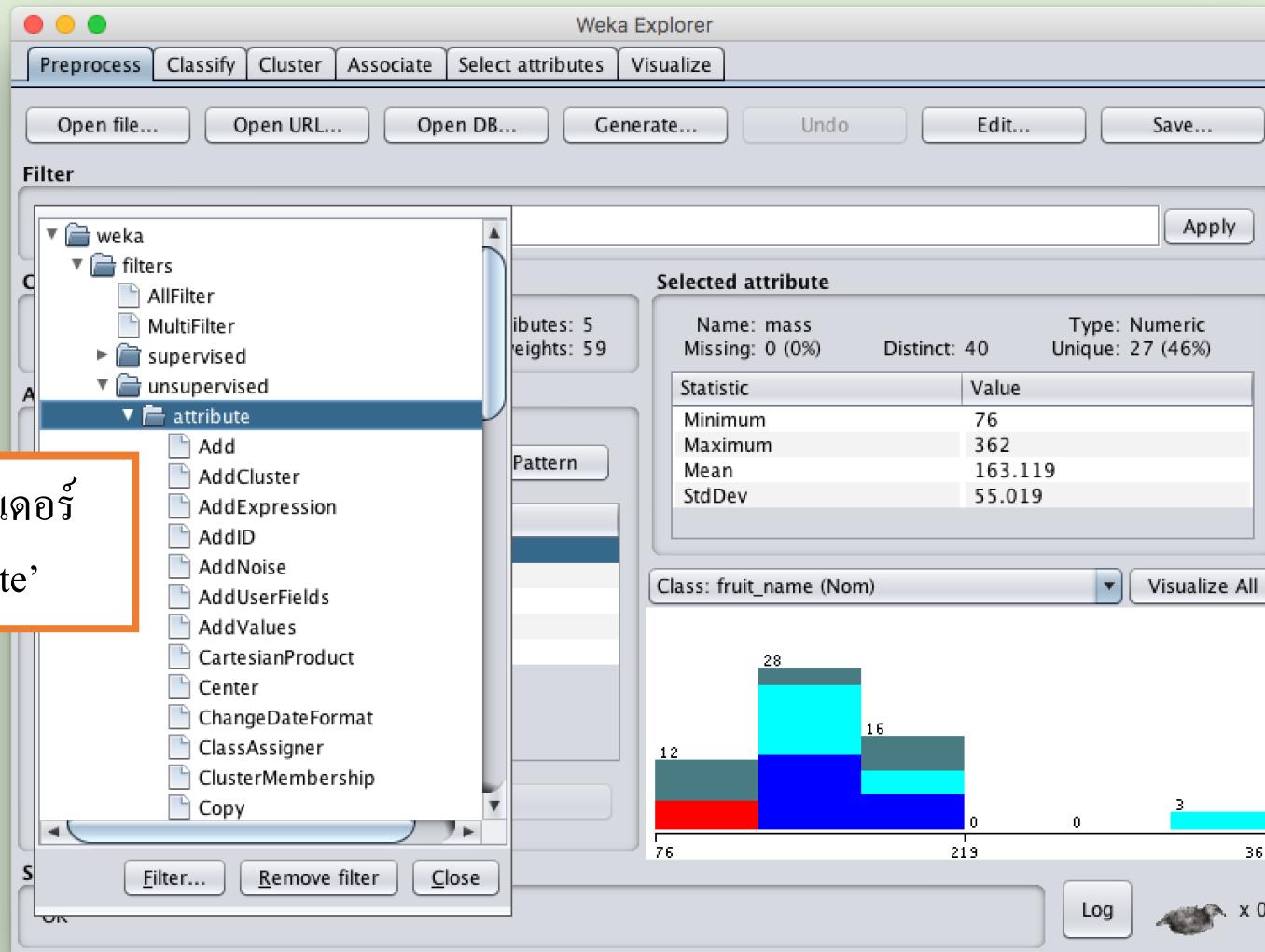
# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

เลือก 'Discretize'

The screenshot shows the Weka Explorer interface. In the center, the 'Selected attribute' panel displays information for the 'mass' attribute: Name: mass, Type: Numeric, Missing: 0 (0%), Distinct: 40, Unique: 27 (46%). Below this is a table of statistics:

Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **Discretize -B 10 -M -1.0 -R first-last** Apply 1

Current relation

Relation: Fruit Attributes: 5 Instances: 59 Sum of weights: 59

Attributes

All None Invert Pattern

No.	Name
1	mass
2	width
3	height
4	color_score
5	fruit_name

Remove

Selected attribute

Name: mass Type: Numeric  
Missing: 0 (0%) Distinct: 40 Unique: 27 (46%)

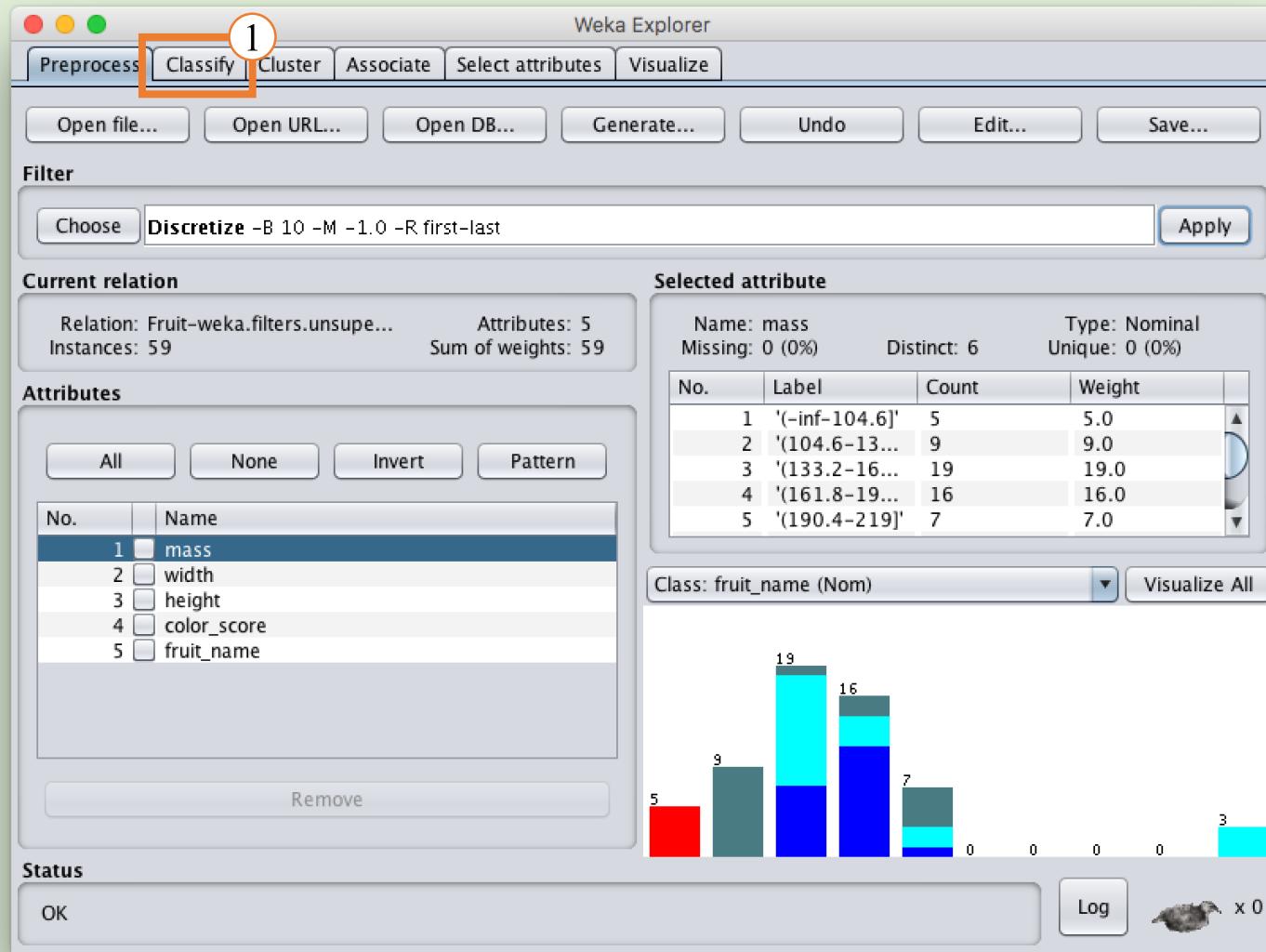
Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

Class: fruit\_name (Nom) Visualize All

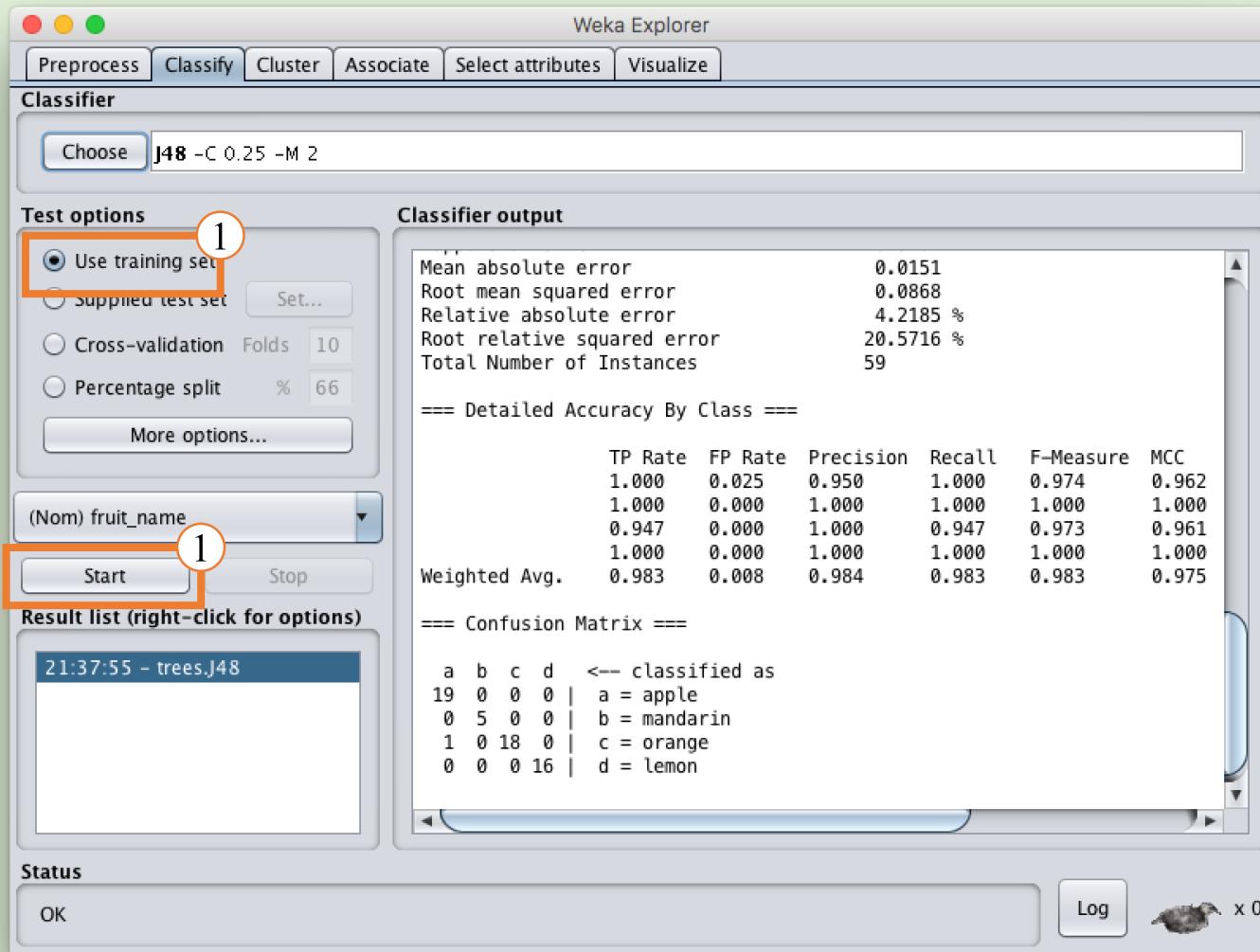
Status

OK Log

# สร้าง Decision tree ใหม่



# สร้าง Decision tree ใหม่



# ผล Classification ของ Decision tree

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

**Test options**

Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66  
More options...

(Nom) fruit\_name

Start Stop

**Result list (right-click for options)**

21:37:55 - trees.J48  
21:43:00 - trees.J48

**Classifier output**

==== Summary ===

Correctly Classified Instances	53	89.8305 %
Incorrectly Classified Instances	6	10.1695 %
Kappa statistic	0.8575	
Mean absolute error	0.065	
Root mean squared error	0.1802	
Relative absolute error	18.1921 %	
Root relative squared error	42.7201 %	
Total Number of Instances	59	

==== Detailed Accuracy By Class ===

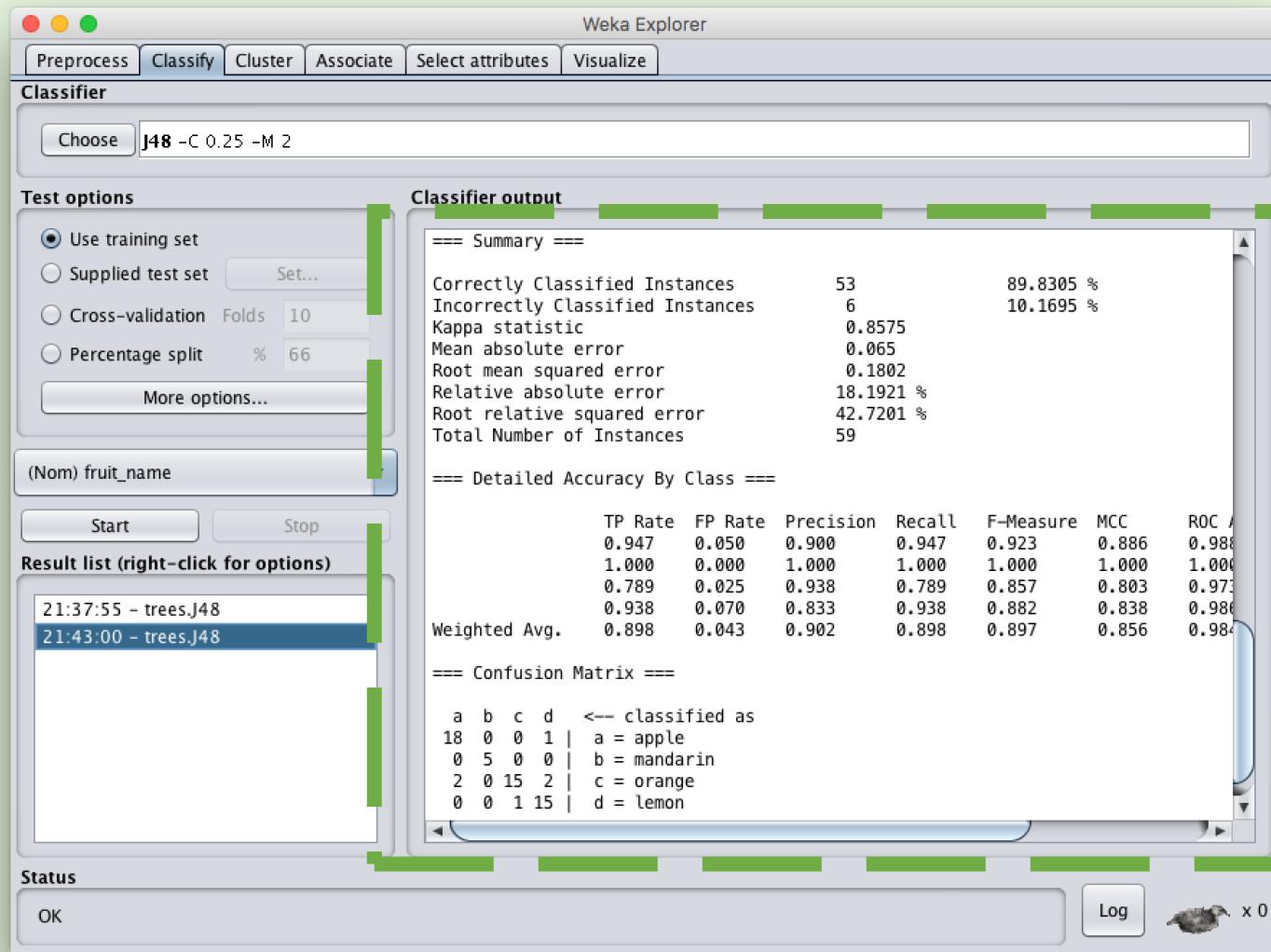
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC A
a = apple	0.947	0.050	0.900	0.947	0.923	0.886	0.988
b = mandarin	1.000	0.000	1.000	1.000	1.000	1.000	1.000
c = orange	0.789	0.025	0.938	0.789	0.857	0.803	0.973
d = lemon	0.938	0.070	0.833	0.938	0.882	0.838	0.986
Weighted Avg.	0.898	0.043	0.902	0.898	0.897	0.856	0.984

==== Confusion Matrix ===

				<-- classified as
a	b	c	d	
18	0	0	1	a = apple
0	5	0	0	b = mandarin
2	0	15	2	c = orange
0	0	1	15	d = lemon

Status

OK Log x 0



# การวิเคราะห์ผลของปัญหา Fruit ด้วย Decision tree

สรุปเป็น Confusion Matrix ได้ดังนี้

classified as =>	apple	mandarin	orange	lemon
apple	18	0	0	1
mandarin	0	5	0	0
orange	2	0	15	2
lemon	0	0	1	15

# ตัวอย่างปัญหาอิน

- Buying a computer
  - 14 Instances
  - 4 Attributes: Age, Income, Student, Credit\_Rating
  - 2 Classes: Yes (9 Instances), No (5 Instances)

No.	1: Age Nominal	2: Income Nominal	3: Student Nominal	4: Credit_Rating Nominal	5: Buy_Computer Nominal
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_age	medium	no	excellent	yes
13	middle_age	high	yes	fair	yes
14	senior	medium	no	excellent	no

# ตัวอย่างปัญหาอิน

- Iris
  - 150 Instances
  - 4 Attributes: sepal length, sepal width, petal length, petal width
  - 3 Classes: setosa, versicolor, virginica

No.	1: sepal length Numeric	2: sepal width Numeric	3: petal length Numeric	4: petal width Numeric	5: class Nominal
1	4.3	3.0	1.1	0.1	Iris-setosa
2	4.8	3.0	1.4	0.1	Iris-setosa
3	4.9	3.1	1.5	0.1	Iris-setosa
4	4.9	3.1	1.5	0.1	Iris-setosa
5	4.9	3.1	1.5	0.1	Iris-setosa

# Naive Bayes

$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

តំវេយោងបែមុហានឹងប្រជុកទៅខ្លួន

## Naive Bayes

1. Sunburn
2. Weather
3. Fruit

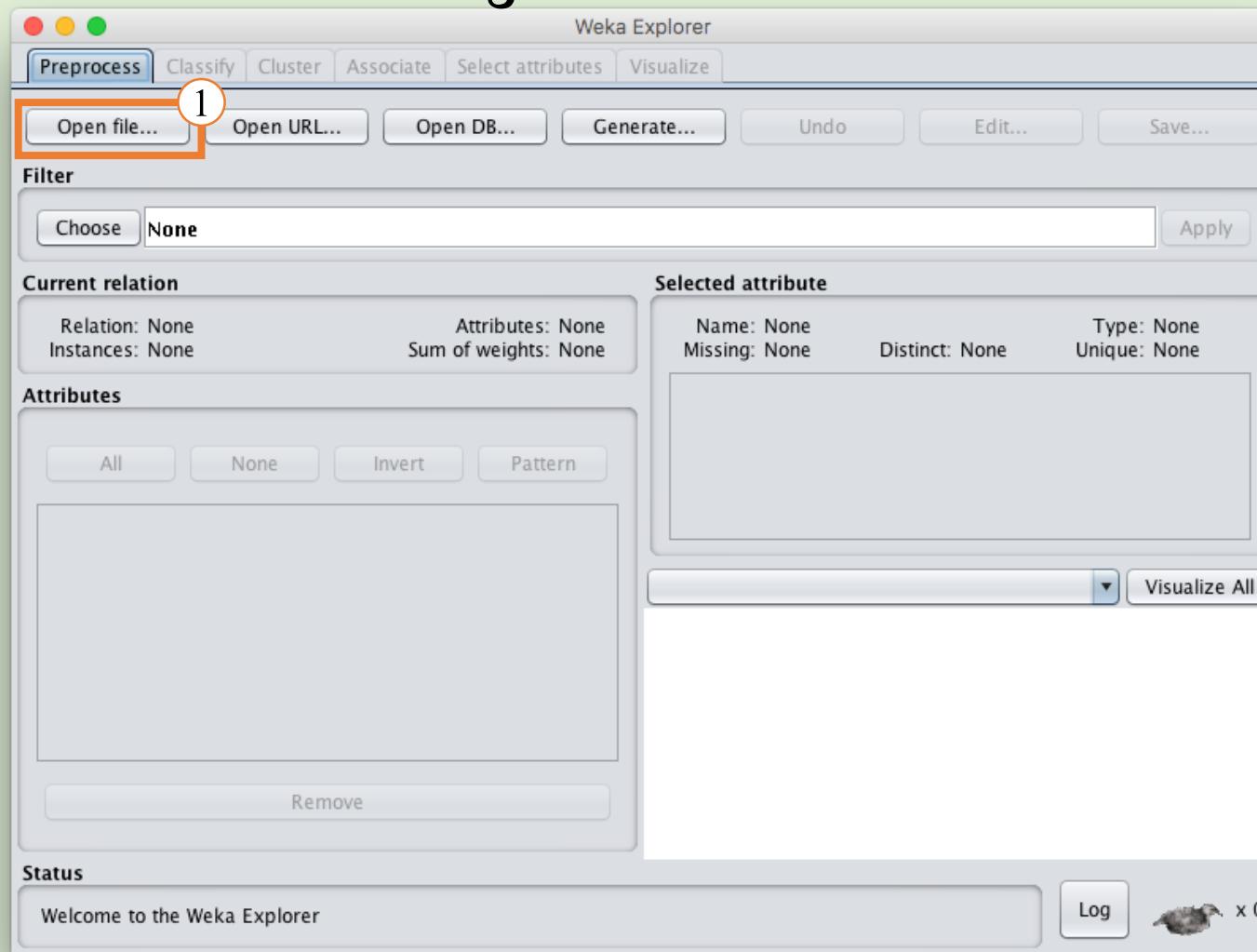
# ตัวอย่างปัญหาที่ 1 : Sunburn

	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
3	brown	short	average	yes	none
4	blonde	short	average	no	sunburned
5	red	average	heavy	no	sunburned
6	brown	tall	heavy	no	none
7	brown	average	heavy	no	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none

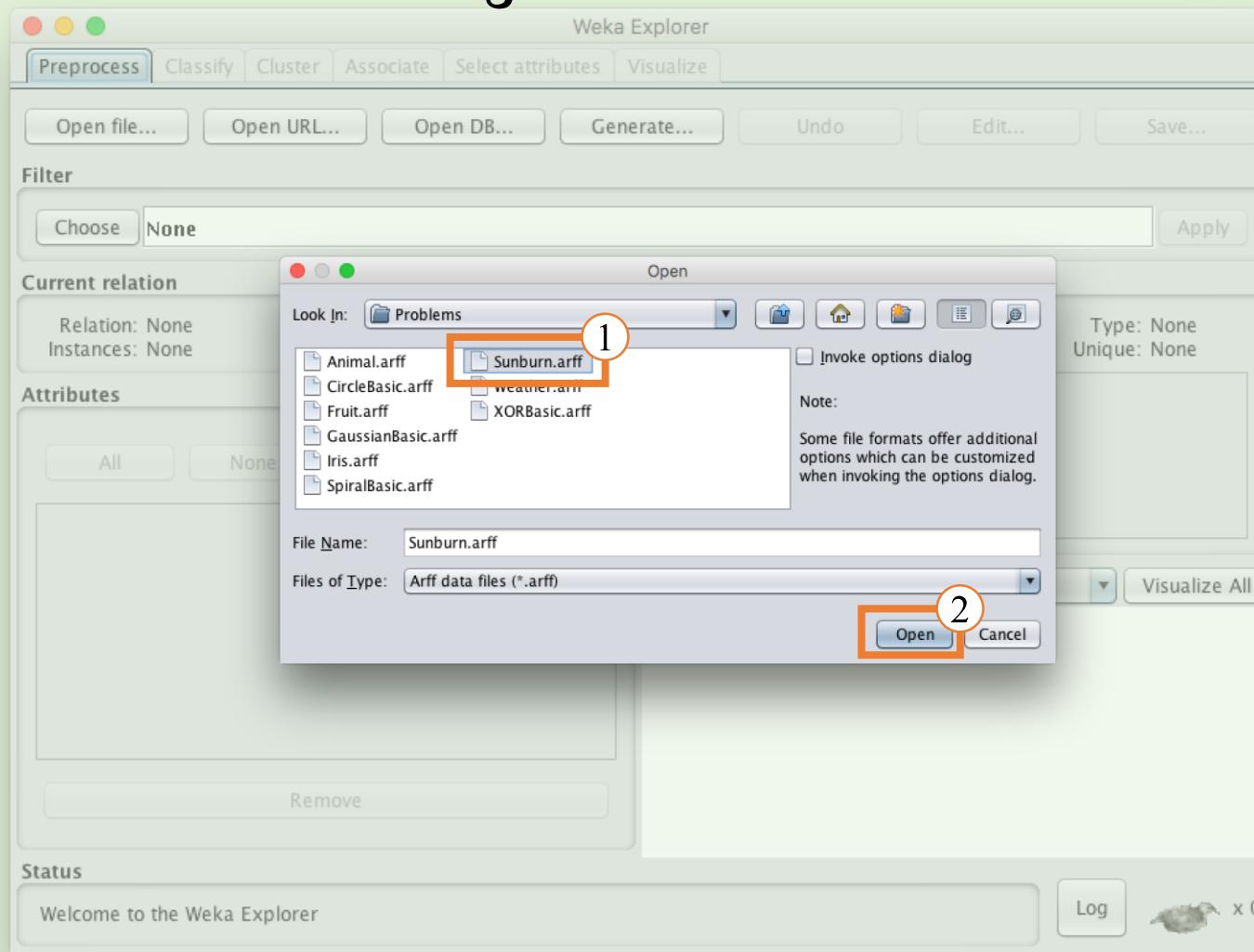
	Attributes				Class
	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
3	brown	short	average	yes	none
4	blonde	short	average	no	sunburned
5	red	average	heavy	no	sunburned
6	brown	tall	heavy	no	none
7	brown	average	heavy	no	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none

	ສືບມ	ສ່ວນສູງ	ນ້ຳຫັນກ	ທາໄລອຸ່ນ	ຜົວໄໝ້
	Hair	Height	Weight	Lotion	Result
1	blonde	average	light	no	sunburned
2	blonde	tall	average	yes	none
3	brown	short	average	yes	none
4	blonde	short	average	no	sunburned
5	red	average	heavy	no	sunburned
6	brown	tall	heavy	no	none
7	brown	average	heavy	no	none
8	blonde	short	light	yes	none
9	red	short	light	yes	sunburned
10	blonde	short	heavy	yes	none
11	red	tall	average	no	sunburned
12	brown	tall	light	yes	none

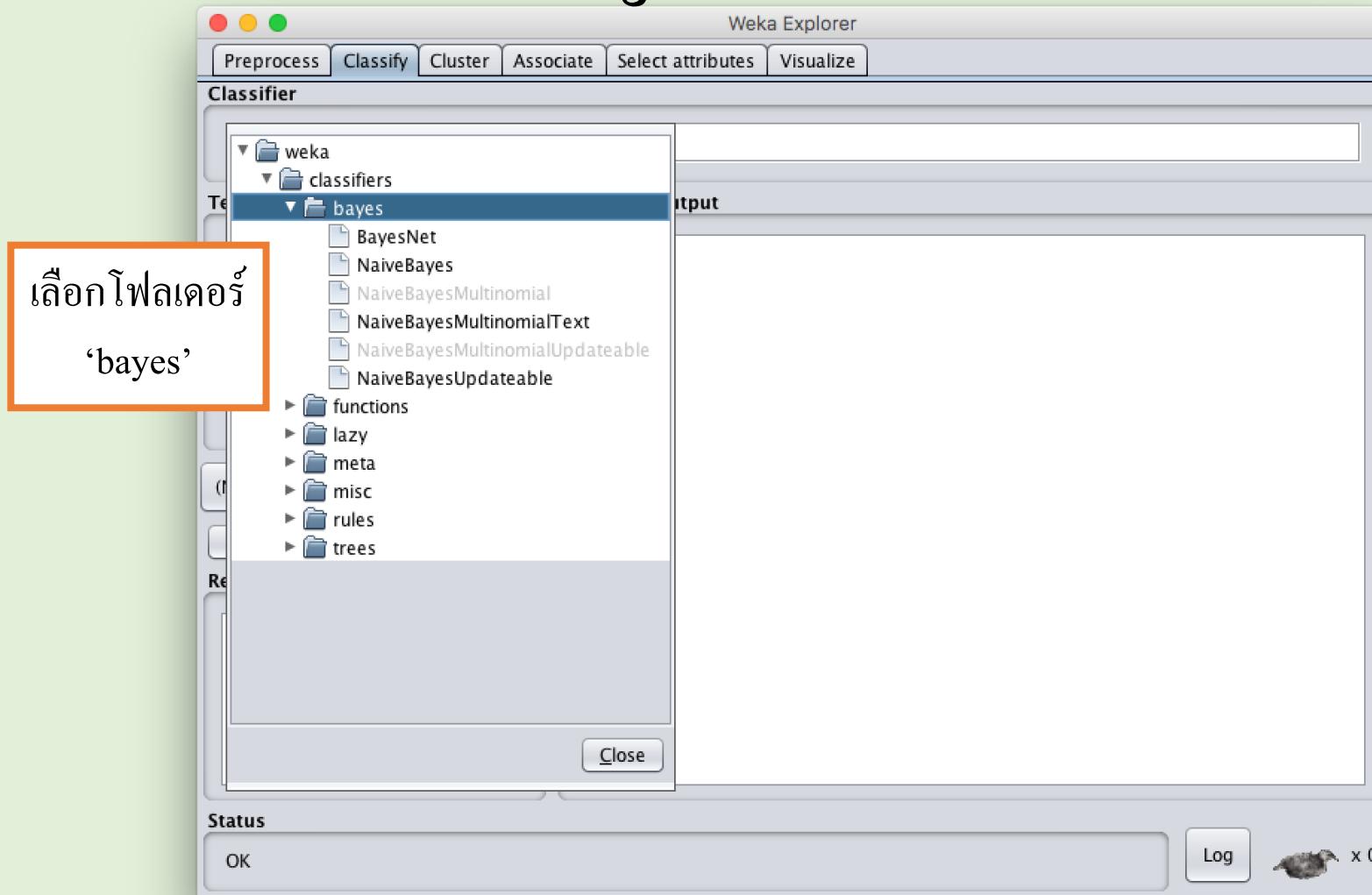
# ขั้นตอนการสร้าง Naive Bayes สำหรับปัญหา Sunburn



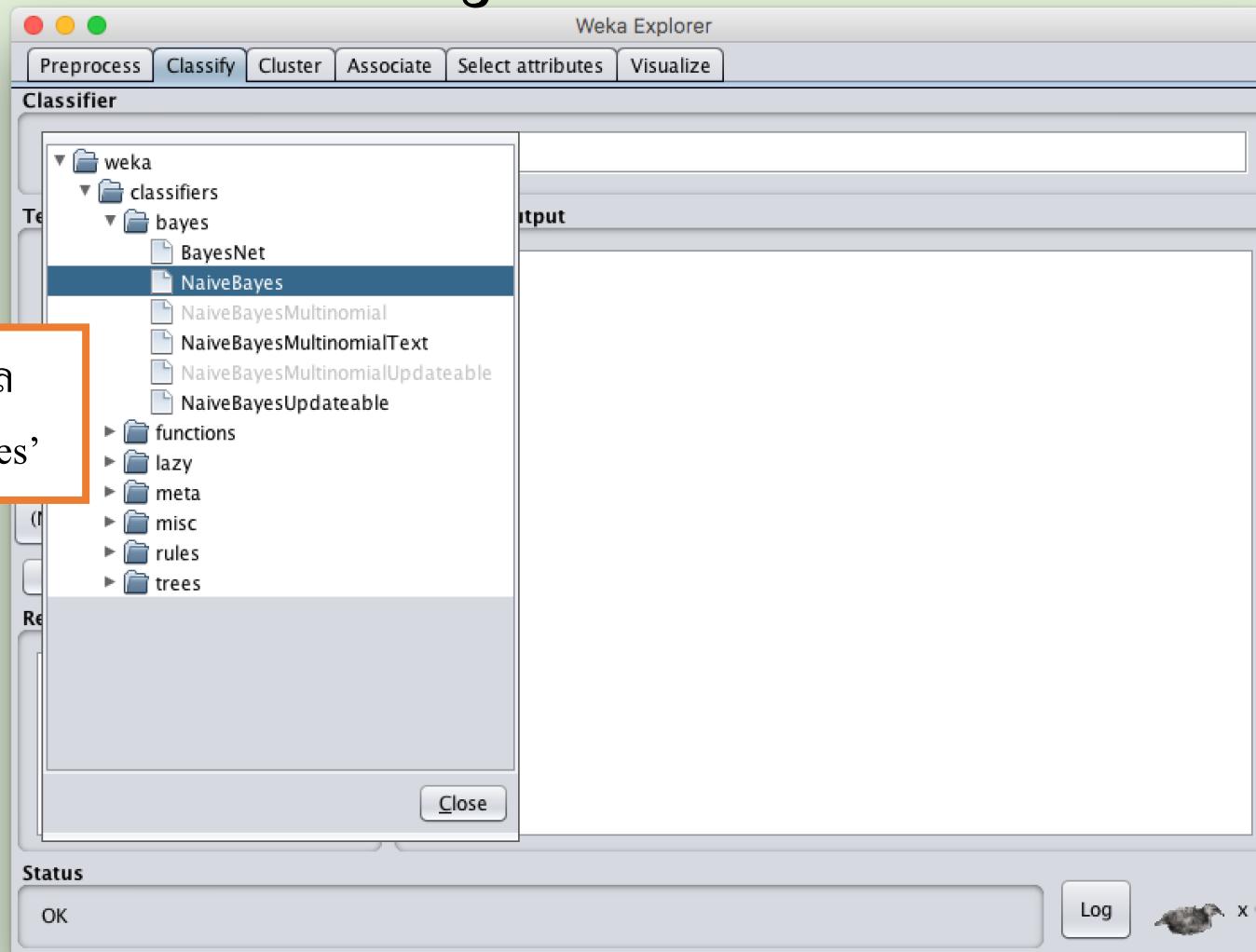
# ขั้นตอนการสร้าง Naive Bayes สำหรับปัญหา Sunburn



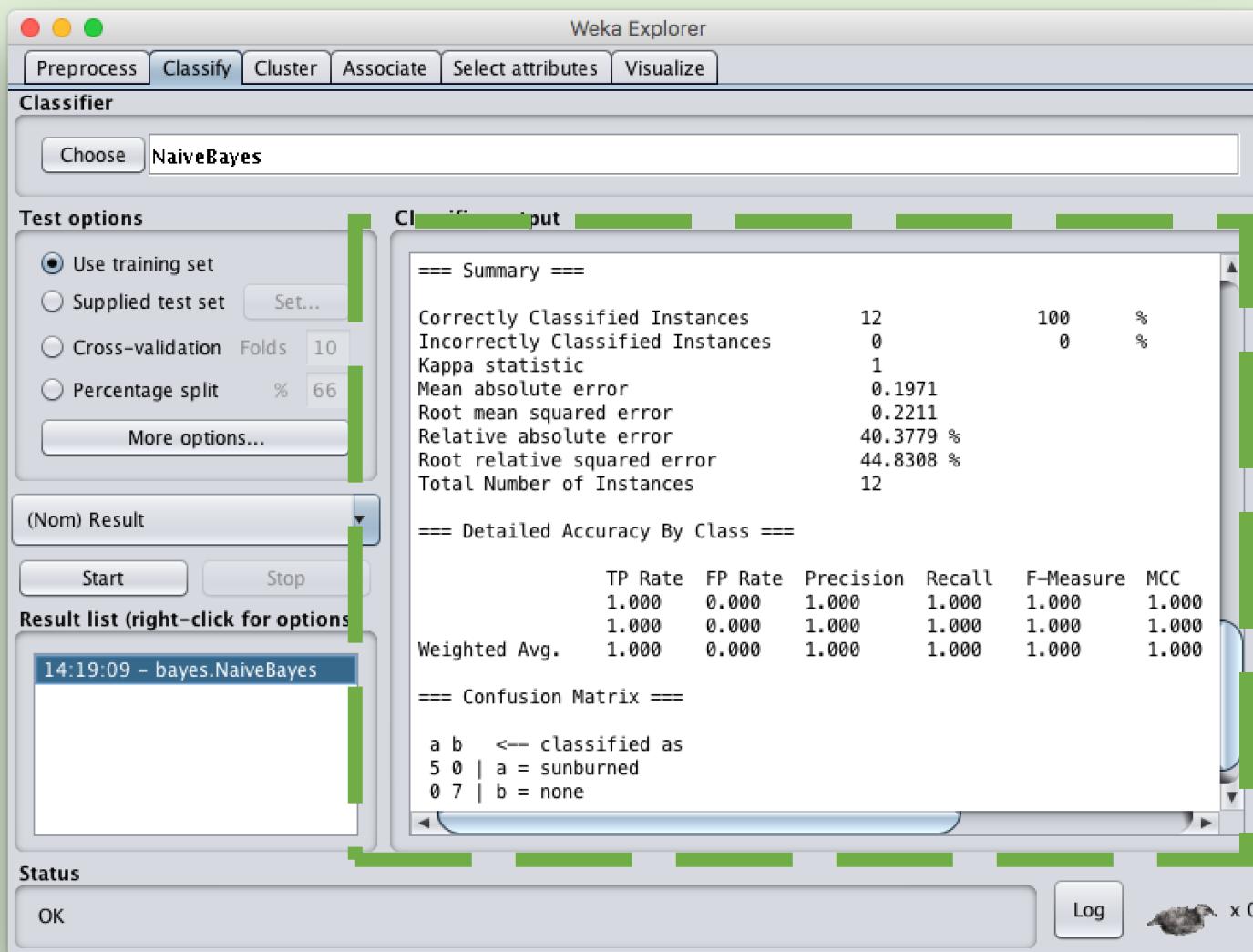
# ขั้นตอนการสร้าง Naive Bayes สำหรับปัญหา Sunburn



# ขั้นตอนการสร้าง Naive Bayes สำหรับปัญหา Sunburn



# ผล Classification ของปัญหา Sunburn ด้วย Naive Bayes



# การวิเคราะห์ผลของปัญหา Sunburn ด้วย Naive Bayes

สรุปเป็น Confusion Matrix ได้ดังนี้

ไม่เดลทำนาย

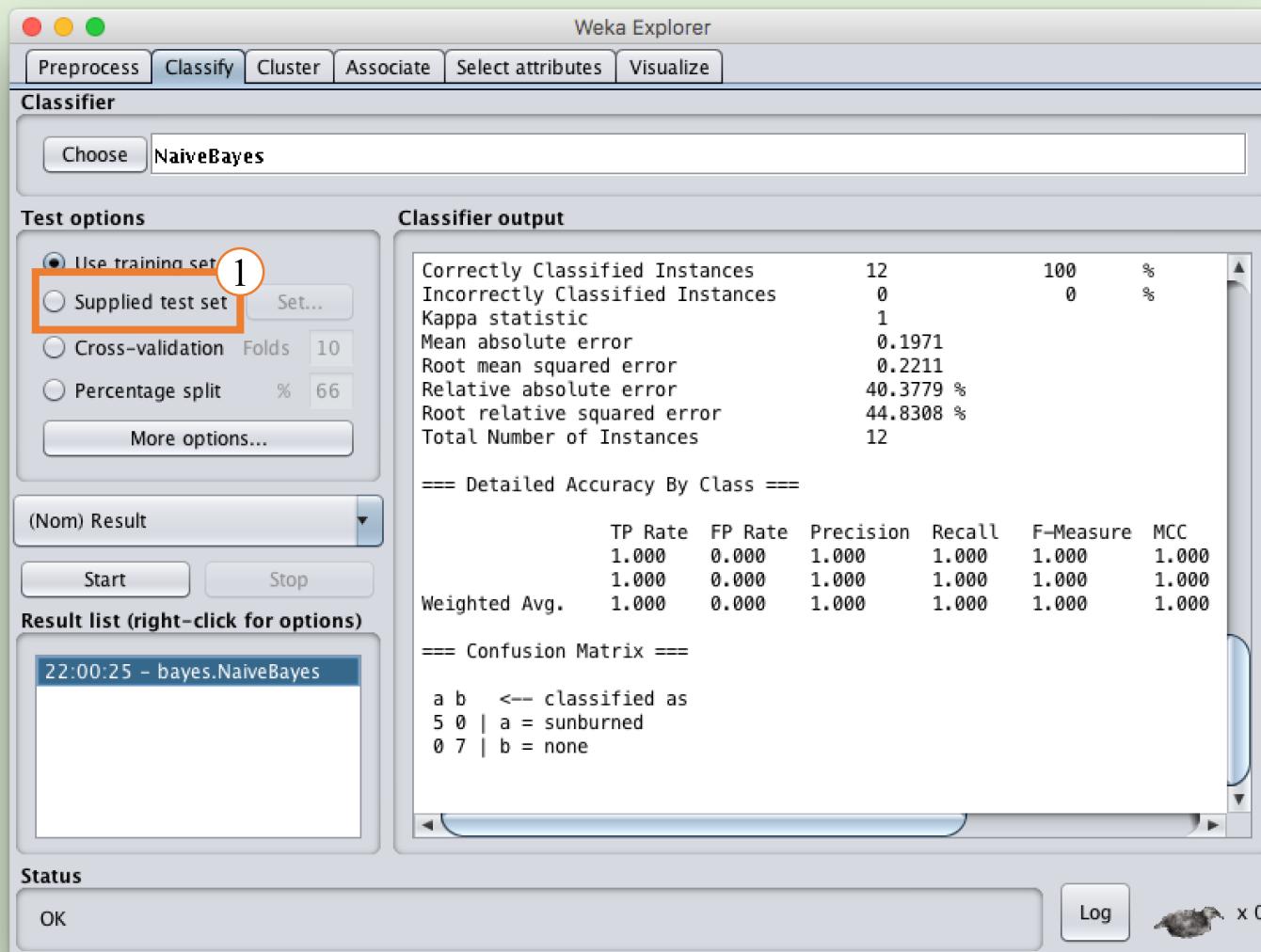
classified as =>	sunburned	none
sunburned	5	0
none	0	7

ค่าของข้อมูลจริง

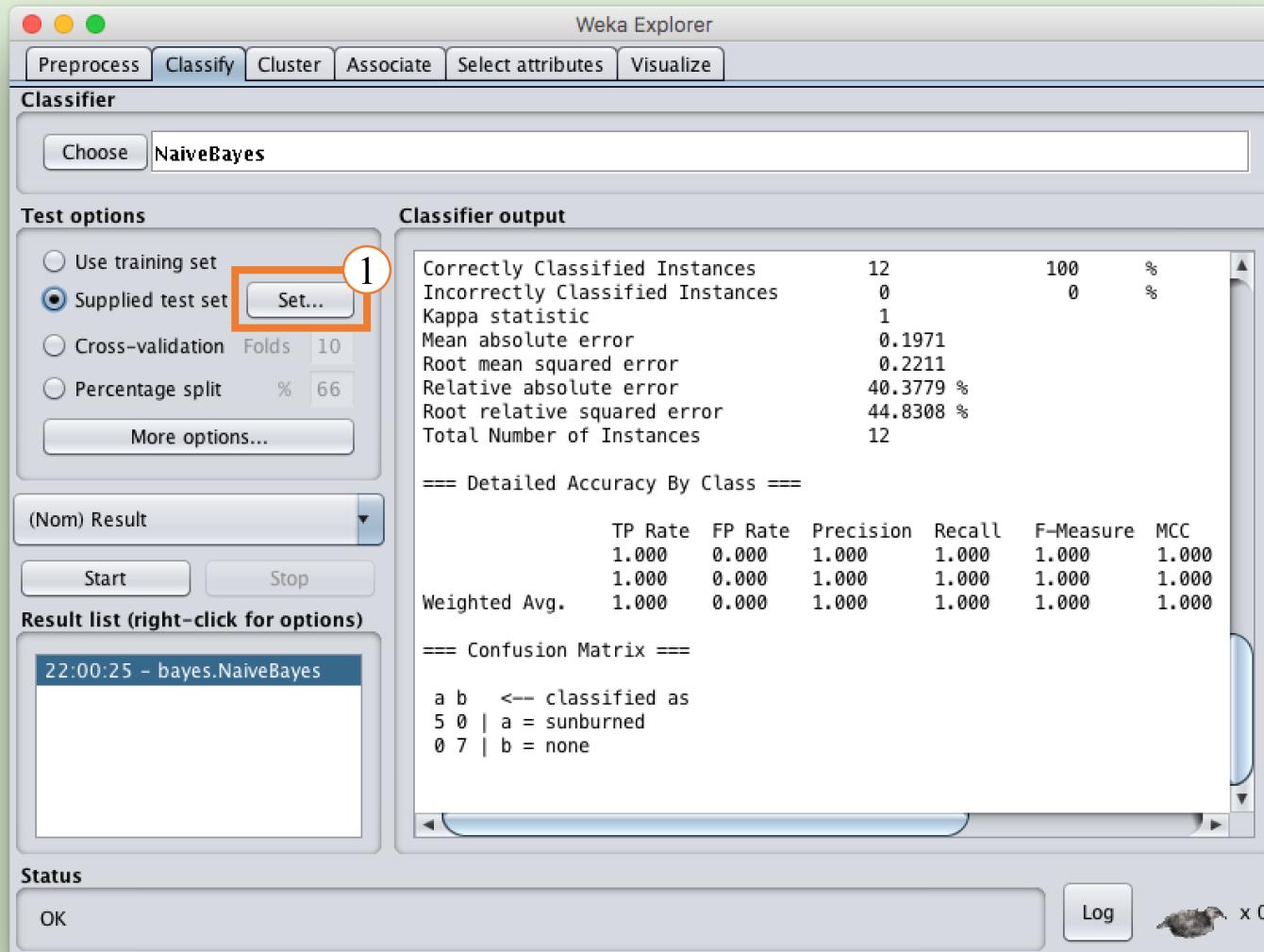
# การคาดการณ์ข้อมูลที่ไม่ทราบคำตอบ

Hair	Height	Weight	Lotion	Result
red	average	light	no	?
brown	tall	light	no	?
blonde	short	heavy	no	?

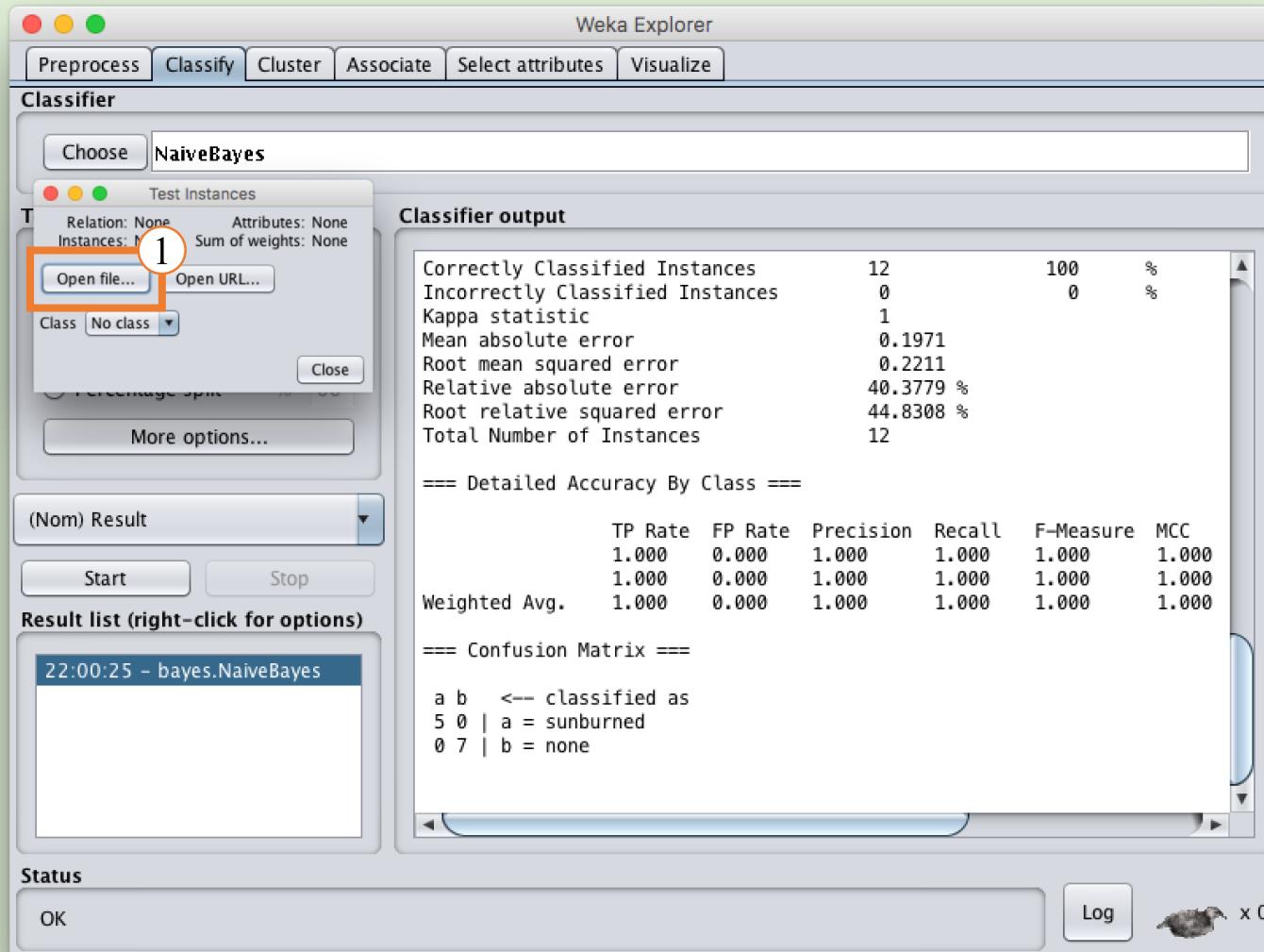
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



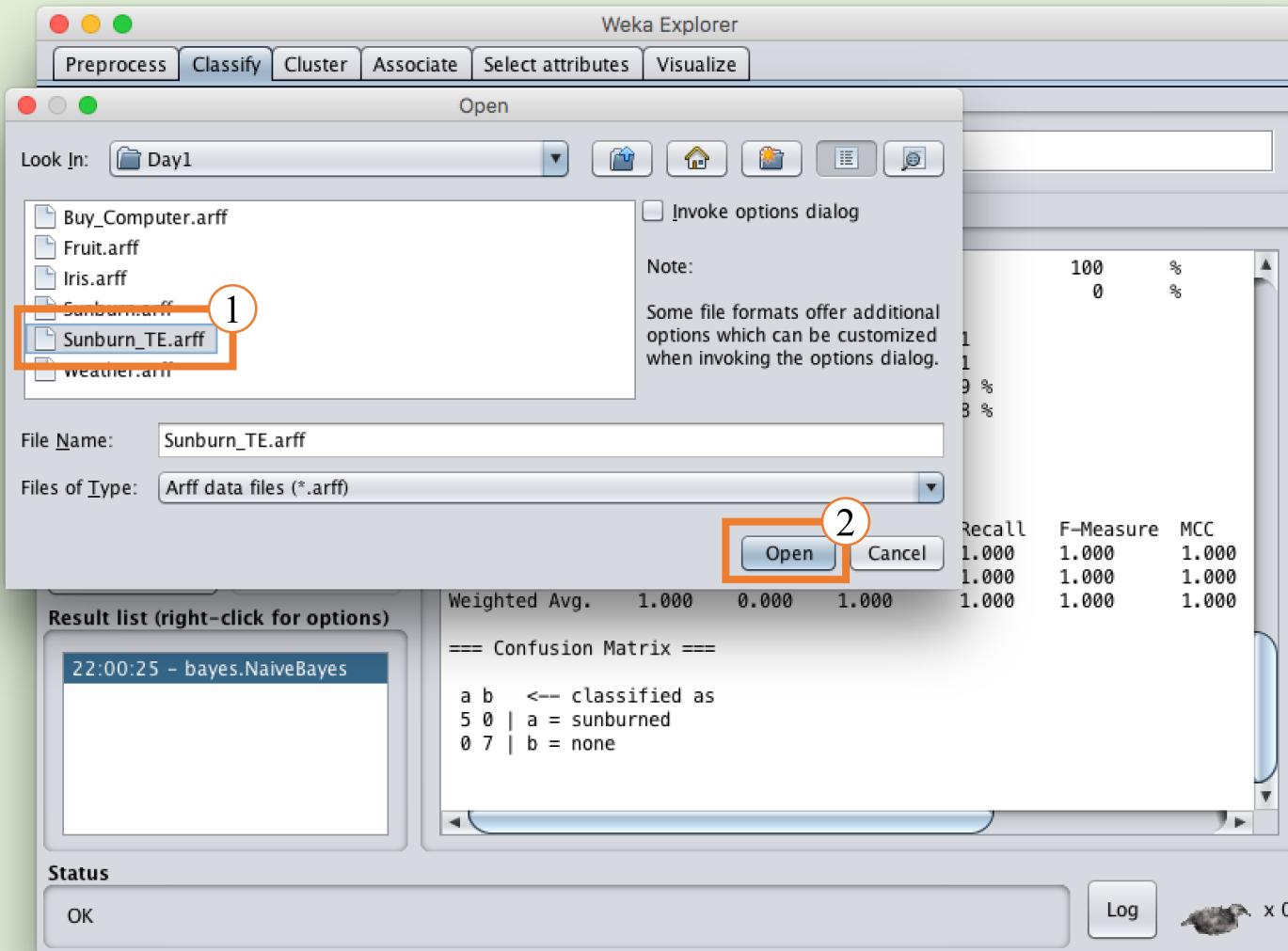
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



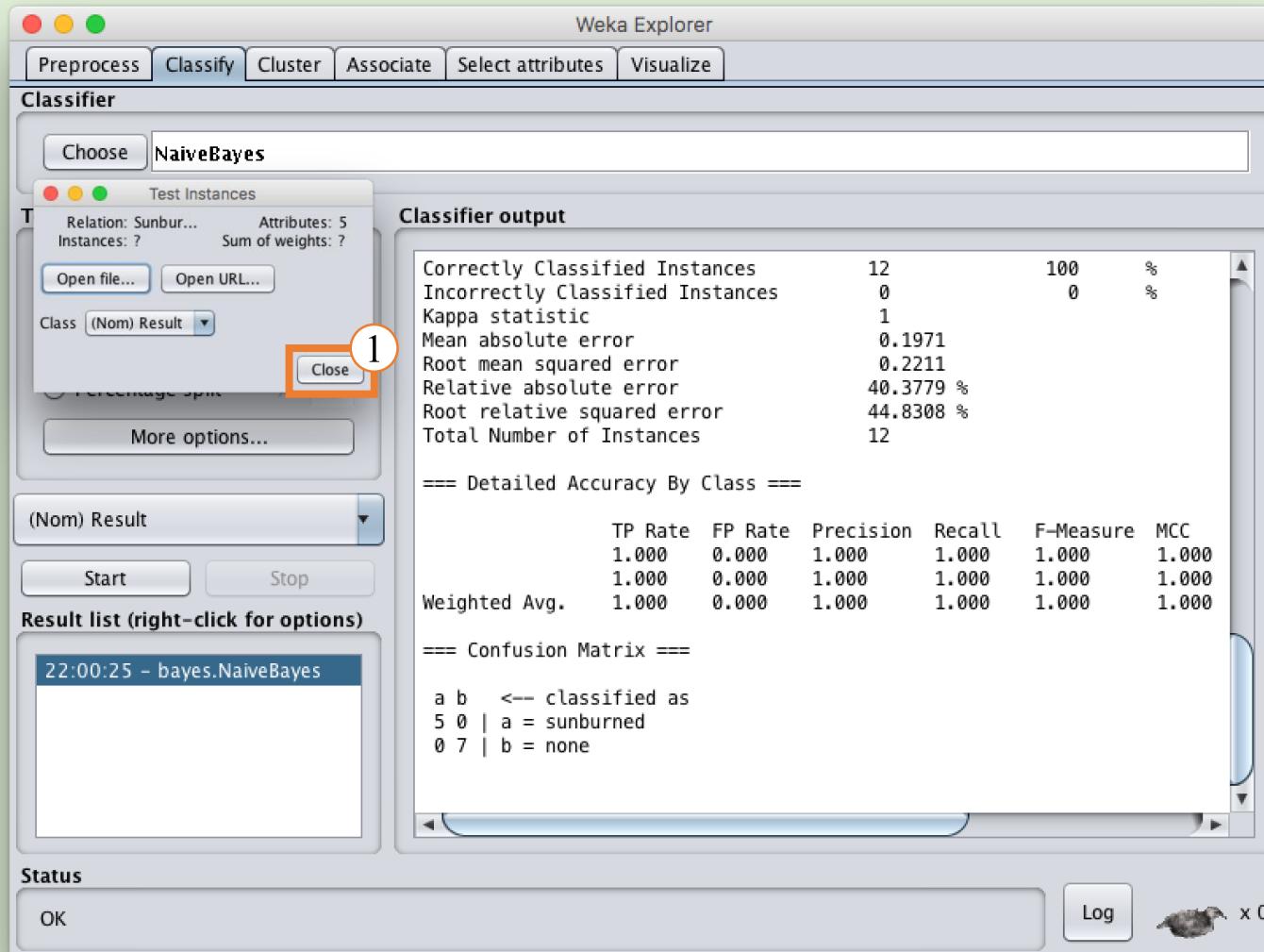
# การคัดการณ์ข้อมูลที่ไม่ทราบคำศوب



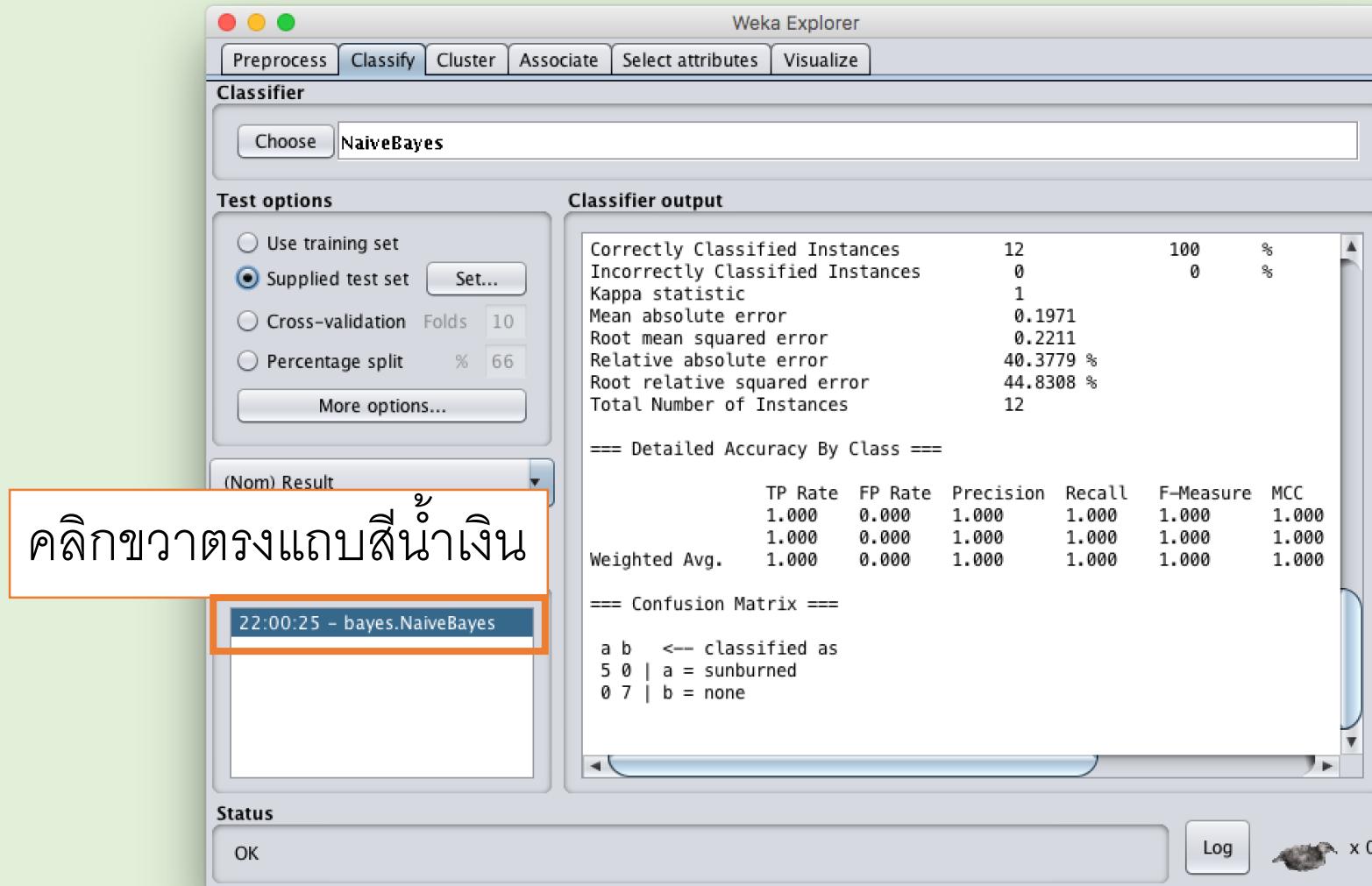
# การคัดกรณ์ข้อมูลที่ไม่ทราบคำต่อหน้า



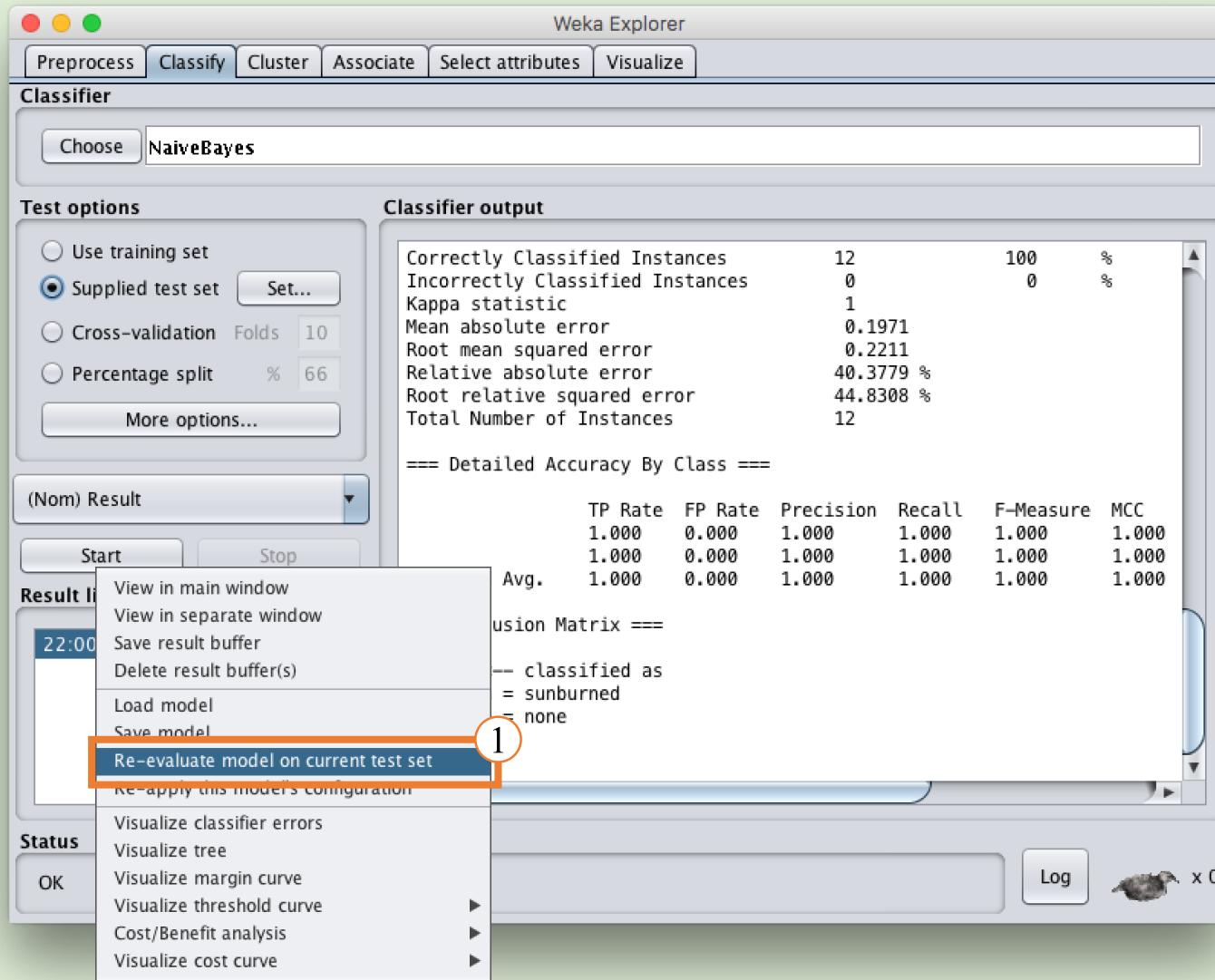
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



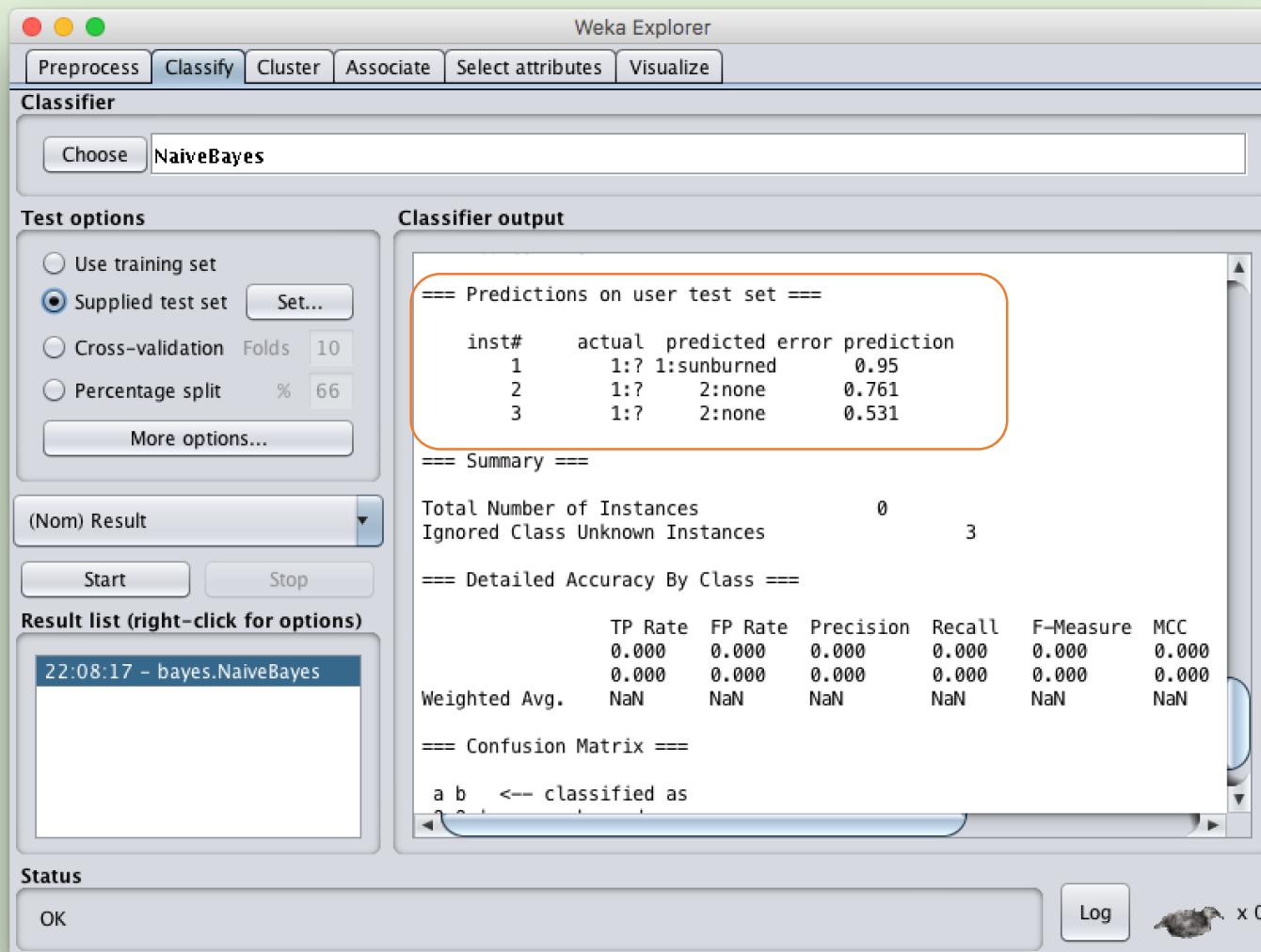
# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب



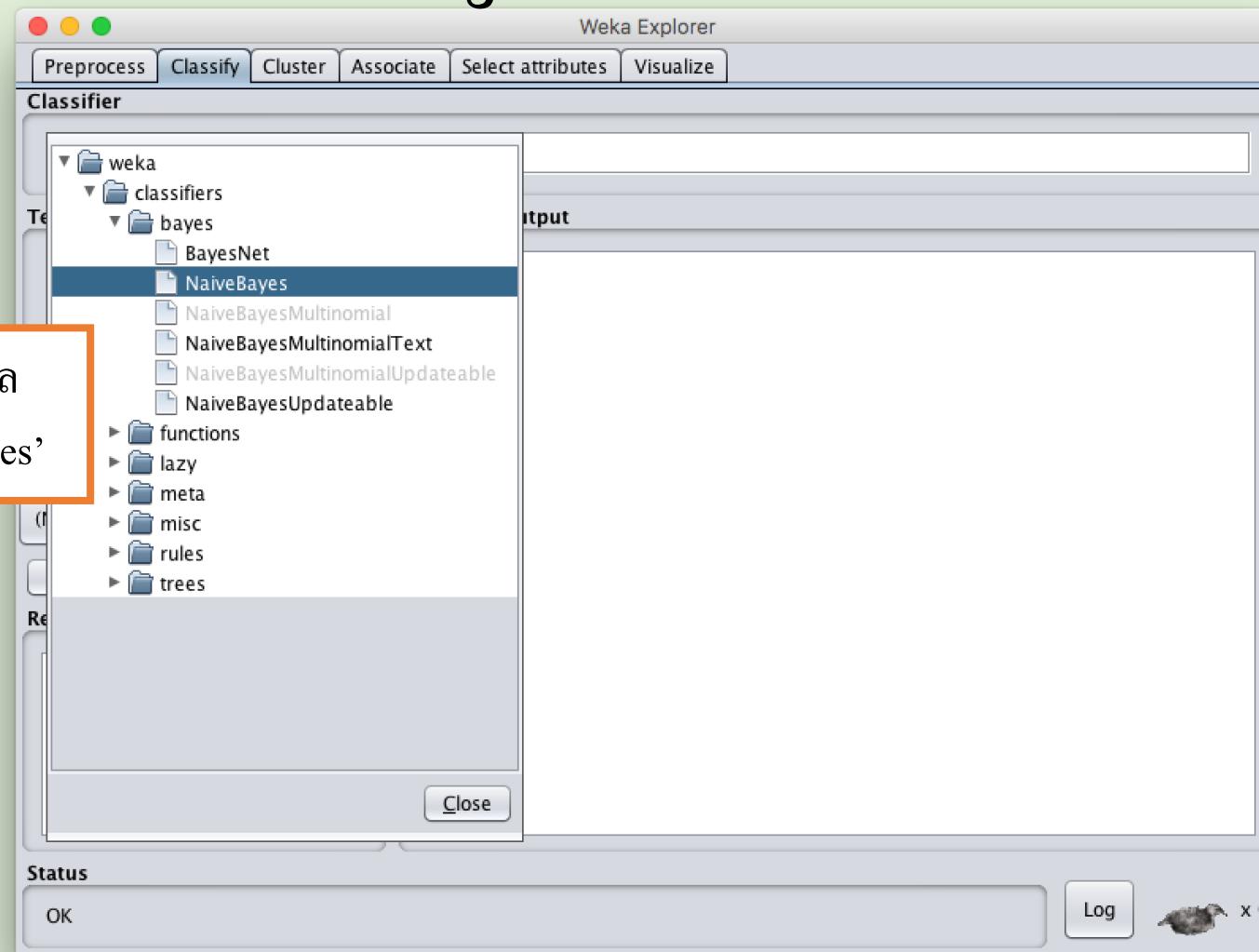
# តាមរយៈរបៀបទូទៅទី 2 : Weather

	outlook	temperature	humidity	windy	play
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

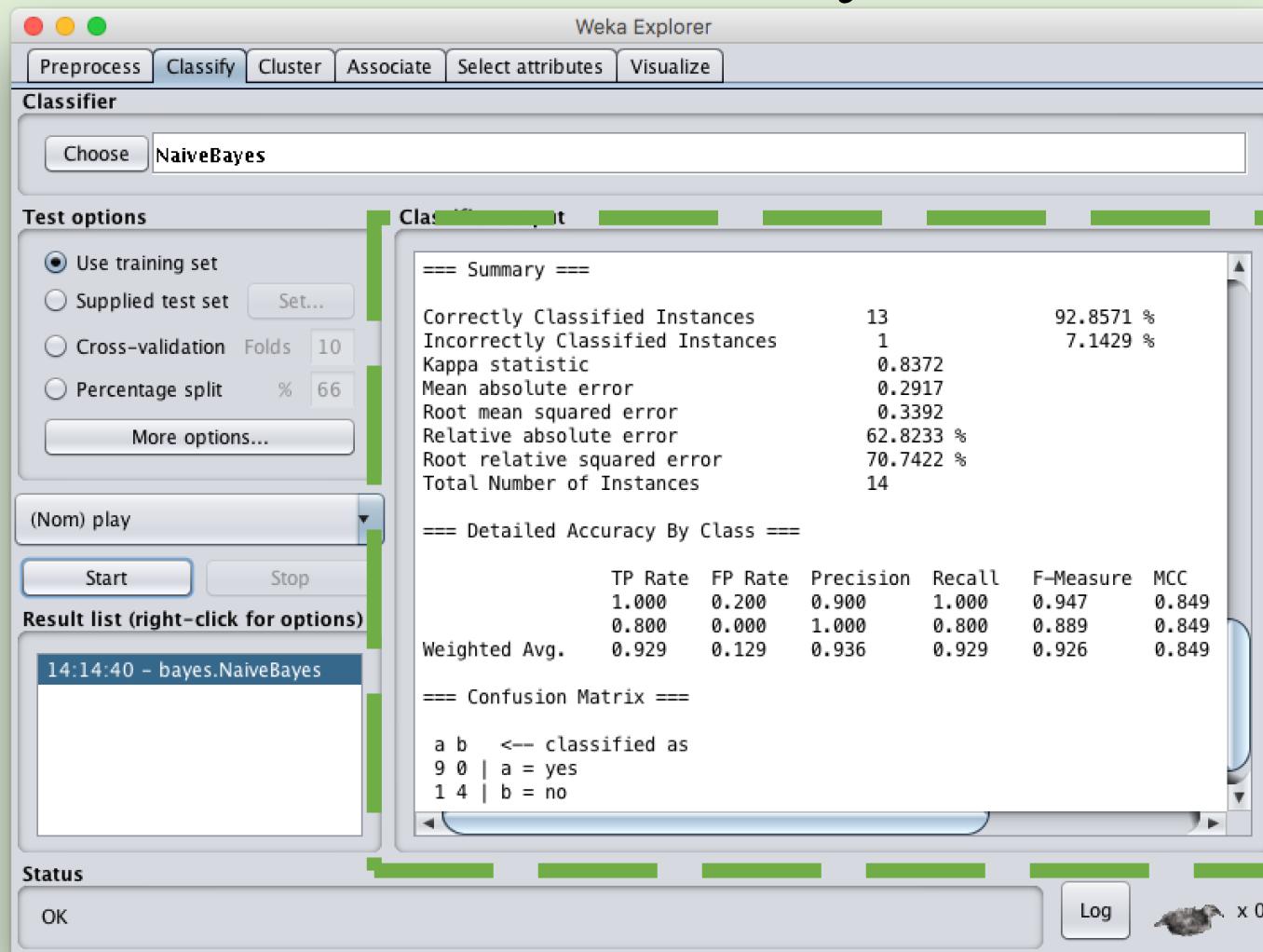
	Attributes				Class
	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	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

สภาพอากาศ		ระดับอุณหภูมิ	ระดับความชื้น	มีลม หรือไม่	ออกป่า <sup>เล่นหรือไม่</sup>
	outlook	temperature	humidity	windy	play
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

# ขั้นตอนการสร้าง Naive Bayes สำหรับ ปัญหา Weather



# ผล Classification ของปัญหา Weather ด้วย Naive Bayes



# ผล Classification ของ Naive Bayes

== Predictions on training set ==

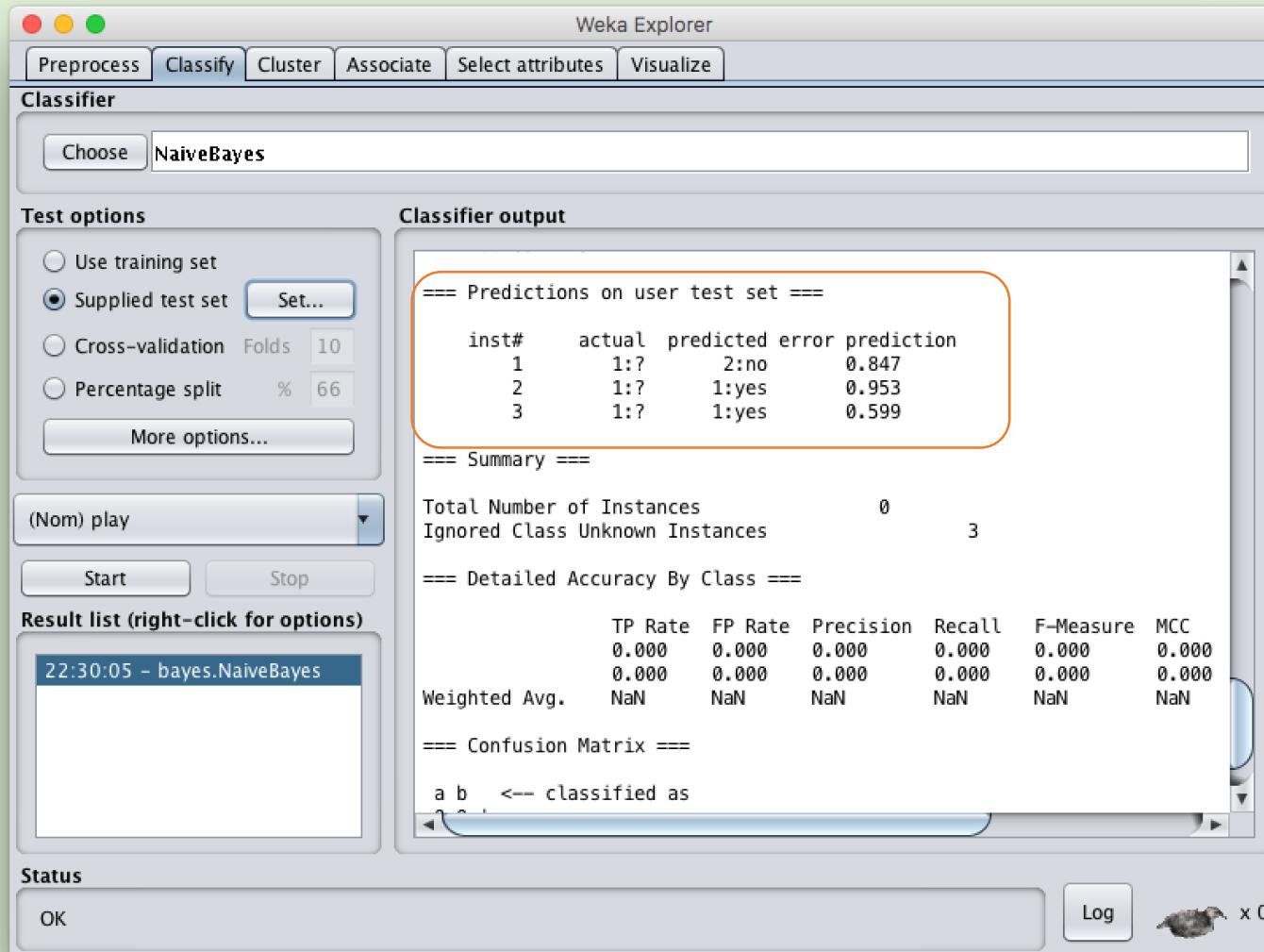
inst#	actual	predicted	error	prediction
1	2:no	2:no		0.704
2	2:no	2:no		0.847
3	1:yes	1:yes		0.737
4	1:yes	1:yes		0.554
5	1:yes	1:yes		0.867
6	2:no	1:yes	+	0.737
7	1:yes	1:yes		0.913
8	2:no	2:no		0.588
9	1:yes	1:yes		0.786
10	1:yes	1:yes		0.845
11	1:yes	1:yes		0.568
12	1:yes	1:yes		0.667
13	1:yes	1:yes		0.925
14	2:no	2:no		0.652

instance ที่ไม่เดล  
คาดการณ์ผิด

# การคาดการณ์ข้อมูลที่ไม่ทราบคำต่อไป

outlook	temperature	humidity	windy	play
sunny	hot	high	TRUE	?
overcast	mild	normal	FALSE	?
rainy	cool	high	FALSE	?

# การคาดการณ์ข้อมูลที่ไม่ทราบคำศوب

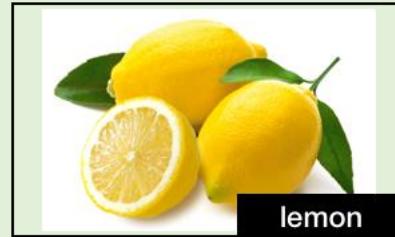


# ตัวอย่างปัญหาที่ 3 : Fruit

มีผลไม้อยู่ 4 ชนิด



mandarin



lemon



apple



orange

ตัวอย่างข้อมูลที่ใช้เคราะห์ชนิดผลไม้

	mass	width	height	color_score	fruit_name
1	192	8.4	7.3	0.55	apple
2	180	8	6.8	0.59	apple
	...	...	...	...	...
58	152	6.5	8.5	0.72	lemon
59	118	6.1	8.1	0.7	lemon

	Attributes				Class
	mass	width	height	color_score	fruit_name
1	<b>192</b>	8.4	7.3	0.55	apple
2	<b>180</b>	8	6.8	0.59	apple
3	<b>176</b>	7.4	7.2	0.6	apple
4	<b>178</b>	7.1	7.8	0.92	apple
5	<b>172</b>	7.4	7	0.89	apple
6	<b>166</b>	6.9	7.3	0.93	apple
	...	...	...	...	...
58	<b>152</b>	6.5	8.5	0.72	lemon
59	<b>118</b>	6.1	8.1	0.7	lemon

ประเภท

มวล

ความกว้าง

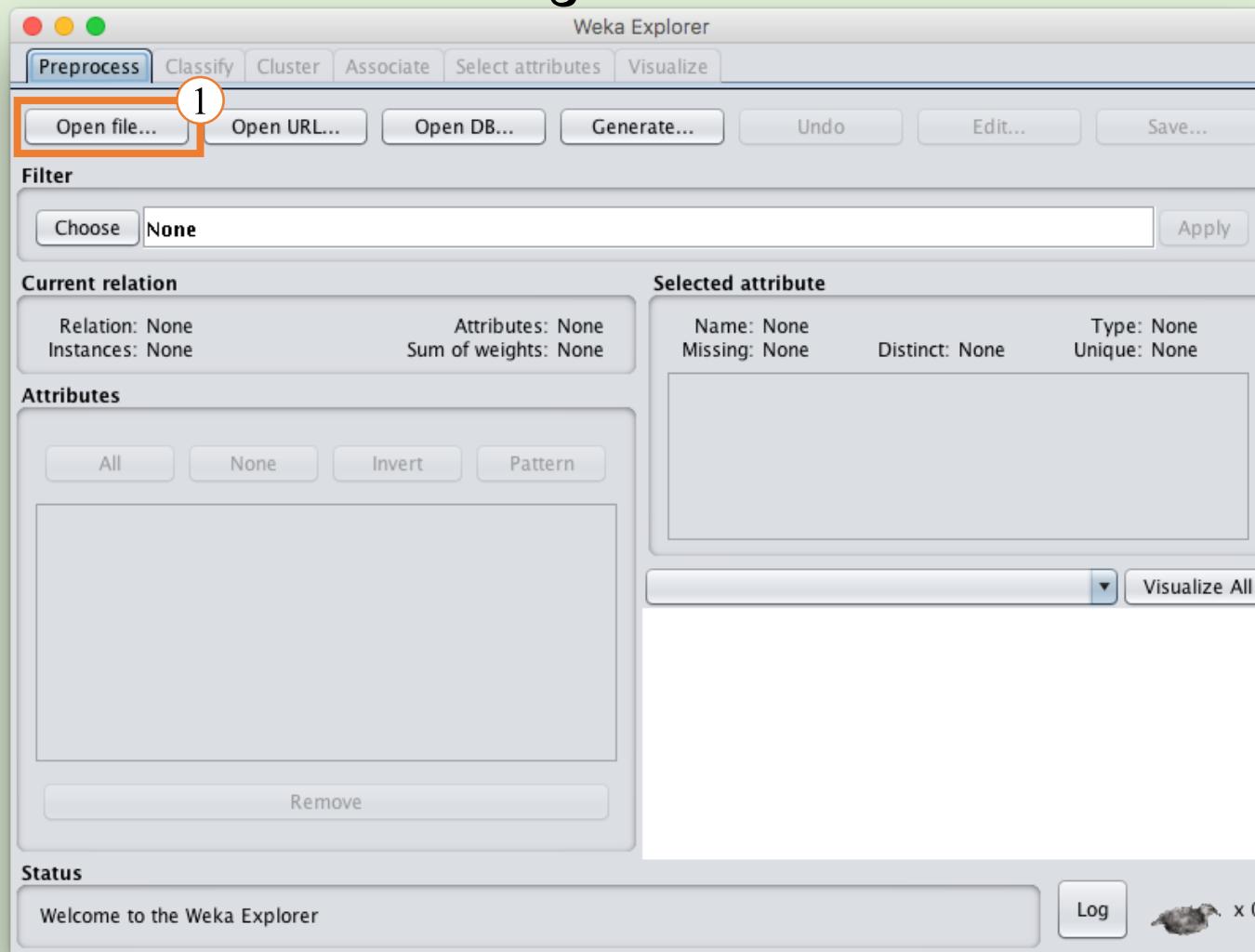
ความสูง

สี

ผลไม้

	mass	width	height	color_score	fruit_name
1	<b>192</b>	8.4	7.3	0.55	apple
2	<b>180</b>	8	6.8	0.59	apple
3	<b>176</b>	7.4	7.2	0.6	apple
4	<b>178</b>	7.1	7.8	0.92	apple
5	<b>172</b>	7.4	7	0.89	apple
6	<b>166</b>	6.9	7.3	0.93	apple
	...	...	...	...	...
58	<b>152</b>	6.5	8.5	0.72	lemon
59	<b>118</b>	6.1	8.1	0.7	lemon

# ขั้นตอนการสร้าง Naive Bayes สำหรับ ปัญหา Fruit

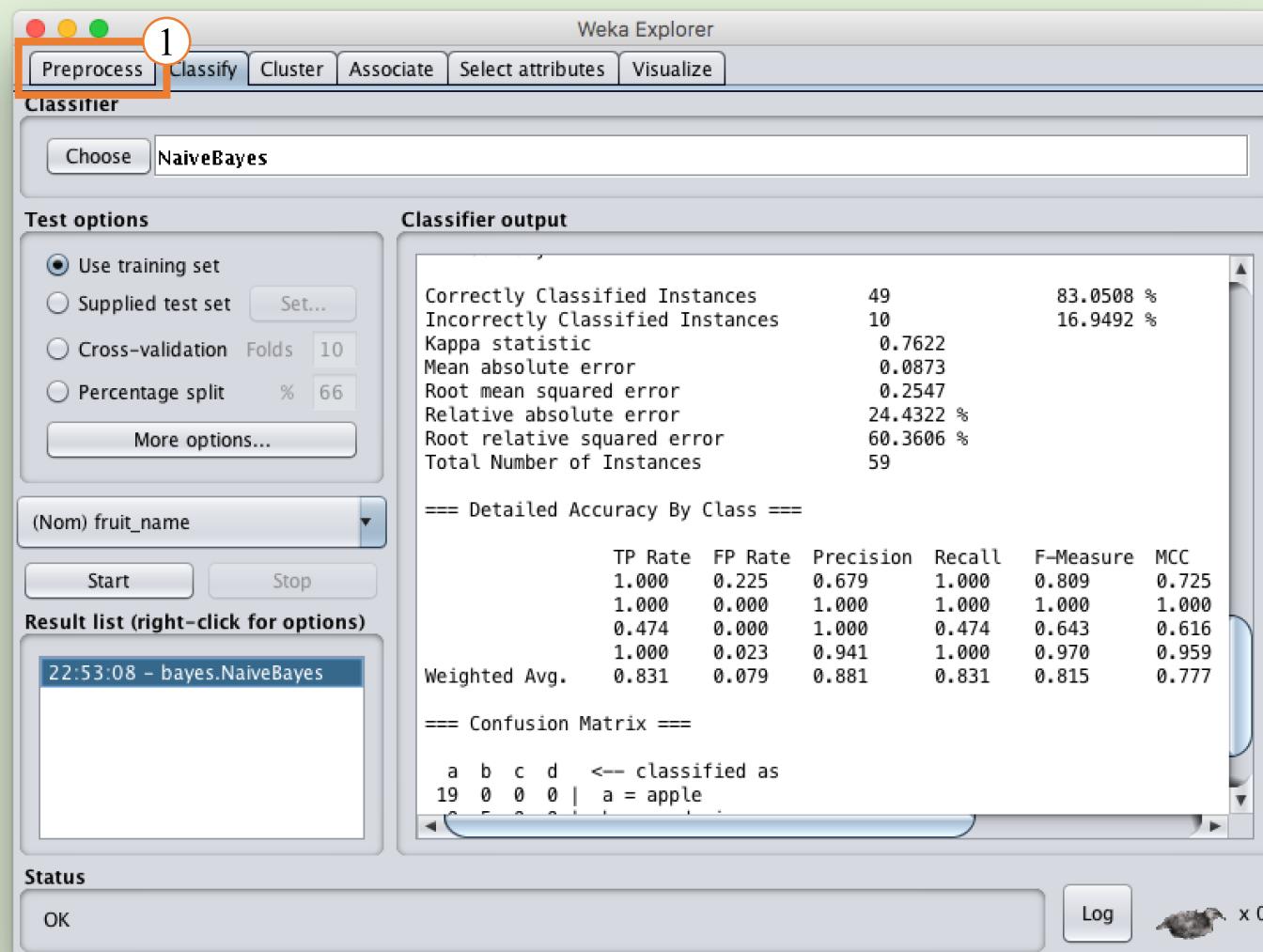


# ผล Classification ของปัญหา Fruit ด้วย Naive Bayes

สรุปเป็น Confusion Matrix ได้ดังนี้

classified as =>	apple	mandarin	orange	lemon
apple	19	0	0	0
mandarin	0	5	0	0
orange	9	0	9	1
lemon	0	0	0	16

# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter 1 Choose None Apply

Current relation

Relation: Fruit Attributes: 5 Instances: 59 Sum of weights: 59

Selected attribute

Name: mass Type: Numeric  
Missing: 0 (0%) Distinct: 40 Unique: 27 (46%)

Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

Attributes

All None Invert Pattern

No.	Name
1	mass
2	width
3	height
4	color_score
5	fruit_name

Remove

Class: fruit\_name (Nom) Visualize All

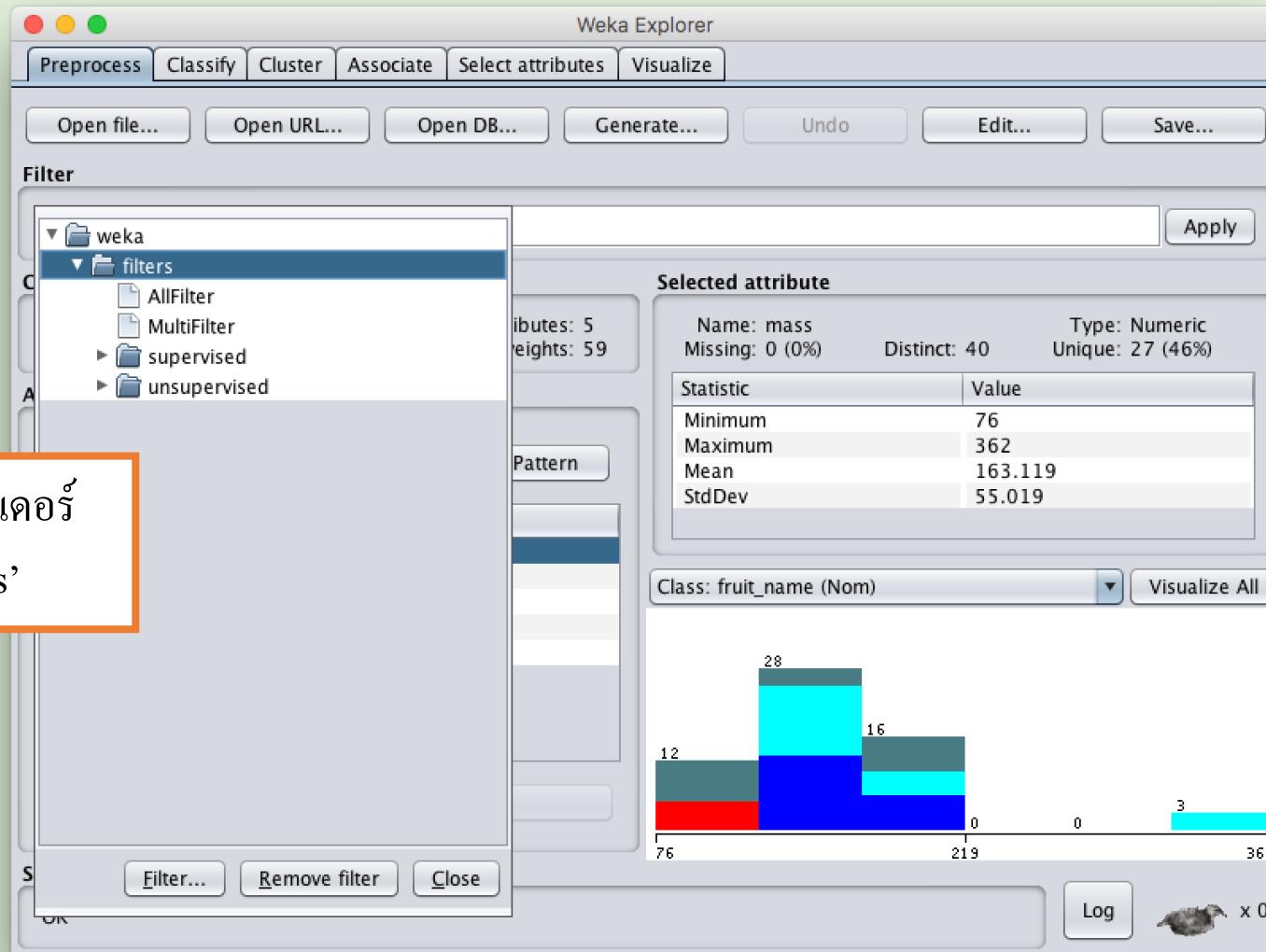
76 219 0 362

Status

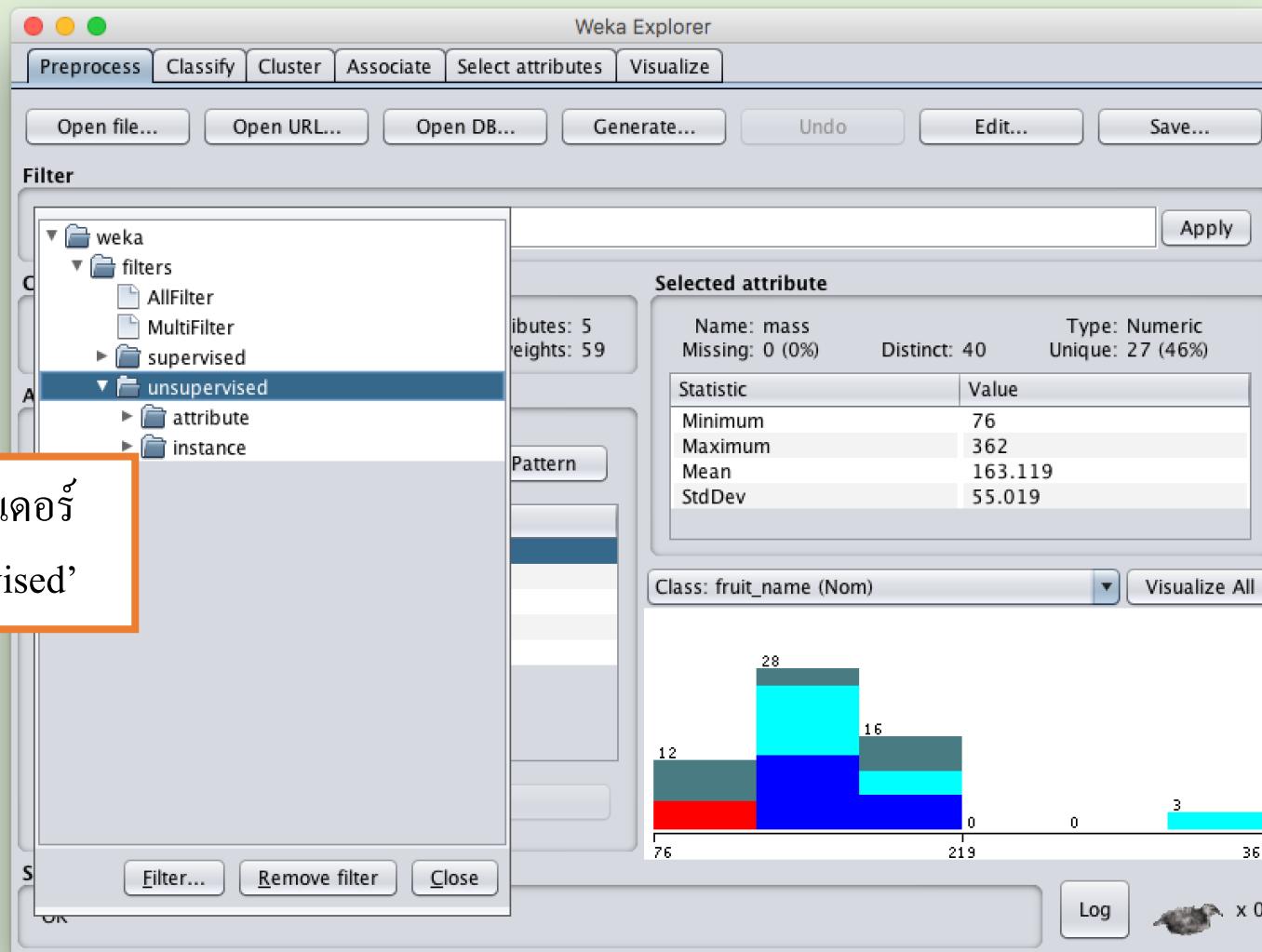
OK Log x 0

The 'Choose' button in the Filter section is highlighted with a red circle and labeled '1'.

# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

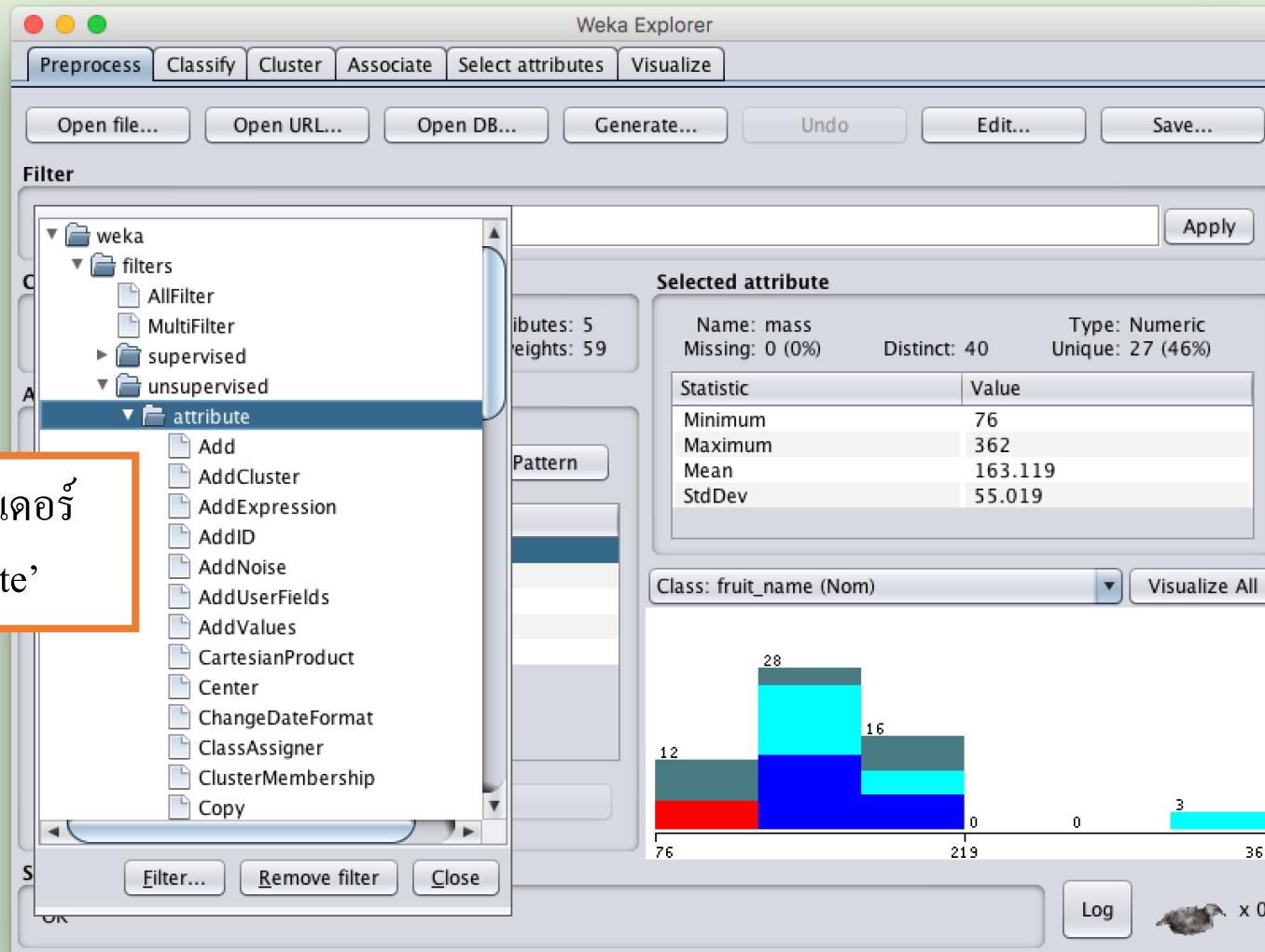


# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



เลือกโฟลเดอร์  
'unsupervised'

# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize



# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

เลือก 'Discretize'

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. A callout box highlights the 'Discretize' option in the 'Selected filter' list. The 'Selected attribute' panel displays statistics for 'mass': Name: mass, Type: Numeric, Missing: 0 (0%), Distinct: 40, Unique: 27 (46%). Below this is a table of statistics:

Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

# ปรับข้อมูลให้มีค่าเป็นช่วงด้วย Discretize

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **Discretize -B 10 -M -1.0 -R first-last** Apply 1

Current relation

Relation: Fruit Attributes: 5 Instances: 59 Sum of weights: 59

Attributes

All None Invert Pattern

No.	Name
1	mass
2	width
3	height
4	color_score
5	fruit_name

Remove

Selected attribute

Name: mass Type: Numeric  
Missing: 0 (0%) Distinct: 40 Unique: 27 (46%)

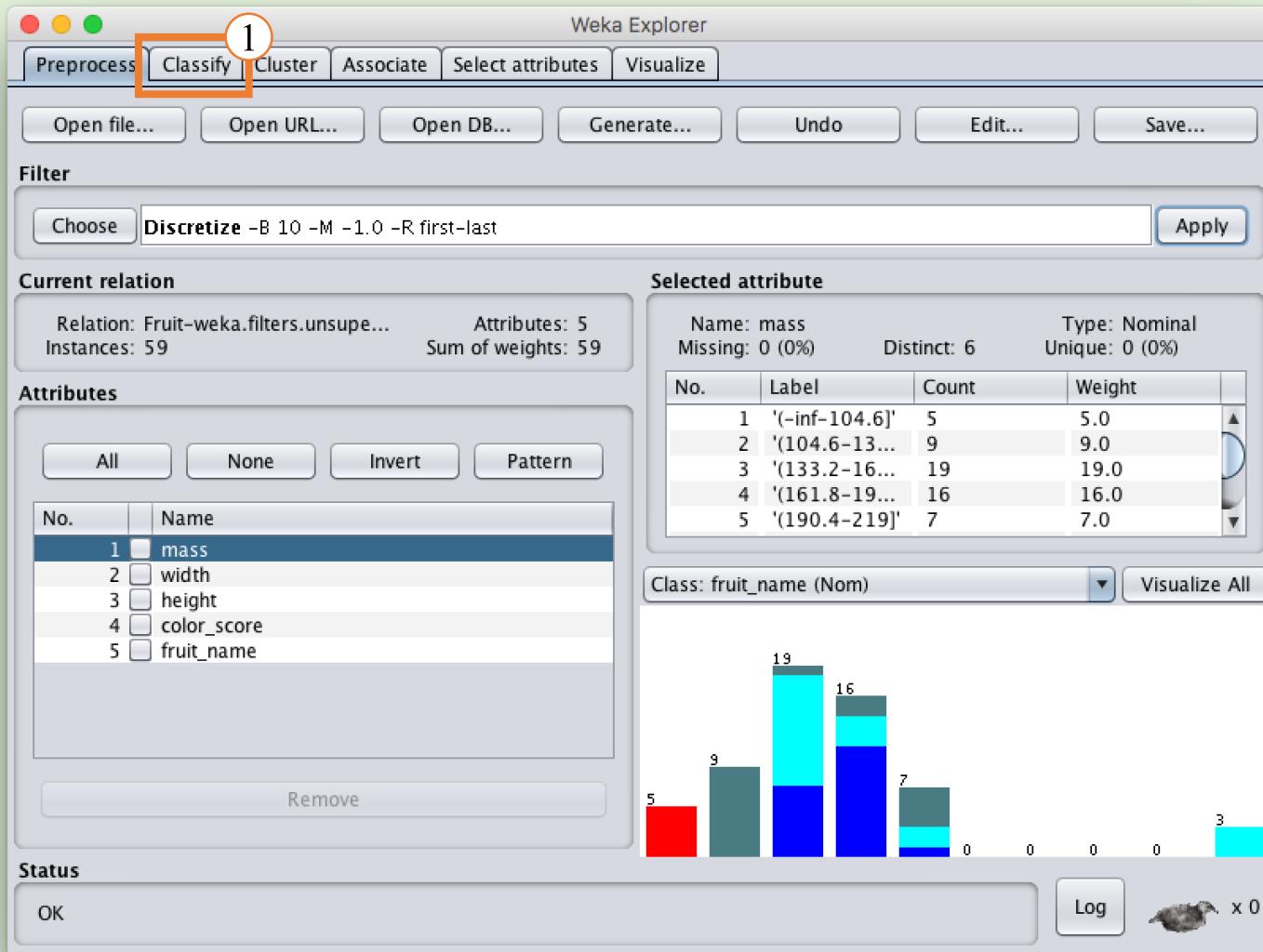
Statistic	Value
Minimum	76
Maximum	362
Mean	163.119
StdDev	55.019

Class: fruit\_name (Nom) Visualize All

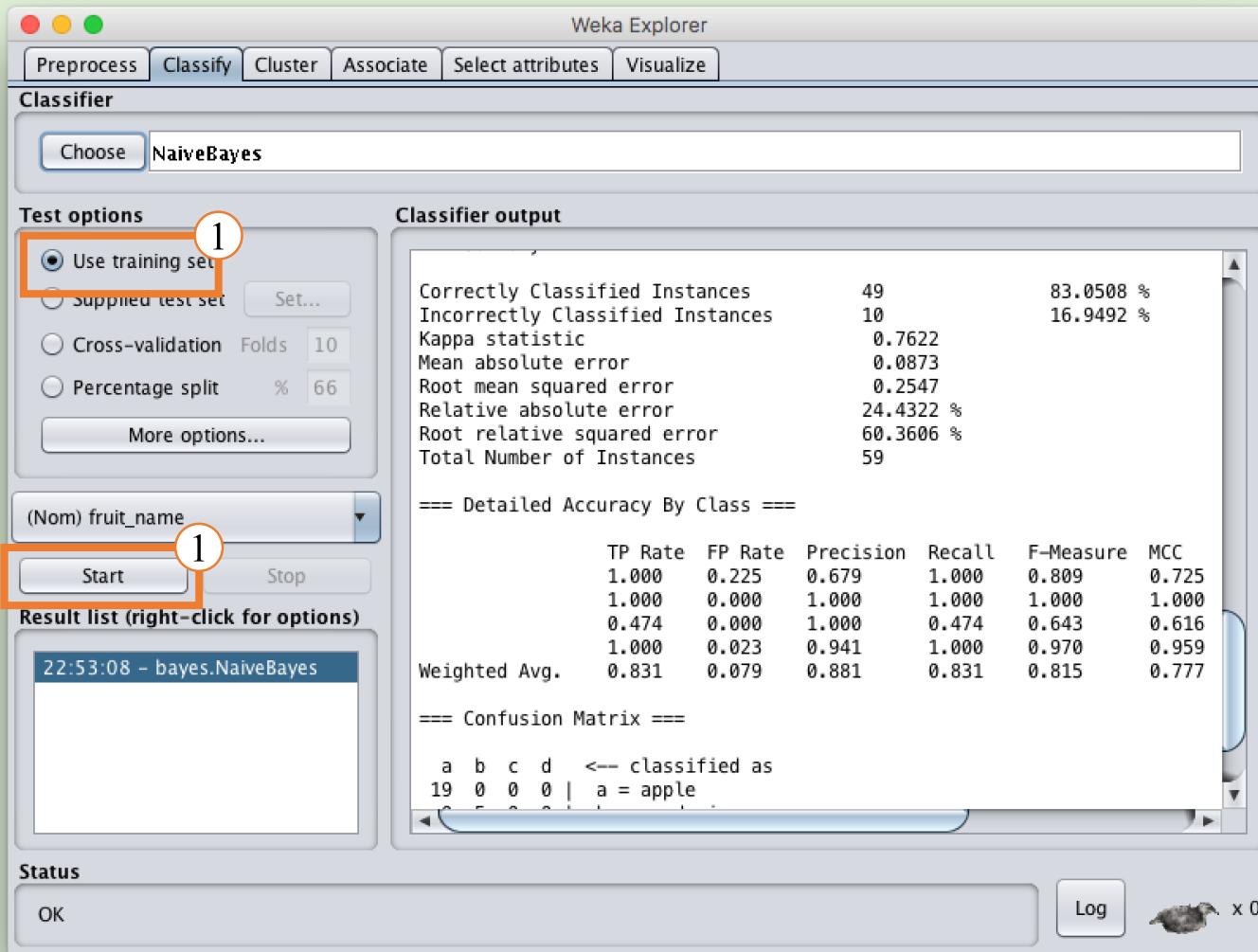
Status

OK Log

# สร้าง Naive Bayes ใหม่



# สร้าง Naive Bayes ใหม่



# ผล Classification ของ Naive Bayes

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose NaiveBayes

**Test options**

Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66  
More options...

(Nom) fruit\_name

Start Stop

**Result list (right-click for options)**

22:53:08 - bayes.NaiveBayes  
22:55:04 - bayes.NaiveBayes

**Classifier output**

== Summary ==

Correctly Classified Instances	57	96.6102 %
Incorrectly Classified Instances	2	3.3898 %
Kappa statistic	0.9524	
Mean absolute error	0.0633	
Root mean squared error	0.1283	
Relative absolute error	17.7219 %	
Root relative squared error	30.3986 %	
Total Number of Instances	59	

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	R
1.000	0.025	0.950	1.000	0.974	0.962	1	
1.000	0.000	1.000	1.000	1.000	1.000	1	
0.895	0.000	1.000	0.895	0.944	0.923	0	
1.000	0.023	0.941	1.000	0.970	0.959	1	
Weighted Avg.	0.966	0.014	0.968	0.966	0.966	0.952	1

== Confusion Matrix ==

a	b	c	d	<-- classified as
19	0	0	0	a = apple
0	5	0	0	b = mandarin
1	0	17	1	c = orange
0	0	0	16	d = lemon

Status Log x 0

OK

# ผล Classification ของปัญหา Fruit ด้วย Naive Bayes

สรุปเป็น Confusion Matrix ได้ดังนี้

classified as =>	apple	mandarin	orange	lemon
apple	19	0	0	0
mandarin	0	5	0	0
orange	1	0	17	1
lemon	0	0	0	16

# เอกสารอ้างอิง

- ปัญญาประดิษฐ์ Artificial Intelligence โดย บัญเสริม กิจ  
ศิริกุล ภาควิชาคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
- Solving A Simple Classification Problem with Python  
([Link](#))
- Kaggle Buy Computer ([Link](#))