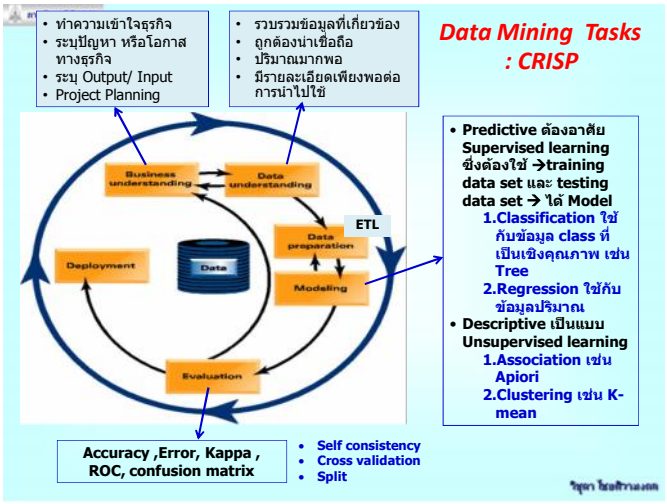
การใช้โปรแกรม WEKA

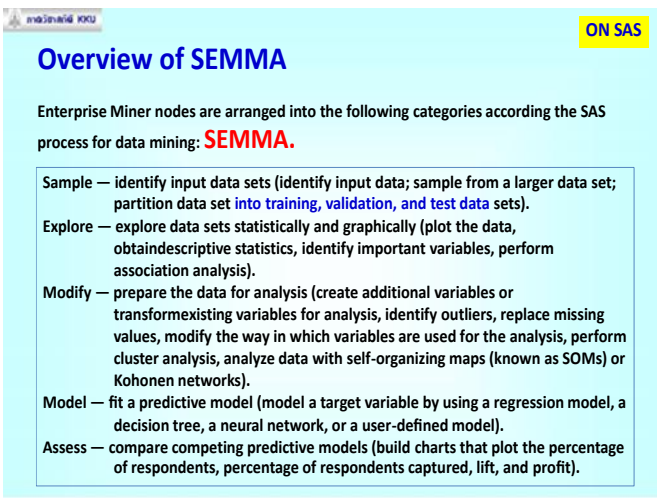
Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The Explorer interface features several panels providing access to the main components of the workbench:

- **The Preprocess panel** has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- **The Classify panel** enables the user to apply classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
- **The Associate panel** provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- **The Cluster panel** gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- **The Select attributes panel** provides algorithms for identifying the most predictive attributes in a dataset.
- **The Visualize panel** shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

โปรแกรม Weka เริ่มพัฒนามาตั้งแต่ปี 1997 โดยมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ เป็นซอฟต์แวร์สำเร็จภาพประกอบประเภทฟรีแวร์ อยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม Weka ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเขียนมาโดยเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และ การทำเหมืองข้อมูล (Data Mining) โปรแกรมจะประกอบไปด้วยโมดูลย่อย ๆ สำหรับใช้ในการจัดการข้อมูล และเป็นโปรแกรมที่สามารถใช้ Graphic User Interface (GUI) โดยมีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูล ได้แก่ Pre-Processing, Classification, Regression, Clustering, Association rules, Selection และ Visualization

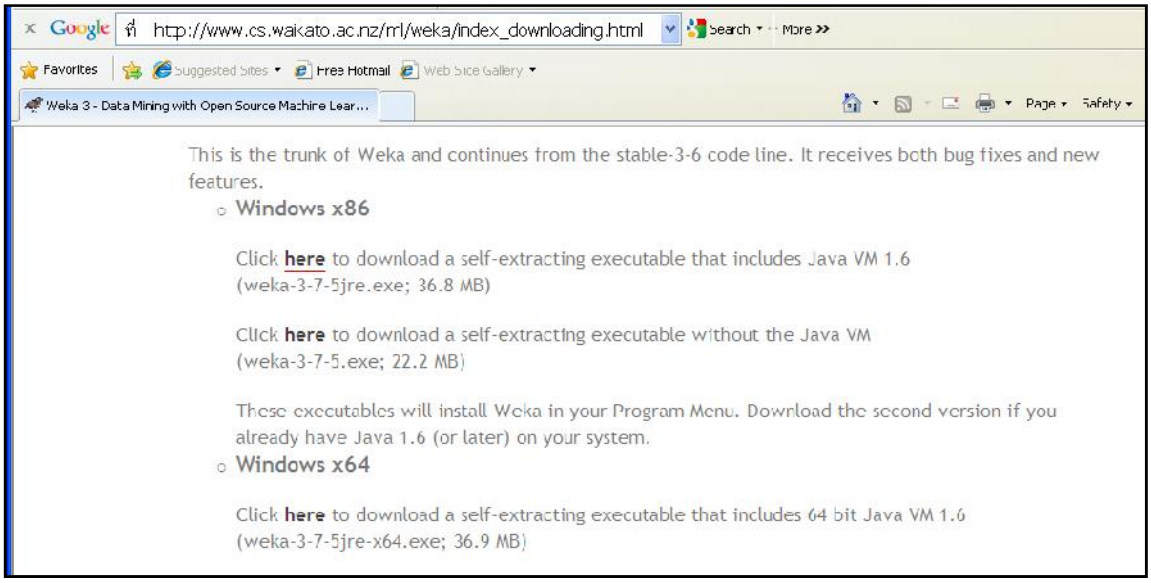




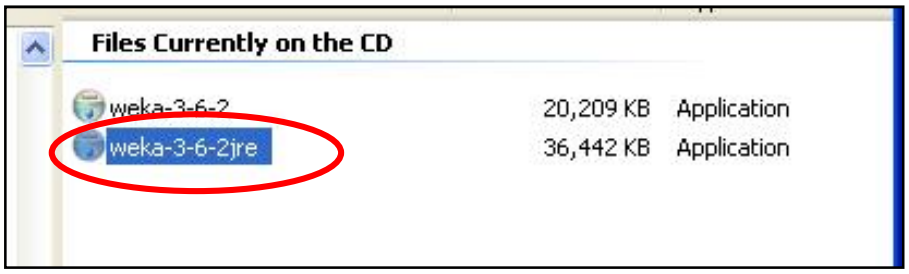
การ Install Software

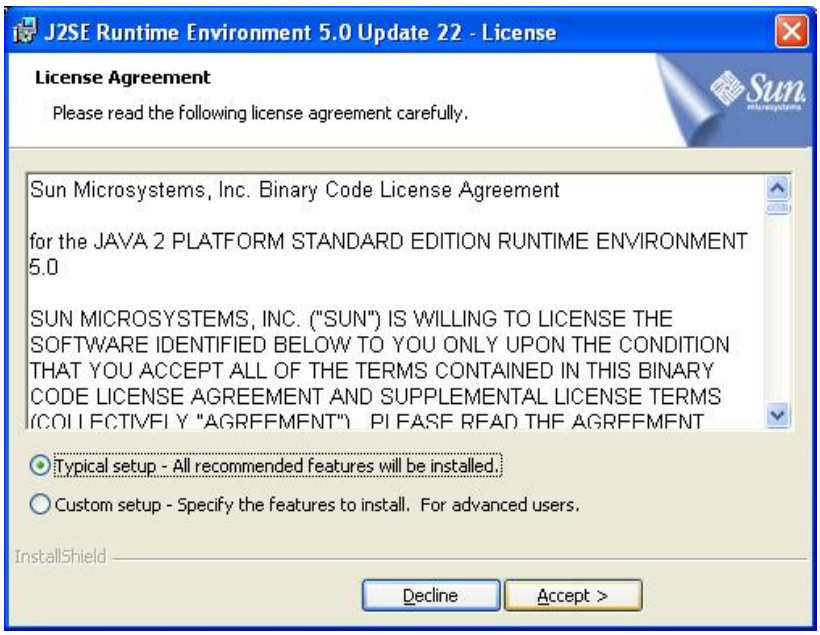
กรณีที่ในเครื่องท่านยัง**ไม่มี JAVA** ให้ download file แบบ **includes JAVA** โดยสามารถเข้าที่ website ต่อไปนี้

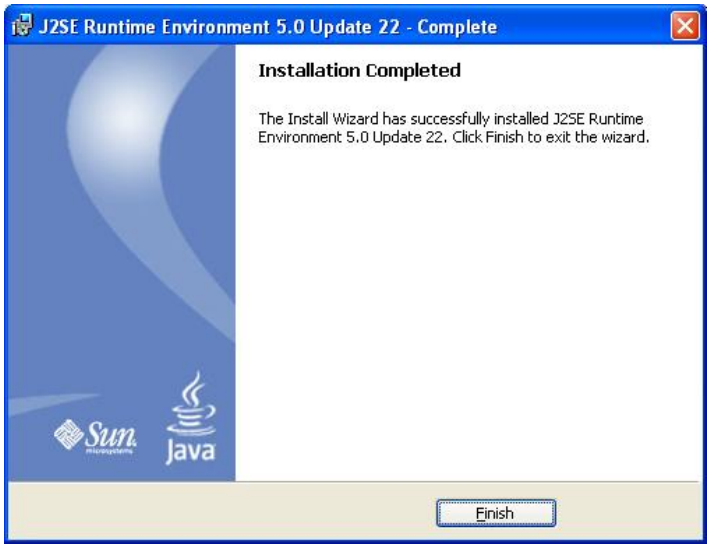
http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html

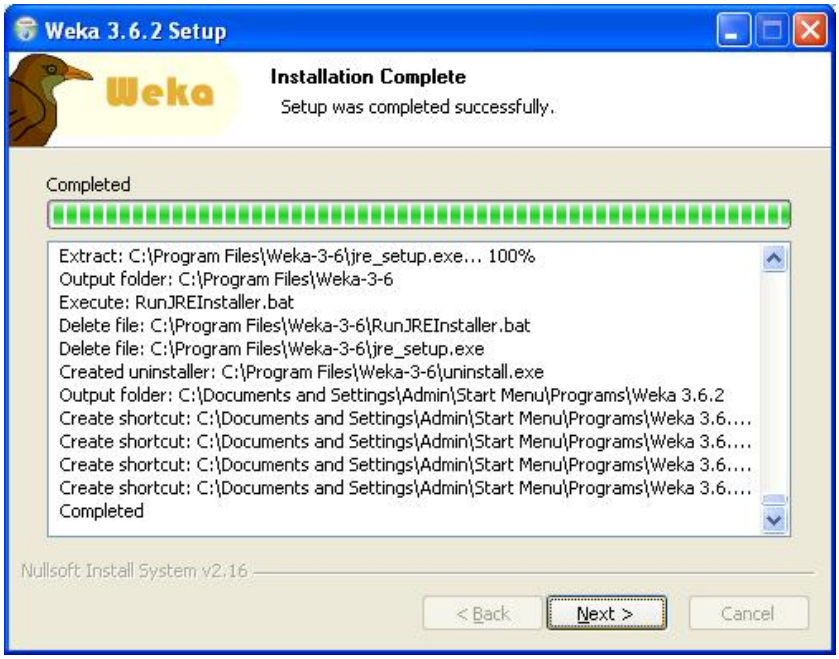


จากนั้นทำการ run

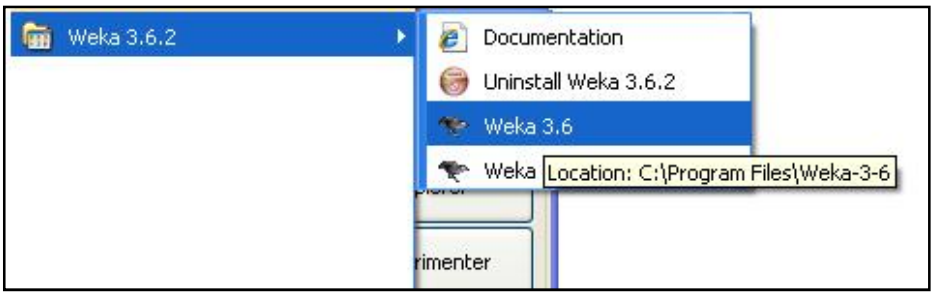




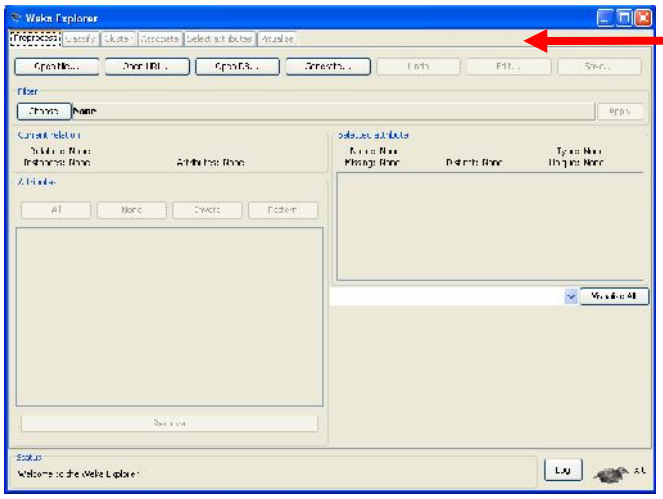

J2SE Runtime Environment 5.0 Update 22 - Complete
Installation Completed
The Install Wizard has successfully installed J2SE Runtime Environment 5.0 Update 22. Click Finish to exit the wizard.
Sun
Java
Finish


Weka 3.6.2 Setup
Weka
Installation Complete
Setup was completed successfully.
Completed
Extract: C:\Program Files\Weka-3-6\jre_setup.exe... 100%
Output folder: C:\Program Files\Weka-3-6
Execute: RunJREInstaller.bat
Delete file: C:\Program Files\Weka-3-6\RunJREInstaller.bat
Delete file: C:\Program Files\Weka-3-6\jre_setup.exe
Created uninstaller: C:\Program Files\Weka-3-6\uninstall.exe
Output folder: C:\Documents and Settings\Admin\Start Menu\Programs\Weka 3.6.2
Create shortcut: C:\Documents and Settings\Admin\Start Menu\Programs\Weka 3.6....
Create shortcut: C:\Documents and Settings\Admin\Start Menu\Programs\Weka 3.6....
Create shortcut: C:\Documents and Settings\Admin\Start Menu\Programs\Weka 3.6....
Create shortcut: C:\Documents and Settings\Admin\Start Menu\Programs\Weka 3.6....
Completed
Nullsoft Install System v2.16
< Back
Next >
Cancel

การเรียกใช้

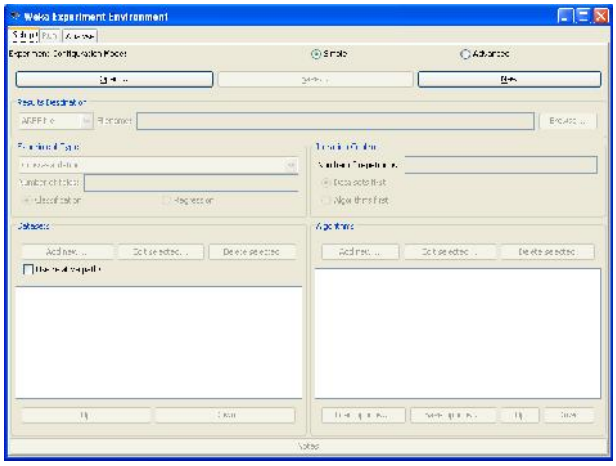

Weka 3.6.2
Documentation
Uninstall Weka 3.6.2
Weka 3.6
Weka
Location: C:\Program Files\Weka-3-6

- Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
- Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
- KnowledgeFlow เป็นโปรแกรมออกแบบผังการไหลของความรู้
- Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งการทำงานผ่านการพิมพ์

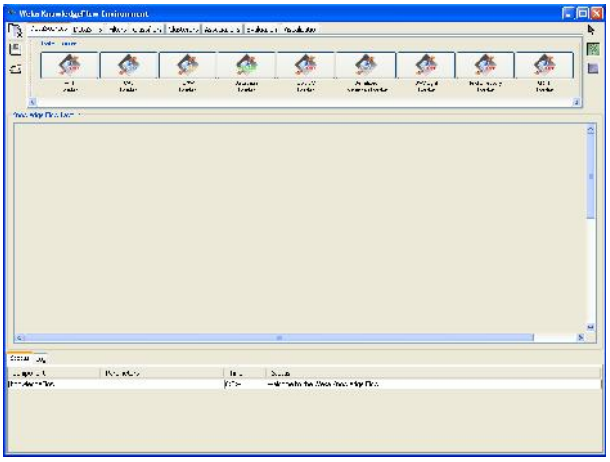


Explorer

- **Preprocess** การเตรียมข้อมูล เลือก file input พิจารณารายละเอียดข้อมูล แก้ไขข้อมูล แปลงข้อมูล
- **Classify** เป็นโมดูลการทำเหมืองข้อมูลแบบการจำแนกประเภท (Classification) จำแนกประเภทข้อมูลทำนายค่าข้อมูลใหม่จาก train model
- **Cluster** เป็นโมดูลการทำเหมืองข้อมูลแบบการแบ่งกลุ่ม (Clustering) แบ่งกลุ่มข้อมูลตามความคล้ายคลึง (Similarity)
- **Associate** เป็นโมดูลการทำเหมืองข้อมูลแบบกฎความสัมพันธ์ (Association rule)
- **Select attribute** คัดเลือกตัวแปรที่สำคัญ
- **Visualize** แสดงผลของข้อมูลในรูปแบบต่างๆ สองมิติ



Experimenter



KnowledgeFlow

การนำข้อมูลเข้าสู่ WEKA

ประเภทของแฟ้มข้อมูลที่รับได้

- แฟ้มข้อมูลที่รับต้องเป็น ARFF หรือ CSV
- ในกรณีที่แฟ้มข้อมูลอยู่ในเครือข่ายสามารถเรียกใช้ผ่าน URL ได้
- สามารถเรียกใช้มูลจากฐานข้อมูลได้ โดยเชื่อมโยงผ่าน JDBC

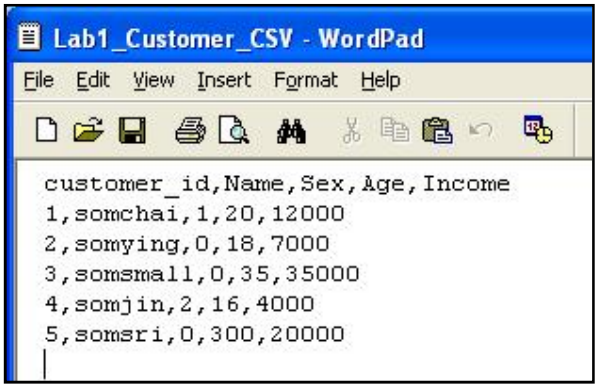
กรณี Attribute Relationship File Format (ARFF) is the text format file used by Weka to store data in a database

```
@relation weather
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
```

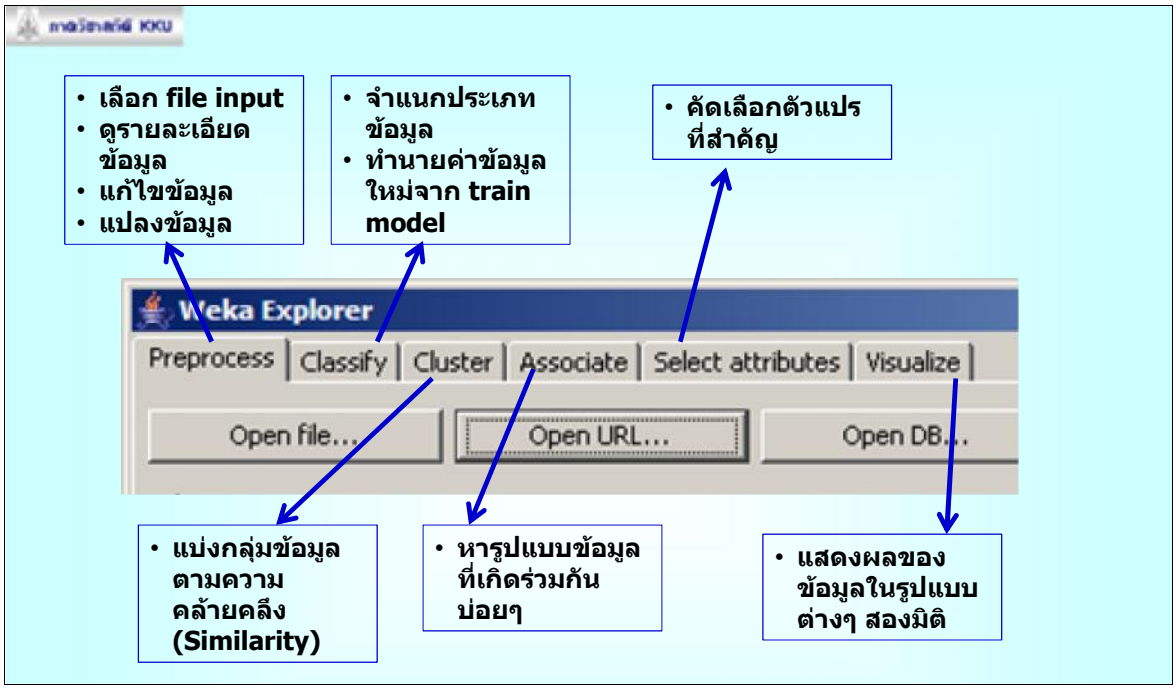
- **@relation *name*** เป็นบรรทัดที่บอกชื่อตารางข้อมูลเชิงสัมพันธ์
- **@attribute *att-name type*** เป็นบรรทัดที่บอกชื่อลักษณะเฉพาะและชนิด
  - numeric หรือ real หมายถึงลักษณะเฉพาะที่เก็บค่าเป็นตัวเลข
  - $\{v_1, v, \ldots, v_n\}$ หมายถึงลักษณะเฉพาะที่เก็บค่าไม่ต่อเนื่อง
- **@data** เป็นบรรทัดที่บอกถึงแถวที่ตามมาจะเป็นข้อมูล โดยแต่ละแถวจะแทนหนึ่งตัวอย่างข้อมูลซึ่งเรียงตามค่าของลักษณะเฉพาะที่บอกไว้ข้างต้น

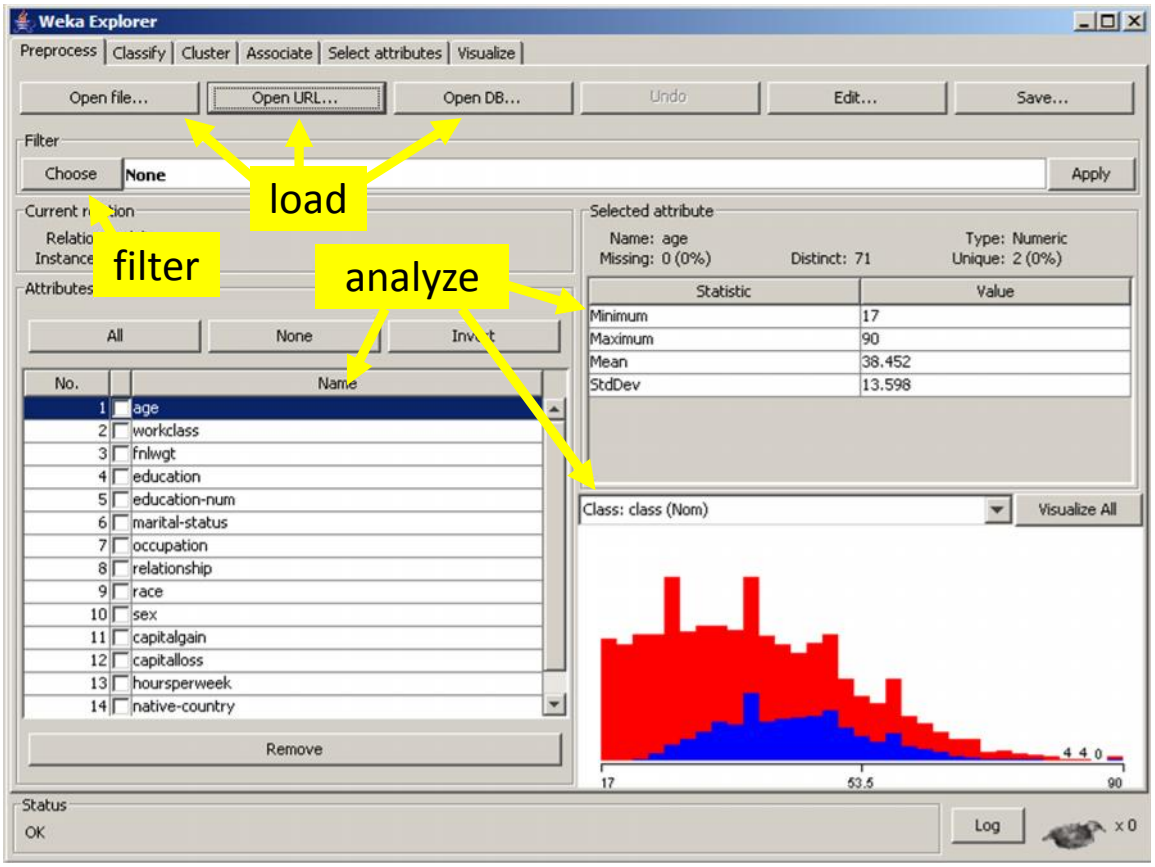
กรณี CVS file

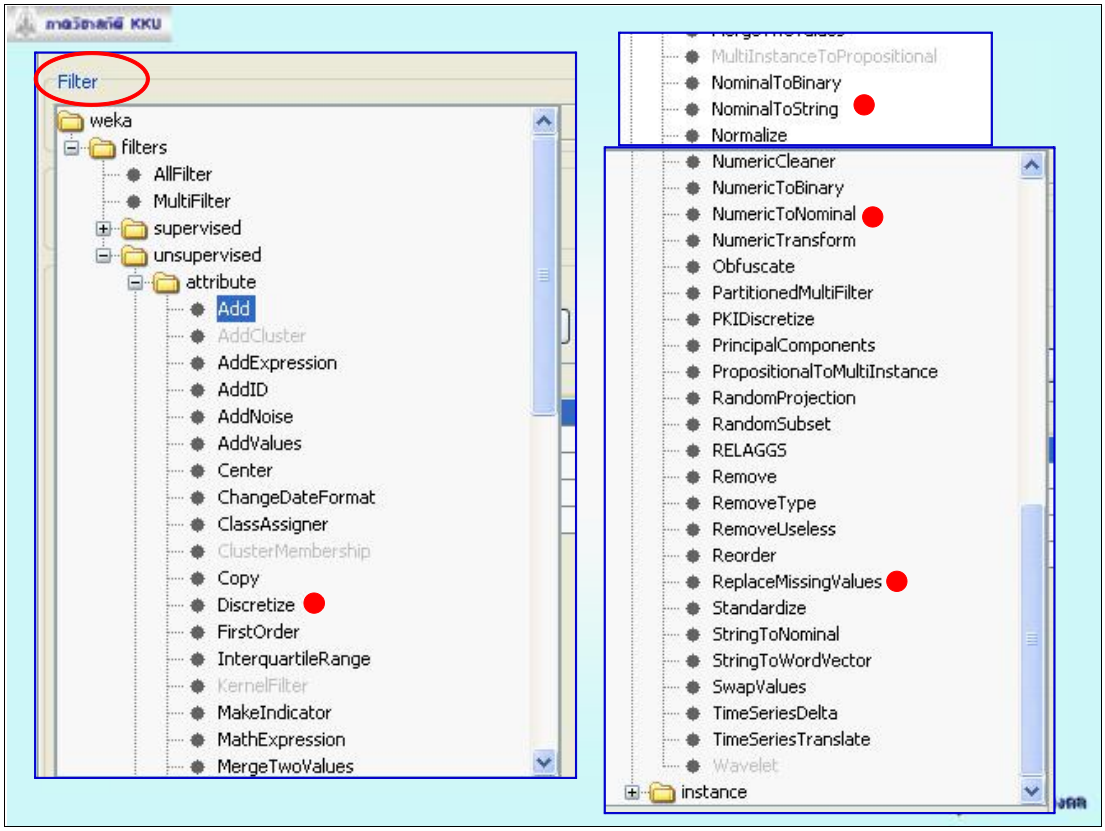


- สร้างแฟ้มแบบ CSV ด้วยโปรแกรม Microsoft Excel โดยบันทึกสกุลเป็น CSV

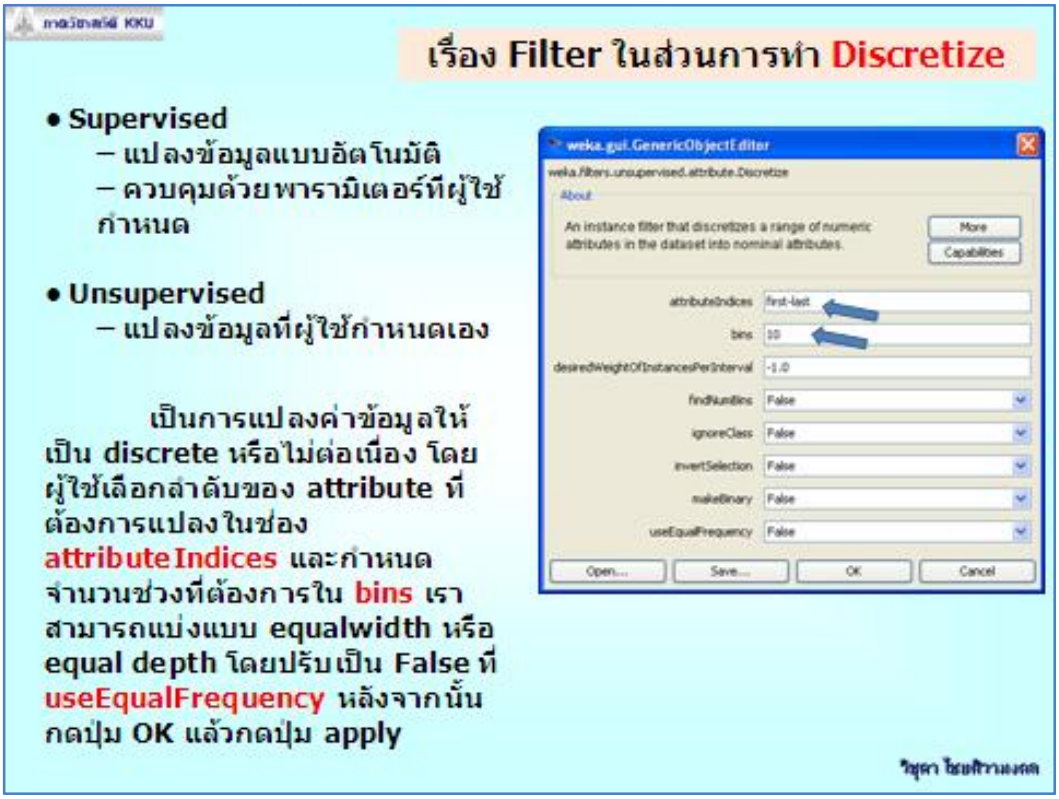
Data Preparation

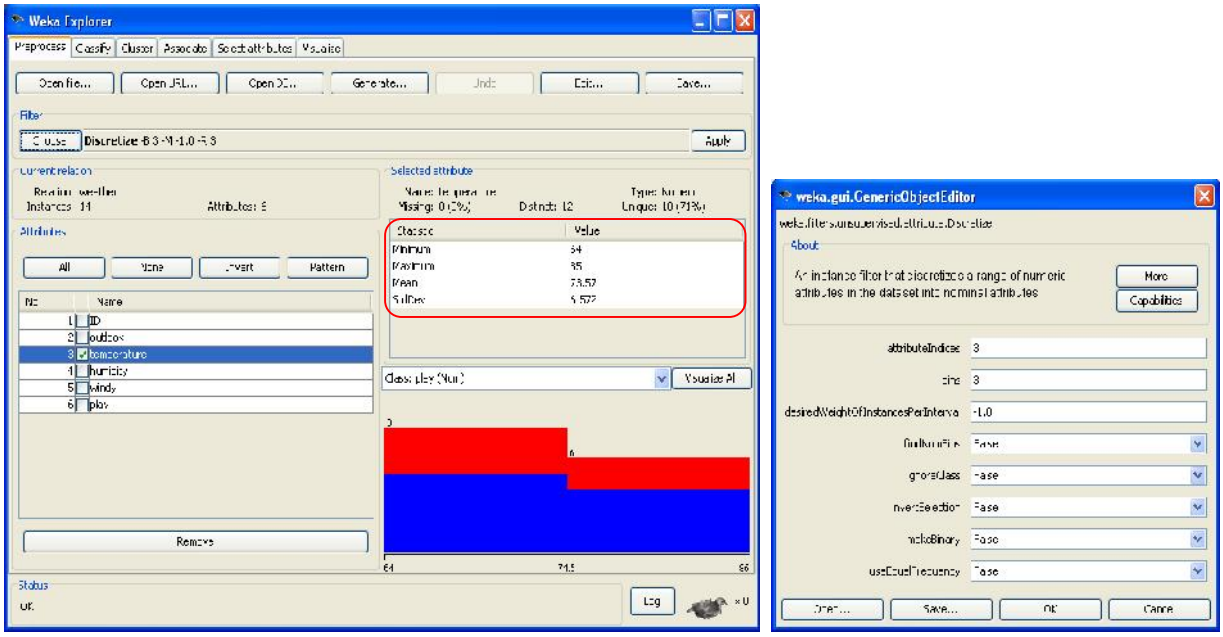


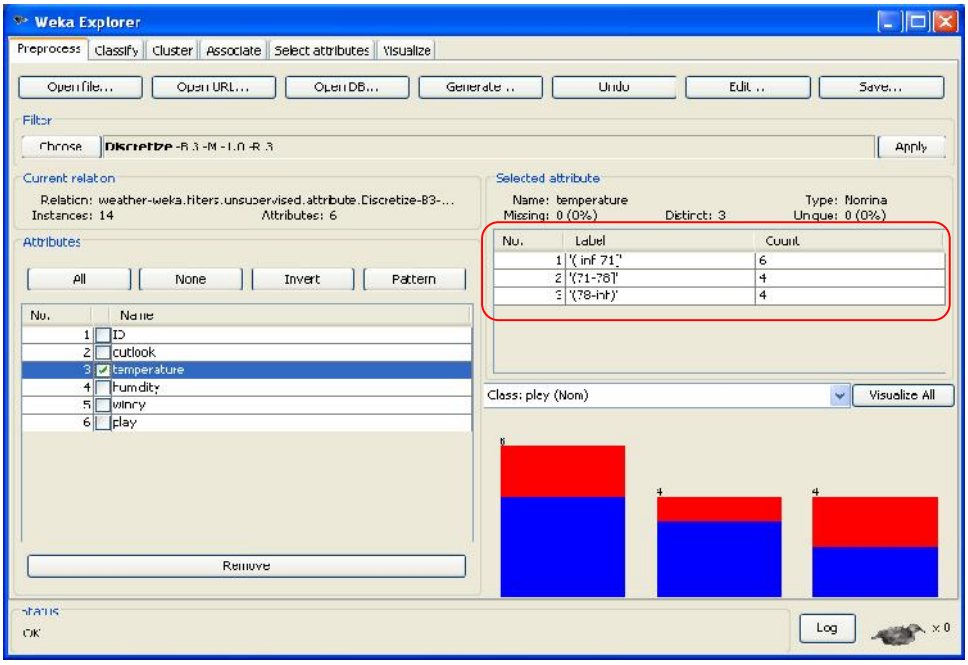


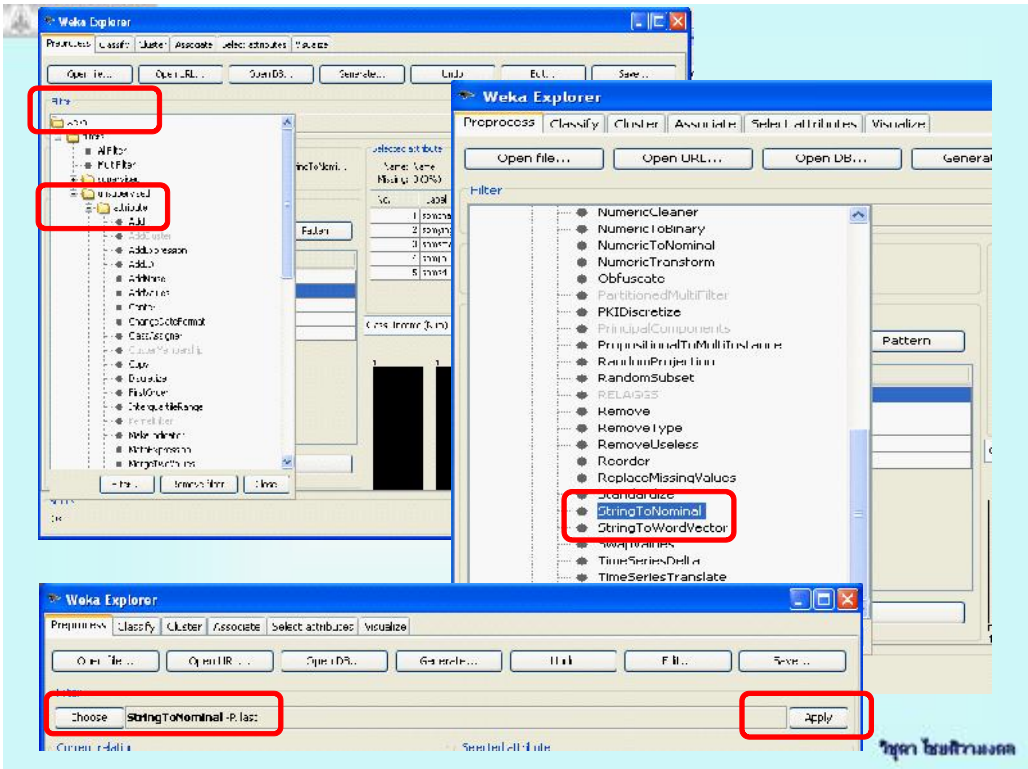


<u>ตัวอย่าง</u> ในการทำ Discredited เป็นการแปลงค่าข้อมูลให้เป็น discrete หรือไม่ต่อเนื่อง เช่น ค่า temperature



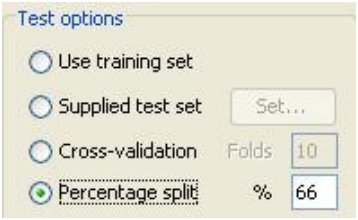

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Apply
Current relation
Selected attribute
Attributes
All
None
Invert
Pattern
Remove
Status
OK
Log
weka.gui.GenericObjectEditor
About
More
Capabilities
OK
Cancel

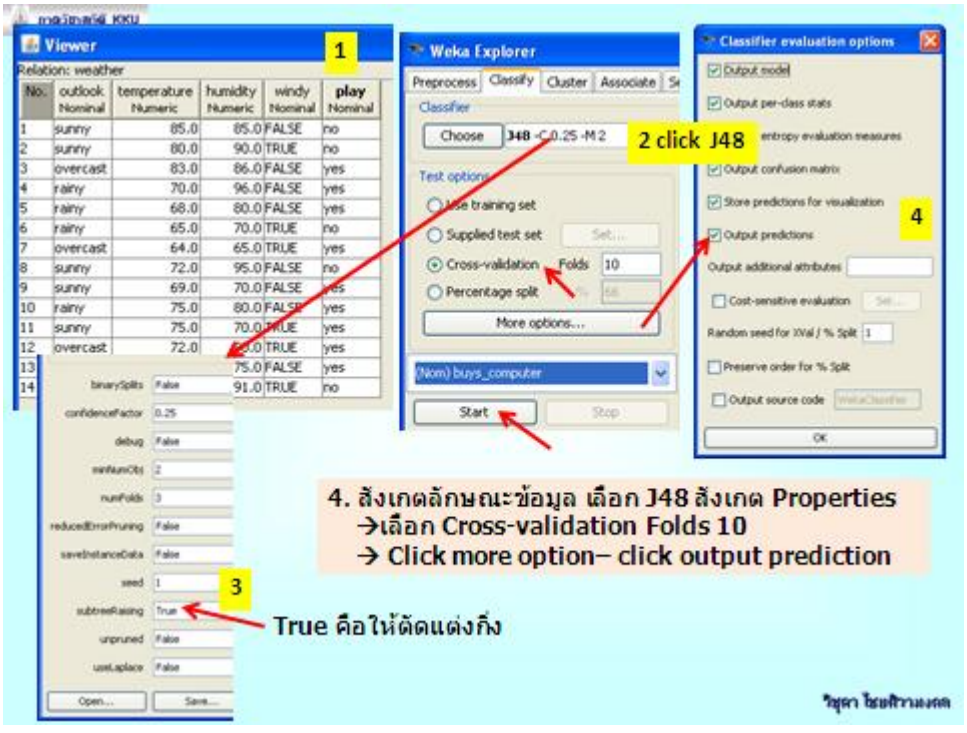

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 3 -M -1.0 -R 3
Apply
Current relation
Relation: weather-weka.filters.unsupervised.attribute.Discretize-B3-...
Instances: 14
Attributes: 6
Selected attribute
Name: temperature
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No. Label Count
1 '(-inf-71]' 6
2 '(71-78]' 4
3 '(78-inf)' 4
Attributes
All
None
Invert
Pattern
No. Name
1 ID
2 outlook
3 temperature
4 humidity
5 windy
6 play
Remove
Class: play (Nom)
Visualize All
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Filter
NumericCleaner
NumericToBinary
NumericToNominal
NumericTransform
Obfuscate
PKIDiscretize
Pattern
RandomSubset
Remove
RemoveType
RemoveUseless
Reorder
ReplaceMissingValues
StringToNominal
StringToWordVector
TimeSeriesTranslate
Choose
StringToNominal -R first
Apply

Supervised


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
1. Click Choose → ID3


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


Viewer
1
Weka Explorer
Classifier evaluation options
Choose
J48 -C 0.25 -M 2
2 click J48
Test options
Use training set
Supplied test set
Cross-validation
Folds
10
Percentage split
More options...
Start
Output model
Output per-class stats
Output confusion matrix
Store predictions for visualization
Output predictions
4
OK
4. สังเกตลักษณะข้อมูล เลือก J48 สังเกต Properties
→เลือก Cross-validation Folds 10
→ Click more option– click output prediction
3
True คือให้ตัดแต่งกิ่ง

Bayes → NaïveBayesSimple

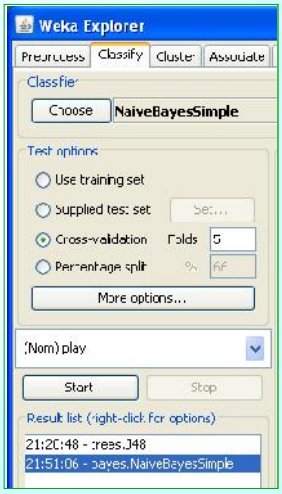
Naïve Bay

File 3weather.arff → Class Label คือ Play→ ข้อมูลเชิงคุณภาพ 2 level

Viewer

Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes



Correctly Classified Instances = 64.2857 %

Lazy → IBk

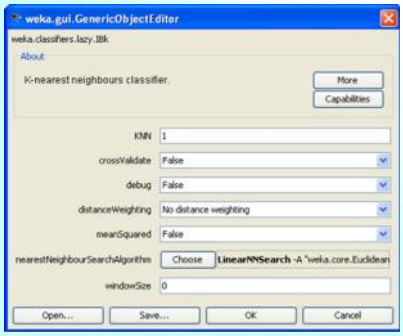
K-nearest neighbors

File 3weather.arff → Class Label คือ Play→ ข้อมูลเชิงคุณภาพ 2 level

Viewer

Relation: weather

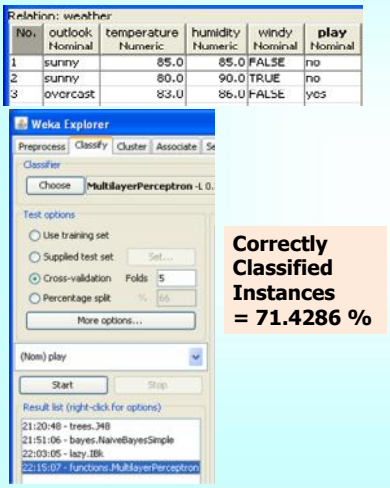
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no



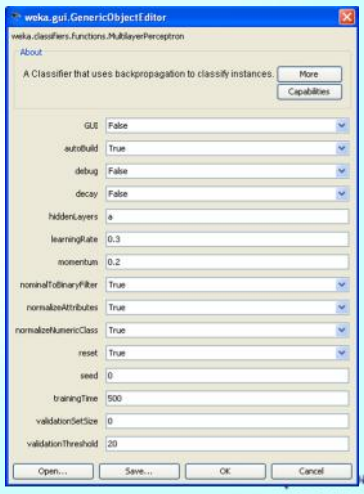
Correctly Classified Instances = 78.5714 %

Functions→ MultilayerPerceptron **Neural Networks**

File 3weather.arff → Class Label คือ Play→ ข้อมูลเชิงคุณภาพ 2 level



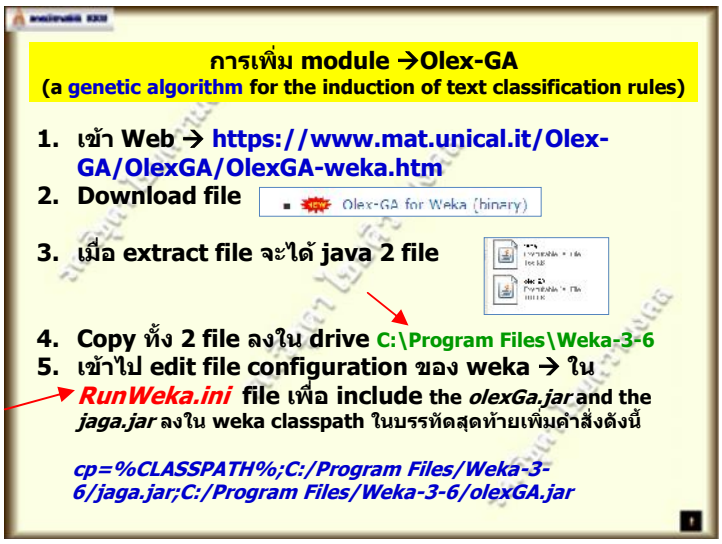
Correctly Classified Instances = 71.4286 %

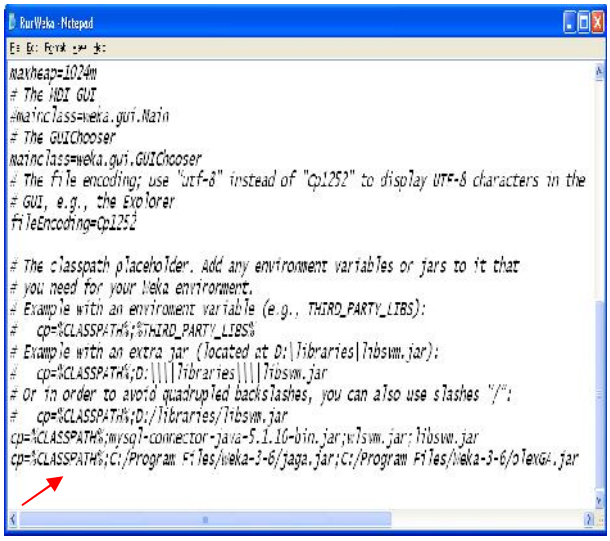


rules→ OlexGA **Genetic Algorithms**

File 3weather.arff → Class Label คือ Play→ ข้อมูลเชิงคุณภาพ 2 level

การเพิ่ม Module →Olex-GA



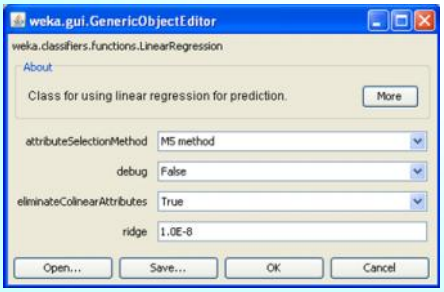


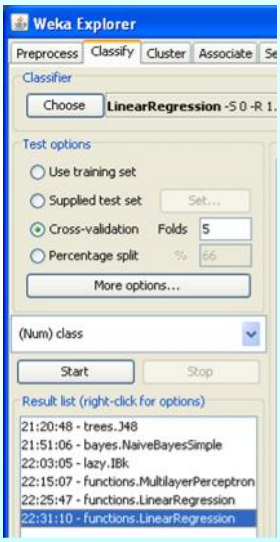
cp=%CLASSPATH%;C:/Program Files/Weka-3-6/jaga.jar;C:/Program Files/Weka-3-6/olexGA.jar

Functions→ LinearRegression

Linear Regression

File CPU→ Class ความเร็ว → ข้อมูลเชิงปริมาณ ซึ่งต้องการพยากรณ์



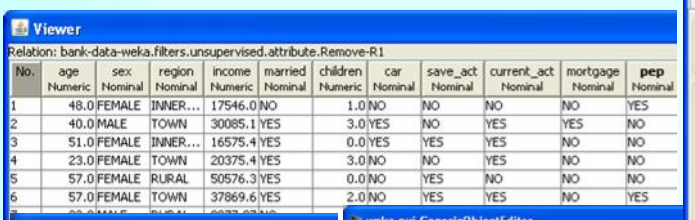


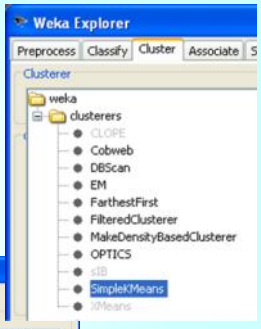
Unsupervised

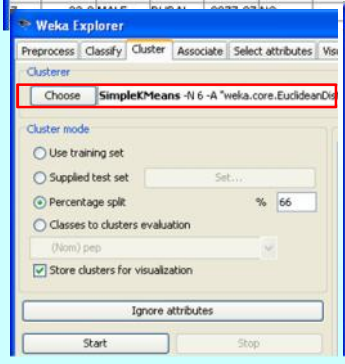
Cluster → SampleKMeans

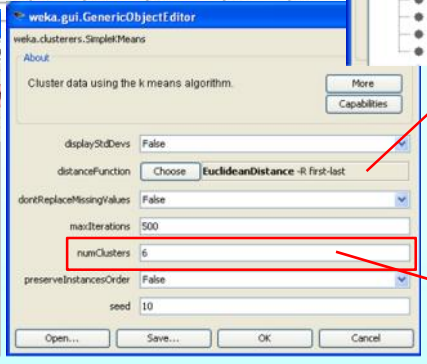
Cluster

File... 1bank_data.arff









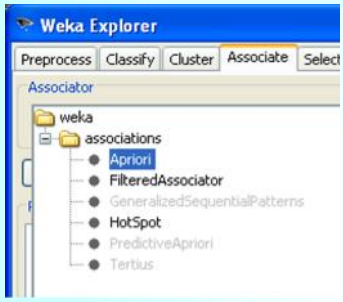
เทคนิคการวัดความห่าง

จำนวน Cluster ที่ต้องการ

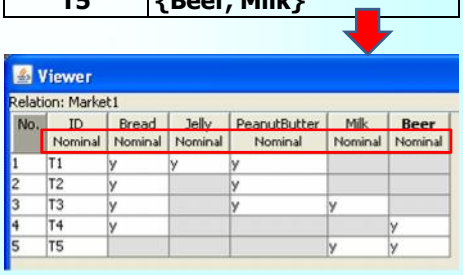
Associate→ Apriori

Association

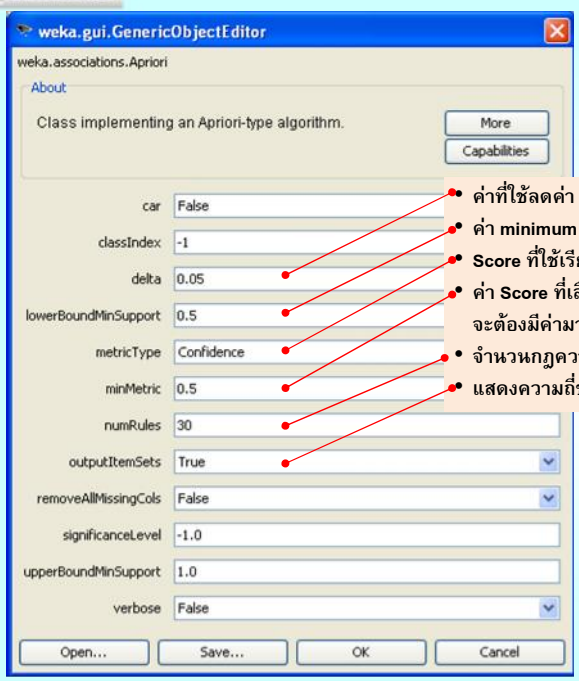
File.. 2market_data.arff



Trans_Id	Items
T1	{Bread, Jelly, PeanutButter}
T2	{Bread, PeanutButter}
T3	{Bread, Milk, PeanutButter}
T4	{Beer, Bread}
T5	{Beer, Milk}



No.	ID	Bread	Jelly	PeanutButter	Milk	Beer
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	T1	y	y	y		
2	T2	y		y		
3	T3	y		y	y	
4	T4	y				y
5	T5				y	y

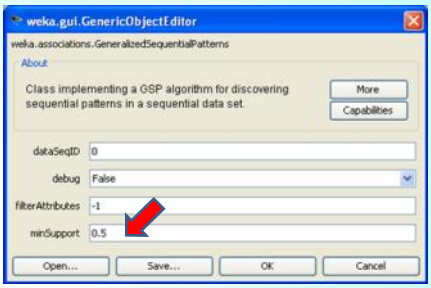


- ค่าที่ใช้ลดค่า minimum support
- ค่า minimum support
- Score ที่ใช้เรียง (rank) กฎความสัมพันธ์
- ค่า Score ที่เลือกใน metric Type โดยกฎที่สนใจจะต้องมีค่ามากกว่าที่กำหนด
- จำนวนกฎความสัมพันธ์ที่ต้องการ
- แสดงความถี่ของสินค้าที่ต้องการซื้อพร้อมกัน

Associate→ GeneralizedSequentialPatterns

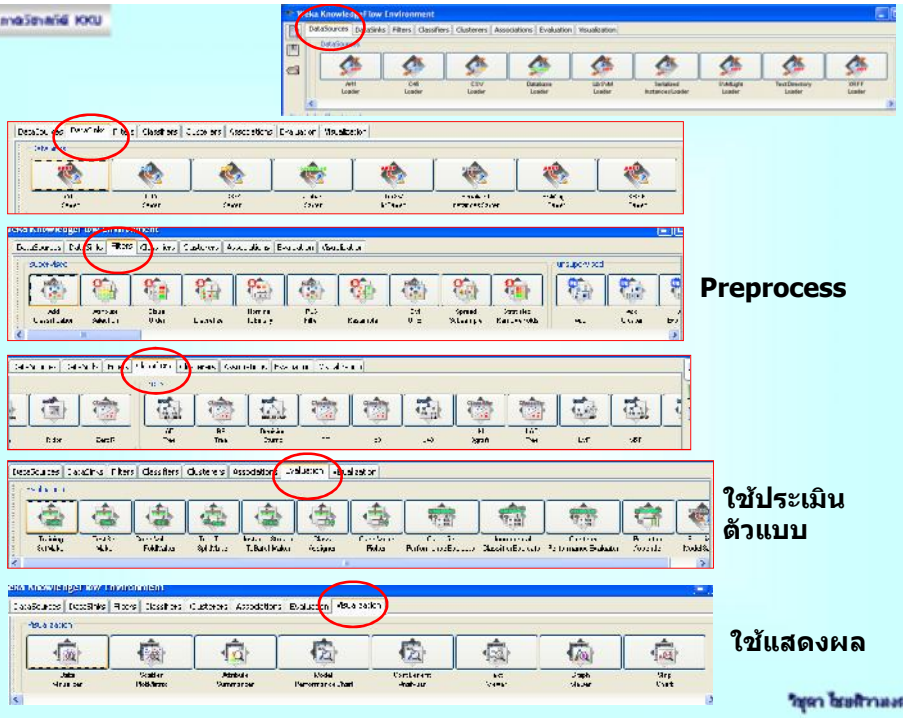
File Asso_Sequential2.csv

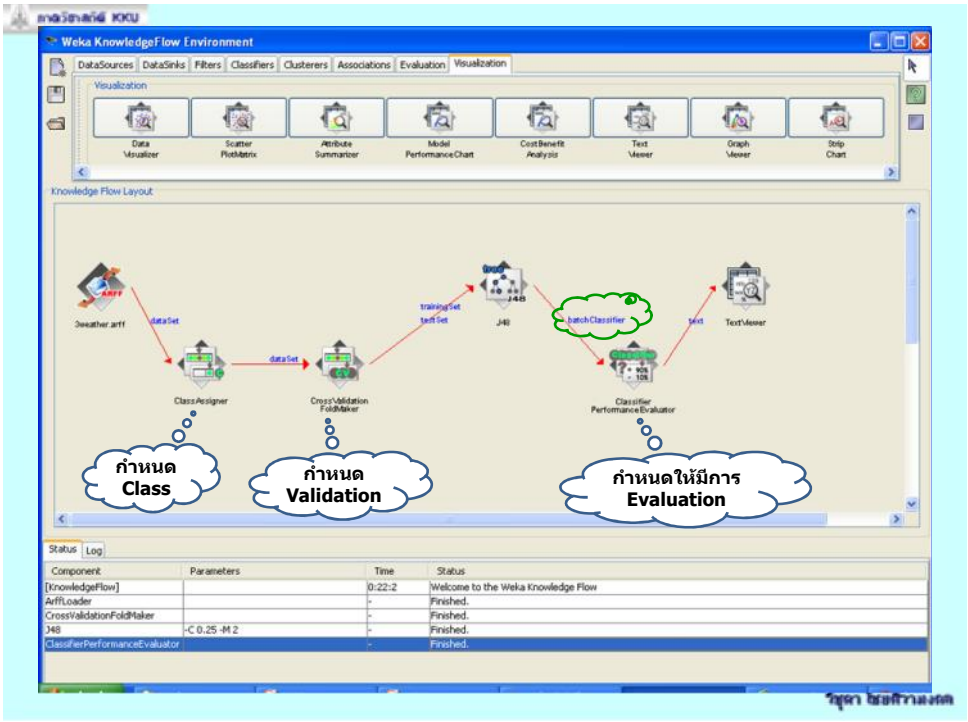
	A	B	C
1	ID	time	type
2	R1	1	fruit
3	R1	2	cloth
4	R2	1	shoes
5	R2	2	cloth
6	R2	3	fruit
7	R3	1	pen
8	R3	2	cloth
9	R4	1	shoes
10	R4	2	cloth
11	R5	1	cloth
12	R5	2	fruit
13	R5	3	pen
14			

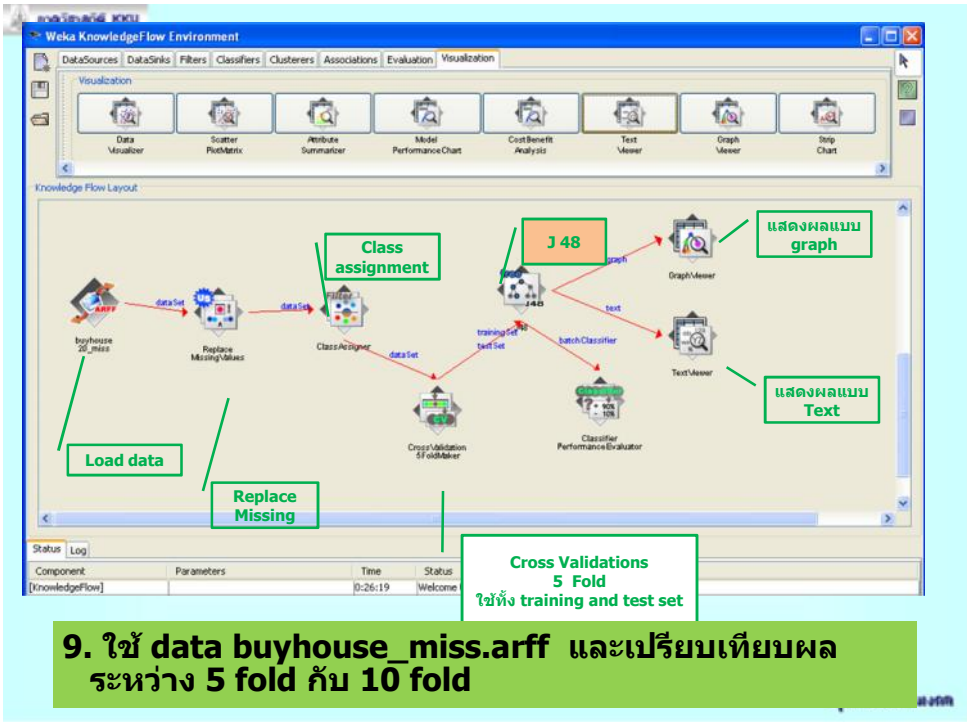


ต้องประกอบด้วย ID และ Time_Stamp

KnowlageFlow







9. ใช้ data buyhouse_miss.arff และเปรียบเทียบผล ระหว่าง 5 fold กับ 10 fold