คู่มือการใช้งาน WEKA

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยรามคำแหง

# การทำเหมืองข้อมูลด้วย Weka (Data Mining with Weka)

# การติดตั้ง Weka

- 1. Download โปรแกรมที่ <u>http://www.cs.waikato.ac.nz/ml/weka/downloading.html</u>
- ดัลเบิลคลิกเพื่อติดตั้ง

ในที่นี้จะใช้ version 3.7.11 (Book version 3.6)

ติดตั้งตาม default และ Next

🕞 Weka 3.7.11 Setup	
	Welcome to the Weka 3.7.11 Setup Wizard
	This wizard will guide you through the installation of Weka 3.7.11.
	It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.
51	Click Next to continue.
	Next > Cancel

จนถึง

🕞 Weka 3.7.11 Setup	
Please wait while Weka 3.7.11 is being inst	alled.
Extract: UnivariateEqualFrequencyHistogramEstimator.html 100%	
Extract: KernelEstimator.html 100% Extract: MahalanobisEstimator.html 100% Extract: MultivariateEstimator.html 100% Extract: MultivariateGaussianEstimator.html 100% Extract: NDConditionalEstimator.html 100% Extract: NNConditionalEstimator.html 100% Extract: NormalEstimator.html 100% Extract: PoissonEstimator.html 100% Extract: UnivariateDensityEstimator.html 100% Extract: UnivariateDensityEstimator.html 100%	•
Nullsoft Install System v08-Mar-2013,cvs — <u>Back N</u> ext	> Cancel

### 3. Install JRE

Java Setup - Welcome	
Java <sup>r</sup>	ORACLE
Welcome to Java	
Java provides safe and secure access to the world From business solutions to helpful utilities and ent your internet experience come	l of amazing Java content. ertainment, Java makes to life.
Note: No personal information is gathered as pa Click here for more information on what	art of our install process. we do collect.
Click Install to accept the license agreement	and install Java now.
Change destination folder	Cancel

กดปุ่ม Install

# 4. ติดตั้งเรียบร้อย

💮 Weka 3.7.11 Setup								
Weka "	stallation Complete Setup was completed successfully.							
Completed								
Output folder: C: \Program Files\Weka-3-7 Execute: RunJREInstaller.bat Delete file: C: \Program Files\Weka-3-7\RunJREInstaller.bat Created uninstaller: C: \Program Files\Weka-3-7\uninstall.exe Output folder: C: \Dsers\Admin \AppData\Roaming\Microsoft\Windows\Start Menu\Pr Create shortcut: C: \Dsers\Admin \AppData\Roaming\Microsoft\Windows\Start Menu\ Create shortcut: C: \Dsers\Admin \AppData\Roaming\Microsoft\Windows\Start Menu\								
Nullsoft Install System v08-Mar-20	013.cvs							
Weka 3.7.11 Setup	Completing the Weka 3.7.11 Setup Wizard Weka 3.7.11 has been installed on your computer. Click Finish to close this wizard.							
	< Back Finish Cancel							

## การเรียกใช้ Weka

1. All Programs > Weka 3.7



หรือ สำเนา short cut Weka 3.7 มาไว้บน desktop เพื่อสะดวกในการเรียกใช้



คลิก Do not show this message again แล้วคลิก OK

#### Weka GUI Chooser



## มี Interface การทำงาน 4 แบบ

### 1. Explorer

O Webs Eveleses		
Wexa Explorer		
Preprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB Gene	ate Undo	Edit Save
Chasse		Apply
Choise None		Арру
Current relation	Selected attribute	Type: None
Instances: None Sum of weights: None	Missing: None Distinct:	None Unique: None
Attributes		
All None Invert Pattern		
		✓ Visualize All
Remove		
Status		
Welcome to the Weka Explorer		Log x 0

#### 2. Experimenter

Weka Experiment Environment			
Setup Run Analyse			
Experiment Configuration Mode:		Simple	<u>A</u> dvanced
Open	5	ave	New
Results Destination			
ARFF file 🔻 Filename:			Browse
Experiment Type		Iteration Control	
Cross-validation	-	Number of repetitions:	
Number of folds:		O Data sets first	
Classification     Classification     Regression		<ul> <li>Algorithms first</li> </ul>	
Datasets		Algorithms	
Add new Edit selected	Delete selected	Add new	Edit selected Delete selected
Use relative paths			
	Down	Load options	Save options Up Down
	N	otes	

### 3. KnowledgeFlow

Construction     C	📿 Weka KnowledgeFlow Enviro	onment														• <b>×</b>	
Design Design Design Design Constructs DataSinks Defilters DetaSinks Defilters Defilters DetaSinks Defilters Defilte	Data mining processes																
DetaSources       B- DataSources       B- DataSources       B- DataSources       B- DataSources       B- DataSources       B- Classifiers       B- Classifiers       B- DataSources       B- DataSources       B- DataSources       B- Classifiers       B- Classifiers       B- Classifiers       B- DataSources       B- DataSour						Q		•	κ 🖿	•	2		<b>-</b>	i.	🗟 🖽	ی 🎲 🧯	D
Balasources	Design	Untitled1 ×															
Bit D GasSurces         Bit D GasSurces         Bit D GasSurces         Bit D Classflers         Bit D Classflers <td></td> <td>ĩ</td>																	ĩ
Pites       Classifiers         Bit Classifiers       Classifiers         Bit Classifiers       Classifiers         Bit Classifiers       File	DataSources     DataSinks																
Component Parameters Time Status RoowledgeFlow]	Filters																
B A Sociations B A Sociations B E Voluation B Voluation B Tools B Flow Status Log Component Parameters Time Status (frowledgeFlow) - Welcome to the Wela Knowledge Flow	Classifiers																
Produzitation Production Pro	Clusterers     Associations															=	
Vsualization P-Tools	Evaluation																
Piov Piov Piov Status Status Component Parameters Time Status Component Parameters Time Status Parameters Time Status Pov Welcome to the Weka Knowledge Flow	Visualization																
	Elow																
III       >         Status       Log         Component       Parameters         Time       Status         [VnowledgeFlow]       -         Welcome to the Weka Knowledge Flow																	
III       >         Status       Jog         Component       Parameters         Time       Status         [YnowledgeFlow]       -         Welcome to the Weka Knowledge Flow																	
III       +         Status       Log         Component       Parameters         Time       Status         [NnowledgeFlow]       -         Welcome to the Weka Knowledge Flow																	
Status Log Component Parameters Time Status [InowledgeFlow] - Welcome to the Weka Knowledge Flow																	
Status Log Component Parameters Time Status [hnowledgeFlow] - Welcome to the Weka Knowledge Flow																	
III     >       Status Log     >       Component     Parameters     Time       Status     Image: Component     Parameters       [NnowledgeFlow]     -     Welcome to the Weka Knowledge Flow																	
Status Log Component Parameters Time Status [VnowledgeFlow] - Welcome to the Weka Knowledge Flow																	
III     Status Log Component Parameters Time Status [KnowledgeFlow] - Welcome to the Weka Knowledge Flow																	
Status Log Component Parameters Time Status [InowledgeFlow] - Welcome to the Weka Knowledge Flow																	
Status Log Component Parameters Time Status [KnowledgeFlow] - Welcome to the Weka Knowledge Flow																	
Status     Log       Component     Parameters       Time     Status       [VnowledgeFlow]     -       Welcome to the Weka Knowledge Flow					1											-	
Status         Log           Component         Parameters         Time         Status           [NnowledgeFlow]         -         Welcome to the Weka Knowledge Flow			III													,	1
Component Parameters Time Status [[inowledgeFlow] - Welcome to the Weka Knowledge Flow		Status Log															
[[[CnowledgeFlow]]  -  Welcome to the Weka Knowledge Flow		Component	Parameter	s		Time	St	tatus									
		[KnowledgeFlow	1			-	We	elcome to	o the We	ka Knowl	edge Flo	W					I

#### 4. Simple CLI

```
SimpleCLI
                                                                    Welcome to the WEKA SimpleCLI
                                                                                .
Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.
> help
                                                                                Ξ
Command must be one of:
          java <classname> <args> [ > file]
          break
          kill
          capabilities <classname> <args>
          cls
          history
          exit
          help <command>
```

## Weka Workshops

### **WS#1: Numeric Prediction**

1 %

Data set: cpu (Weka\_data จะมีไฟล์ format เป็น .arff)

2 % As used by Kilpatrick, D. & Cameron-Jones, M. (1998). Numeric prediction 3 % using instance-based learning with encoding length selection. In Progress % in <u>Connectionist</u>-Based Information Systems. Singapore: Springer-<u>Verlag</u>. 4 5 6 % Deleted "vendor" attribute to make data consistent with with what we 7 % used in the data mining book. 읗 8 9 @relation 'cpu' 10 @attribute MYCT numeric 11 @attribute MMIN numeric 12 @attribute MMAX numeric 13 @attribute CACH numeric 14 @attribute CHMIN numeric 15 @attribute CHMAX numeric 16 @attribute class numeric 17 (data 18 125,256,6000,256,16,128,198 19 29,8000,32000,32,8,32,269 20 29,8000,32000,32,8,32,220 29,8000,32000,32,8,32,172 21 22 29,8000,16000,32,8,16,132 23 26 8000 32000 64 8 32 318

1.	Preprocess > Open file	เลอก cpu
	C Ones	

Open			
Look in:	🕕 Weka_dat	3	- 🦻 📂 🖽 📾
<b>S</b>	<ul> <li>breast-car</li> <li>contact-le</li> </ul>	ncer Segment-test nses Soybean	Invoke options dialog
Recent Items	<ul> <li>cpu.with.v</li> <li>credit-g</li> <li>diabetes</li> </ul>	vendor vote	Note: Some file formats offer additional options which can be customized
Desktop	glass ionospher	e weather.numeric	when invoking the options dialog.
Mu Deservente	iris.2D		
	<ul> <li>ReutersCo</li> <li>ReutersCo</li> </ul>	rn-test rn-train	
Computer	<ul> <li>ReutersGr.</li> <li>ReutersGr.</li> <li>segment-</li> </ul>	sin-test sin-train -hallenge	
	File name:	cpu.arff	 
Network	Files of type:	Arff data files (*.arff)	

ได้

🖉 Weka Explorer		
Preprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB Gener	rate Undo	Edit Save
Choose None		Apply
Current relation Relation: cpu Attributes: 7 Instances: 209 Sum of weights: 209	Selected attribute Name: MYCT Missing: 0 (0%) Distinct:	Type: Numeric 60 Unique: 19 (9%)
Attributes All None Invert Pattern	Statistic Minimum Maximum	Value 17 1500
	Mean	203.823
1 MYCT 2 MMIN 3 MMAX 4 CACH 5 CHMIN	Class: class (Num)	▼ Visualize All
6 CHMAX 7 dass	34	
Remove	15 17 17	130 _2 58.5 1500
Status OK		Log 💉 X 0

 แอททริบิวต์ตัวสุดท้ายจะเป็นตัวแปร class หากต้องการให้ตัวแปรอื่นเป็นตัวแปร class ให้ Edit แล้ว คลิกขวาตัวแปรที่ต้องการ เลือก Attribute as class ตัวแปรนั้นก็จะมาอยู่เป็นตัวสุดท้าย และแสดงเป็น

## ตัวหนา

<u>ب</u>	iewer		-	-	_	-		x			
Relatio	on: cpu										
No.	1: MYCT	2: MMIN	3: MMAX	4: CACH	5: CHMIN	6: CHMAX	7: class				
	Numeric	Numeric	Numeric	Numeric	Nu	<u>.</u>					
123	1500.0	768.0	1000.0	0.0		Get mean					
124	1500.0	768.0	2000.0	0.0							
125	800.0	768.0	2000.0	0.0		Set all value	es to				
208	480.0	512.0	8000.0	32.0		Set missing	values to				
209	480.0	1000.0	4000.0	0.0		Replace val	ues with				
11	400.0	1000.0	3000.0	0.0		replace val	ues within				
12	400.0	512.0	3500.0	4.0		Rename att	ribute				
13	60.0	2000.0	8000.0	65.0		itenanie au	moutem				
14	50.0	4000.0	16000.0	65.0		Attribute as	s class				
15	350.0	64.0	64.0	0.0		Delete attri	bute				
25	320.0	128.0	6000.0	0.0		Delete etteil					
26	320.0	512.0	2000.0	4.0		Delete attri	outes				
27	320.0	256.0	6000.0	0.0		Sort data (a	scending)				
28	320.0	256.0	3000.0	4.0							
29	320.0	512.0	5000.0	4.0		Optimal co	lumn widt	th (current)			
30	320.0	256.0	5000.0	4.0		Optimal co	lumn widt	th (all)			
37	50.0	500.0	2000.0	8.0		optimar co	LOIO				
38	50.0	1000.0	4000.0	8.0	1.0	5.0	29.0				
39	50.0	2000.0	8000.0	8.0	1.0	5.0	71.0				
47	810.0	512.0	512.0	8.0	1.0	1.0	18.0				
48	810.0	1000.0	5000.0	0.0	1.0	1.0	20.0				
49	320.0	512.0	8000.0	4.0	1.0	5.0	40.0				
50	200.0	512.0	8000.0	8.0	1.0	8.0	62.0				
51	700.0	384.0	8000.0	0.0	1.0	1.0	24.0	-			
	Undo OK Cancel										

## <u>หมายเหตุ</u> จะใช้ว่า แอททริบิวต์ หรือ ตัวแปร ก็ได้

ถ้าต้องการได้ไฟล์ลง Excel ก็ ctrl+A (Select All) แล้วไป paste ลงใน Excel แล้วเพิ่มบรรทัดแรกเป็น ชื่อตัวแปร Weka จะสรุปค่า descriptive statistics ของแต่ละตัวแปรให้ พร้อมแสดงกราฟ (histogram) หรือคลิก
 Visualize All เพื่อด histogram ของตัวแปรทกตัวพร้อมกัน



4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ Visualize

Weka Explorer						
Preprocess Classify	Cluster Associate	Select attributes Visuali	ze			
Plot Matrix	МУСТ	MMIN	MMAX	CACH	CHMIN	CHMA)
class						
	teste kontra anti-					
СНМАХ			··· .	· . ·		
↓						
PlotSize: [105]				Fast scrolling (uses mo	re memory)	
PointSize: [1]			[	Update		
Jitter:				Select Attributes		
Colour: class (Num)			▼]	SubSample % : 10	00	
Class Colour						
6			578			1150
Status OK					L	.og 💉 x 0

### 5. Fit linear regressiom model

Classify > Choose Classifier = functions > Linear Regression

Test options = Use training ser (default)

Weka Explore	r	Weka Explorer	
Preprocess Class	sify Cluster Associate !	Preprocess Classify Cluster Associate Select attri	
Classifier		Classifier	
weka	rs	Choose LinearRegression -5 0 -R 1.0E-8	
📑 📑 🛄 bay	es	Test options	
📄 📄 🕌 func	tions	Use training set	
	GaussianProcesses LinearRegression	Supplied test set Set	
•	Logistic MultilayerPerceptron	Cross-validation Folds 10	
•	SGD	Percentage split % 66	
	SGDText SimpleLinearRegression	More options	
-	SimpleLogistic		
	SMO	(Num) class 🗸 🗸	
	<b>SMOreg</b> VotedPerceptron	Start Stop	
f ⊕… blazy f ⊕… bmeta		Result list (right-click for options)	
🔲 🖾 🛄 misc			

## กด Start จะได้ผลล*ั*พธ์

🔮 Weka Explorer	
Preprocess Classify Cluster Associate S	elect attributes Visualize
Classifier	
Choose LinearRegression -5 0 -R 1	.0E-8
Test options	Classifier output
Ose training set	Linear Regression Model
Supplied test set Set Cross-validation Folds 10	class =
Percentage split % 66	0.0491 * MYCT +
	0.0152 * MMIN +
More options	0.0056 * CACH +
	1.4599 * CHMAX +
(Num) class 🔹 👻	-56.075
Start Stop	Time taken to build model: 0.03 seconds
18:15:55 - functions.LinearRegression	=== Evaluation on training set ===
	Time taken to test model on training data: 0.02 seconds
	=== Summary ===
	Correlation coefficient 0.93
	Mean absolute error 37.9748
	Root mean squared error 58.9899
	Relative absolute error 39.592 %
	Root relative squared error 36.7663 %
Status	
ок	Log 💉 x0

- 6. การแปลผล
  - 6.1 Model: Class/PRP = -56.075 + 0.0491\*MYCT + 0.0152\*MMIN + 0.0056 \*MMAX

```
+ 0.6298*CACH + 1.4599*CHMAX
```

ตัวแปรใดที่ไม่เข้าในโมเดล?

```
Linear Regression Model

class =

0.0491 * MYCT +

0.0152 * MMIN +

0.0056 * MMAX +

0.6298 * CACH +

1.4599 * CHMAX +

-56.075
```

6.2 Evaluation: Correlation coefficient R=0.93 or R<sup>2</sup>=0.8649

ตัวแปร predictor 5 ตัว คือ MYCT, MMIN, MMAX, CACH, CHMAX สามารถอธิบายความ แปรปรวนของตัวแปร Class/PRP ได้ประมาณ 81%

Correlation coefficient 0.93

6.3 Deployment: เครื่องคอมพิวเตอร์เครื่องใหม่ 1 เครื่อง มีคุณสมบัติดังนี้ MYCT=200, MMIN=1000, MMAX=2000, CASH=0, CHMAX=64 จะมีประสิทธิภาพ (PRP) = 73.58

คลิกขวาที่ Result แล้ว Save model

 ให้ทดสอบทั้ง 3 test options จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของ โมเดลจะเหมือนกัน ซึ่งสอดคล้องกับ Stepwise Regression ใน SPSS และใน R แต่ผลลัพธ์ของ Evaluation จาก Correlation coefficient ของทั้ง 3 test mode คือ 1) Use training set, 2) Crossvalidation (10 folds), 3) Percentage split (66%) จะไม่เท่ากัน

### WS#2: Classification by Logistic Regression

#### Data set: diabetes

1. Preprocess > Open file เลือก diabetes

Weka Explorer		
Preprocess Classify Cluster Associate Select attributes Visualize     Open file     Open URL     Open DB     Gene	rate Undo	Edit Save
Filter Choose None		Apply
Current relation Relation: pima_diabetes Attributes: 9 Instances: 768 Sum of weights: 768	Selected attribute Name: preg Missing: 0 (0%) Distinct:	Type: Numeric 17 Unique: 2 (0%)
Attributes	Statistic	Value
	Minimum	0
All None Invert Pattern	Maximum	17
	Mean	3.845
No. Name	StdDev	3.37
3         pres           4         skin           5         insu           6         mass           7         ped           8         age           9         class	Class: class (Nom)	✓ Visualize All
Remove		24 <u>11_19</u> _2_1_1
Status OK		Log 🛷 X

2. Edit เพื่อดู type และ ค่าของแต่ละตัวแปร

🔊 Viewer						×				
Relation: pima_diabetes										
No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal	
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive	A
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative	
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive	
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative	
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive	
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative	
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive	
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative	
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive	
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive	
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative	
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive	
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative	
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive	
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive	
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive	
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive	
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive	
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative	
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive	
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative	
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_negative	
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_positive	
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive	
	Undo OK Cancel									

ตัวแปร class (dependent/target) มี type เป็น nominal มี 2 ค่า คือ tested negative/tested positive ตัวแปร classifier (independent/predictor) มี 8 ตัว มี type เป็น numeric ทั้งหมด ดู descriptive statistics ของแต่ละตัวแปรที่ Weka สรุปให้ พร้อมแสดงกราฟ (histogram) หรือคลิก
 Visualize All เพื่อดู histogram ของตัวแปรทุกตัวพร้อมกัน

🥥 Weka Explorer	
Preprocess Classify Cluster Associate Select attributes Visualize	
Open file Open ORL Open DB Gener	rate Undo Edit Save
Filter	
Choose None	Apply
Current relation	Selected attribute
Relation: pima_diabetes Attributes: 9	Name: preg Type: Numeric
Instances: 768 Sum of weights: 768	Missing: 0 (0%) Distinct: 1/ Unique: 2 (0%)
Attributes	Statistic Value
All None Invert Pattern	Minimum 0
	Maximum 17 Mean 3.945
No. Name	StdDev 3.37
1 preg	
2 plas	
4 skin	
5 minsu	Class: dass (Nom)
6 mass	
/ ped	246
9 dass	
	125
	103
	50 45
Remove	24 11 19 2 1 1
	0 85 17
Status	
OK	Log x0

ข้อมูลนี้มีทั้งหมด 768 ตัว (Instances/Cases/N)

ตัวแปร 1 ถึง 8 เป็นตัวแปร classifier มี type เป็น numeric ดังนั้น descriptive statistics จึงเป็น Minimum, Maximum, Mean, Standard Deviation (StdDev)

ตัวแปร class จะเป็นตัวแปรสุดท้ายเสมอ มี type เป็น nominal

ลองคลิก Visualize All จะเห็นกราฟความสัมพันธ์ของตัวแปร class กับตัวแปร classifier แต่ละตัว

- 4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ Visualize
- 5. Fit logistic regressiom model

เลือกแท็บ Classify > Choose Classifier = functions >Logistic (Regression) Default Test option = Cross validation Fold 10

🜍 Weka Explorer	🜍 Weka Explorer		
Preprocess Classify Cluster Associate Select	Preprocess Classify Cluster Associate		
Classifier	Classifier		
weka	Choose Logistic -R 1.0E-8 -M -1		
the bayes	Test options		
inctions	<ul> <li>Use training set</li> </ul>		
GaussianProcesses	Supplied test set Set		
Logistic     MultilaverPercentron	Cross-validation Folds 10		
SGD	Percentage split % 66		
SGDText	More options		
- SimpleLogistic			
( • SMO	(Nom) class		
SMOreg			
VotedPerceptron	Start Stop		
lazy	Result list (right-click for options)		
n meta			
🕒 📁 📁 rules			

## กด Start จะได้ผลล*ั*พธ์

## 6. การแปลผล

```
6.1 Model: Class = e<sup>-0.1232</sup>*preg + e<sup>-0.0352</sup>*plas + e<sup>0.0133</sup>*pres + e<sup>-0.0006</sup>*skin + e<sup>0.0012</sup>*insu +
e<sup>-0.0897</sup>*mass + e<sup>-0.9452</sup>*pedi + e<sup>-0.0149</sup>*age + 8.4047 ਅਤਿੰ
Class = 0.8841*preg + 0.9654*plas + 1.0134*pres + 0.9994*skin + 1.0012*insu +
0.9142*mass + 0.3886*pedi + 0.9852*age
```

Classifier output	t	Odds Ratios	
Logistic Re	egression with ridge	Variable	tested_negative
Coefficient	ts		
	Class	preg	0.8841
Variable	tested_negative	plas	0.9654
		pres	1.0134
preg	-0.1232	skin	0.9994
plas	-0.0352	insu	1.0012
pres	0.0133	mass	0.9142
skin	-0.0006	pedi	0.3886
insu	0.0012	age	0.9852
mass	-0.0897		
pedi	-0.9452		
age	-0.0149		
Intercept	8.4047		

### 6.2 Evaluation:

Correctly Classified Instances	593	77.2135 %	
Incorrectly Classified Instances	175	22.7865 %	

## 6.3 **Deployment:** แทนค่าลงในสมการข้อ 6.1

คลิกขวาที่ Result แล้ว Save model

7. ให้ทดสอบทั้ง 2 test options (Use training set, Cross-validation Folds 10 จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของโมเดลจะเหมือนกัน แต่ accuracy จะไม่เหมือนกัน ในวิธีการ Cross-validation Folds 10 Weka จะแบ่งข้อมูลทั้งหมดออกเป็น 10 ส่วน และรัน 10 ครั้ง โดยครั้งที่ 1 ใช้ ส่วนที่ 1 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set ครั้งที่ 2 ใช้ ส่วนที่ 2 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set โดย evaluation accuracy จะเป็นค่าเฉลี่ยของทั้ง 10 ครั้ง จากนั้นจะรันโมเดลอีกครั้งโดยใช้ full training set ทั้งหมดในการรันโมเดล

#### WS#3: Classification by ML algorithm

#### Data set: diabetes

- 1. Preprocess > Open file เลือก diabetes
- 2. Fit classification model Classifier = rules > ZeroR (baseline accuracy)

#### การแปลผล

#### Model: Class = tested\_negative

```
=== Classifier model (full training set) ===
ZeroR predicts class value: tested negative
```

#### Evaluation:

```
Correctly Classified Instances 500 65.1042 %
Incorrectly Classified Instances 268 34.8958 %
=== Confusion Matrix ===
a b <-- classified as
500 0 | a = tested_negative
268 0 | b = tested_positive
```

Deployment: tested negative

3. Fit classification model Classifier = rules > OneR

#### การแปลผล

Model: ไม่ make sence

Evaluation: (cross-validation)

Correctly Classified Instances	549	71.4844 %	
Incorrectly Classified Instances	219	28.5156 %	

```
a b <-- classified as
433 67 | a = tested_negative
152 116 | b = tested_positive</pre>
```

**Deployment:** ใช้กฎ (apply rule)

4. Fit classification model Classifier = trees > J48

<u>หมายเหตุ</u>ตาม default parameter จะได้จะได้ 20 กฏ ซึ่งบางกฏ ก็เป็นจริงสำหรับข้อมูลจำนวนน้อย ดังนั้นควรจะปรับ parameter "minNumObj" เช่นถ้าต้องการให้กฏเป็นจริงสำหรับข้อมูล 10% ก็ปรับ minNumObj = 77 โดยคลิกที่ **J48** แก้ minNumObj

(	Choose 348 -C 0.25 -M 77						
1	🔮 weka.gui.GenericObjectEditor						
	weka.dassifiers.trees.J48 About						
l	Class for generating a prun	ed or unpruned C4.					
	binarySplits	False =					
	collapseTree	True					
1	confidenceFactor	0.25					
1	debug	False					
	doNotCheckCapabilities	False					
l	doNotMakeSplitPointActualValue	False					
l	minNumObj	77					
l	numFolds	3					
	۰ III.	•					
	Open Save	OK Cancel					

### การแปลผล

Model: คลิกขวาที่ Result เลือก Visualize tree



Evaluation: (cross-validation)

<pre>=== Stratified cross-validation === === Summary ===</pre>				
Correctly Classified Instances	570	74.2188 %		
Incorrectly Classified Instances	198	25.7813 %		
=== Confusion Matrix ===				
a b < classified as				
445 55   a = tested_negative				
143 125   b = tested_positive				

Deployment: คลิกขวาที่ Result แล้ว Save model

 ให้เปรียบเทียบผลการทดลองด้วย Classifier ตัวอื่น ๆ เช่น rules > PART, rules > JRIP (จะดำเนินการใน Experimenter Interface)

## WS#4: Association Problem Type

## Data set: supermarket

1. Preprocess > Open file เลือก supermarket

## Edit เพื่อดูการจัดเก็บข้อมูล

-								
4	🛃 Viewer							
Re	Relation: supermarket							
at	20: canned fruit Nominal	21: canned vegetables Nominal	22: breakfast food Nominal	23: cigs-tobacco pkts Nominal				
		t		*				
	t	t						
	t							
		t	t					
		t	t					
	t	t	t					
		t						
	t	t	-					
	t	t	t					
	t	t						
			t	t				
			1					
			t					
	t •	t	t					
	L.							
			+	+				
			د 	L				
				+				
		t	t					
			-					
-	E							
<b>H</b>	E							
			Undo	OK Cancel				

แถวหนึ่ง ๆ คือตะกร้าสินค้าหนึ่ง ๆ t (=true) มีสินค้าชนิดนั้น

2. Fit Association model โดยเลือกแท็บ Associate

Preprocess Classify Cluster Associate	Preprocess Classify Cluster Associate Select attributes Visuali
Associator	Associator
weka	Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2 -S -1.
Apriori	🕼 weka.gui.GenericObjectEditor
FilteredAssociator	weka associations. Apriori
FPGrowth	About
	Class implementing an Apriori-type algorithm
เลือก Associator algorithm เป็น	Class implementing an option type agentation.
Apriori	
คลิกที่ Apriori (รูปทางขวา)	car False 🗉
ปรับ parameter ที่สำคัญ 3 ตัวคือ	dassIndex -1
-lowerBoundMinSupport = 0.2	delta 0.05
-minMetric (min Confidence) = 0.9	lowerBoundMinSupport 0.2
-numRules = 10	metricType Confidence
ผลลัพธ์: ไม่พบกฎใด ๆ เลย	
	minMetric 0.9
NDAT 1T	numRules 10
-minMetric (min Confidence) = 0.85	· · · · · · · · · · · · · · · · · · ·
	Open Save OK Cancel

### การแปลผล

#### Model:



Evaluation: Min Support และ Confidence

Deployment: เอา กฏ ไปใช้

### WS#5: Clustering

#### Data set: iris

- Preprocess > Open file เลือก iris
   ในที่นี้ iris มีตัวแปร class แต่เราจะไม่ใช้ในการ clustering แต่จะใช้ในการเปรียบเทียบกับค่าจากโมเดล
- 2. Fit Clustering model โดยเลือกแท็บ Cluster



Weka Explorer Preprocess Classify Cluster Associate Select a	attributes Visualize	คลิก SimpleKMeans
Choose SimpleKMeans -init 0 -max-candid	idates 100 -periodic-pruning 10000 -min-density	151 pumCluster = 3
weka.gui.GenericObjectEditor	<b>X</b>	
weka.clusterers.SimpleKMeans About Cluster data using the k means algorith		
	Capabi	
canopyMaxNumCanopiesToHoldInMemory 1	100	
canopyMinimumCanopyDensity 2	2.0	
canopyPeriodicPruningRate 1	10000	
canopyT1 -	-1.25	
canopyT2 -	-1.0	
debug [	False	
displayStdDevs F	False	
distanceFunction	Choose EuclideanDistance -R f	
doNotCheckCapabilities F	False	
dontReplaceMissingValues	False	
fastDistanceCalc F	False	
initializationMethod R	Random	
maxIterations 5	500	
numClusters 3	3	
<		
S Open Save	OK Cancel	

## การแปลผล

## Model:

Final cluster	centroids:			
		Cluster#		
Attribute	Full Data	0	1	2
	(150)	(61)	(50)	(39)
sepallength	5.8433	5.8885	5.006	6.8462
sepalwidth	3.054	2.7377	3.418	3.0821
petallength	3.7587	4.3967	1.464	5.7026
petalwidth	1.1987	1.418	0.244	2.0795

#### Evaluation:

Within cluster sum of squared errors: 6.998114004826762

Deployment: คลิกขวาที่ Result แล้ว Save model

## WS#6 จงเปรียบเทียบการวิเคราะห์โจทย์ Classification โดยใช้หลาย algoithms

## Experiment 1

#### Data set: iris

Algoritms: trees > J48, rules > OneR, rules > ZeroR

1. Weka GUI Chooser เลือก Experimenter

Weka Experiment Environment		and the second second		
Setup Run Analyse				
Experiment Configuration Mode:		Simple	Advance	ed
		ave		New
Results Destination				
ARFF file v Filename:				Browse
Experiment Type		Iteration Control		
Cross-validation	~	Number of repetitions:		
Number of folds:		Oata sets first		
Classification     Classification     Regression	n	<ul> <li>Algorithms first</li> </ul>		
Datasets		Algorithms		
Add new Edit selected	Delete selected	Add new	Edit selected	Delete selected
Use relative paths				
Up	Down	Load options	Save options	Up Down
	No	tes		
	110			

## คลิก New

Weka Experiment Environment	the local balls	-	A country of	
Setup Run Analyse				
Experiment Configuration Mode:		Simple	<u>A</u> dvanced	ł
Open	Sa	ve	N	ew
Results Destination				
ARFF file				Browse
Experiment Type		Iteration Control		
Cross-validation	•	Number of repetitions: 10		
Number of folds: 10		<ul> <li>Data sets first</li> </ul>		
Classification     Classification		Algorithms first		
Datasets		Algorithms		
Add new Edit selected	Delete selected	Add new	Edit selected	Delete selected
Use relative paths				
Up	Down	Load options	Save options	Up Down
Notes				
<u></u>				

3. ที่ Datasets คลิก Add new แล้วเปิดเลือก iris

ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)

G Weka Experiment Environment	
Setup Run Analyse	
Experiment Configuration Mode:	<u>Simple</u>
<u>O</u> pen <u>S</u>	ave <u>N</u> ew
Results Destination	
ARFF file   Filename:	Browse
Experiment Type	Iteration Control
Cross-validation 👻	Number of repetitions: 10
Number of folds: 10	O Data sets first
Classification     Classification	<ul> <li>Algorithms first</li> </ul>
Datasets	Algorithms
Add ne Edit s Delete	Add Edit sel Delete s
🔲 Use relati	J48 -C 0.25 -M 2
C:\Program Files\Weka-3-7\data\jris.arff	OneR -B 6
Up Down	Load op Save op
No	otes

## 4. ไปที่แท็บ Run คลิก Start

Weka Experiment Environment		🔮 Weka Experiment Environment
Setup Run Analyse		Setup Run Analyse
Start	Stop	Stort Stop
Log		Log
		18:48:02: Started
		18:48:02: Finished
		18:48:02: There were 0 errors
Status		Status
Not running		Not running

5. ไปที่แท็บ Analyze คลิก Experiment

Weka Experiment Er	nvironment	
Setup Run Analyse		
Source		
Got 300 results		<u>Fi</u> le <u>D</u> atabase <u>Experiment</u>
Configure test		Test output
Testing <u>w</u> ith	Paired T-Tester (corrected)	Available resultsets
Select <u>r</u> ows and cols	Rows Cols Swap	<pre>(1) trees.J48 '-C 0.25 -M 2' -21773316839364444 (2) rules.OneR '-B 6' -3459427003147861443 (3) rules.ZeroR '' 48055541465867954</pre>
Comparison field	Percent_correct	
Significance	0.05	
Sorting (asc.) by	<default></default>	
Test <u>b</u> ase	Select	
Displayed Columns	Select	
Show std. devi <u>a</u> tions		
Output Format	Select	
Perform <u>t</u> est Result list	Save output	
		راد ــــــــــــــــــــــــــــــــــــ

### 6. คลิก Perform test

🕝 Weka Experiment E	nvironment		
Setup Run Analyse			
Source			
Got 300 results			Eile Database Experiment
Configure test		Test output	
Testing <u>w</u> ith	Paired T-Tester (corrected)	Tester:	weka.experiment.PairedCorrectedTTester
Select <u>r</u> ows and cols	Rows Cols Swap	Analysing: Datasets: Resultsets:	Percent_correct 1 3
Comparison field	Percent_correct	Confidence:	0.05 (two tailed)
Significance	0.05	Date:	26/12/2557, 19:01 u.
Sorting (asc.) by	<default></default>	Dataset	(1) trees.J4   (2) rules (3) rules
Test <u>b</u> ase	Select	iris	(100) 94.73   92.53 33.33 *
D <u>i</u> splayed Columns	Select		(v/ /*)   (0/1/0) (0/0/1)
Show std. deviations			
<u>O</u> utput Format	Select	Key: (1) trees.J (2) rules.O	48 '-C 0.25 -M 2' -217733168393644444 neR '-B 6' -3459427003147861443
Perform <u>t</u> est Result list	Save output	(3) rules.Z	eroR '' 48055541465867954

7. การแปลผลจาก Witten

Dataset	<li>(1) trees.J4</li>	(2) rules (3) rules
iris	(100) 94.73	92.53 33.33 *
	(∇/ /*)	(0/1/0) (0/0/1)
Key: (1) trees.J48		
(2) rules.OneR		
(3) rules.ZeroR		

- v significantly better
- \* significantly worse

- ZeroR (33.3%) is significantly worse than J48 (94.7%)
- Cannot be sure that OneR (92.5%) is significantly worse than J48
- ✤ ... at the 5% level of statistical significance
- ↔ J48 seems better than ZeroR: pretty sure (5% level) that this is not due to chance

#### Experiment 2:

Data set: iris, breast-cancer, glass, ionosphere, segment-challenge

Algoritms: trees > J48, rules > OneR, rules > ZeroR

- 1. ไปที่แท็บ Set up คลิก New
- ที่ Datasets คลิก Add new แล้วเปิดเลือก iris, breast-cancer, glass, ionosphere, segmentchallenge (Add new 5 ครั้ง)

ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)

Weka Experiment Environment		
Setup Run Analyse		
Experiment Configuration Mode:	Simple	Advanced
Open	Save	New
Results Destination		
ARFF file   Filename:		Browse
Experiment Type	Iteration Control	
Cross-validation	<ul> <li>Number of repetition</li> </ul>	ns: 10
Number of folds: 10	Oata sets first	
Classification     Classification	on 💿 Algorithms first	
Datasets	Algorithms	
Add new Edit sele	Delete s Add n	Edit select Delete sele
Use relative	J48 -C 0.25 -M 2	
C:\Program Files\Weka-3-7\data\iris.arff	OneR -B 6	
C:\Program Files\Weka-3-7\data\breast-ca	incer.arff	
C: \Program Files \Weka-3-7\data \glass.arti	re.arff	
C:\Program Files\Weka-3-7\data\segment	challenge.arff	
Up De	wn Load opti	Save opti
	Notes	

- 3. ไปที่แท็บ Run คลิก Start
- 4. ไปที่แท็บ Analyze คลิก Experiment

## 5. คลิก Perform test

🕝 Weka Experiment Er	nvironment	
Setup Run Analyse		
Source		
Got 1500 results		File Database Experiment
Configure test		Test output
Testing with	Paired T-Tester (corrected)	Tester: weka.experiment.PairedCorrectedTTester Analysing: Percent correct
Select rows and cols	Rows Cols Swap	Datasets: 5 Resultsets: 3
Comparison field	Percent_correct	Confidence: 0.05 (two tailed) Sorted by: -
Significance	0.05	Date: 26/12/2557, 19:26 u.
Sorting (asc.) by	<default></default>	Dataset (1) trees.J4   (2) rules (3) rules
Test base	Select	iris (100) 94.73   92.53 33.33 *
Displayed Columns	Select	breast-cancer (100) 74.28   66.91 * 70.30 Glass (100) 67.63   57.40 * 35.51 *
Show std. deviations		ionosphere (100) 89.74   82.28 * 64.10 * segment (100) 95.71   64.35 * 15.73 *
Output Format	Select	(v/ /*)   (0/1/4) (0/1/4)
Perform test	Save output	Key:
Result list		(1) trees.J48 '-C 0.25 -M 2' -217/33168393644444 (2) rules.OneR '-R 6' -3459427003147861443
19:00:39 - Available res	ultsets	(3) rules.ZeroR '' 48055541465867954
19:01:51 - Percent_corr 19:26:06 - Available res	ect - trees.J48 -C 0.25 -M 2 -21/73316839 ultsets	
19:26:35 - Percent_corr	rect - trees.J48 '-C 0.25 -M 2' -21773316839	
	4	

6. เปลี่ยน **Test base** Select เป็น rules.OneR

Weka Experiment E	nvironment			
Setup Run Analyse				
Source				
Got 1500 results			Eile	<u>E</u> xperiment
Configure test		Test output		
Testing <u>w</u> ith	Paired T-Tester (corrected)	Tester: weka.expe Analysing: Percent_c	riment.PairedCorrectedTTester orrect	<u>^</u>
Select <u>r</u> ows and cols	Rows Cols Swap	Datasets: 5 Resultsets: 3		
Comparison field	Percent_correct 🔹	Confidence: 0.05 (two tailed) Sorted by: -		
Significance	0.05	Date: 26/12/2557, 19:50 %.		
Sorting (asc.) by	<default></default>	Dataset	(2) rules.On   (1) trees (3	3) rules =
Test <u>b</u> ase	Select	iris	(100) 92.53   94.73	33.33 *
D <u>i</u> splayed Columns	Select	breast-cancer Glass	(100) 66.91   74.28 v (100) 57.40   67.63 v	70.30 35.51 *
Show std. devi <u>a</u> tions		ionosphere segment	(100) 82.28   89.74 v (100) 64.35   95.71 v	64.10 * 15.73 *
<u>O</u> utput Format	Select		(v/ /*)   (4/1/0)	(0/1/4)
Perform <u>t</u> est	Save output	Key:		
Result list		(1) trees.J48 '-C 0.25 -M 2' -217733168393644444 (2) rules.OneR '-B 6' -3459427003147861443		
■ Internet Context = # CES.546 = C 0.25 +1 2 = 21775510		• • • • • • • • • • • • • • • • • • •		