

บทที่ 3

ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการประยุกต์ใช้โมเดลที่เกิดการเรียนรู้รูปแบบการตัดสินใจด้วยวิธีการจำแนกข้อมูล ของโปรแกรม WEKA เวอร์ชัน 3.6.2 โดยทำการเรียนรู้รูปแบบการตัดสินใจของข้อมูลบัญชีเงินกู้เบิกเกินบัญชีที่เคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้า ระหว่างเดือนมกราคม 2556 ถึงเดือนกันยายน 2557 จากนั้นได้พัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชี จะได้รับการทบทวนวงเงิน ที่มีลักษณะการทำงานแบบ Web Application โดยใช้ภาษา PHP และเรียกใช้รูปแบบโมเดลการตัดสินใจจากโปรแกรม WEKA โดยตรง และมีการประเมินผลว่า ผลการพยากรณ์ที่ได้ นั้น มีความถูกต้องน่าเชื่อถือ และสอดคล้องกับความเป็นจริงทางธุรกิจหรือไม่ มากน้อยเพียงใด รายละเอียดเกี่ยวกับระเบียบวิธีวิจัย มีดังต่อไปนี้

3.1 ขั้นตอนการดำเนินงาน

1. กำหนดประเด็นปัญหาในการทำวิจัย
2. เก็บรวบรวมข้อมูลที่เกี่ยวข้อง
3. ศึกษาและทำความเข้าใจกับข้อมูล
4. ศึกษารูปแบบและแนวทางการทำวิจัย
5. ศึกษาหาความรู้เพิ่มเติมเกี่ยวกับเทคนิคการทำเหมืองข้อมูล
6. จัดเตรียมข้อมูลให้อยู่รูปแบบที่พร้อมใช้งาน
7. ประมวลผลข้อมูลและคัดเลือกรูปแบบโมเดลที่เหมาะสม
8. พัฒนาโปรแกรมพยากรณ์
9. ทดสอบโปรแกรมพยากรณ์กับการปฏิบัติงานจริง
10. ประเมินผลประสิทธิภาพของโปรแกรมพยากรณ์
11. สรุปผลการวิจัย
12. เก็บรวบรวมเนื้อหา นำเสนอและจัดพิมพ์เล่มงานวิจัย

3.2 อุปกรณ์และเครื่องมือที่ใช้ในการวิจัย

1. ฮาร์ดแวร์ ได้แก่ ชุดคอมพิวเตอร์ Notebook
 - 1.1 OS : Windows 8 (64-bit)
 - 1.2 Processor : Intel Core i5 CPU 1.80 GHz
 - 1.3 RAM : 4.00 GB
 - 1.4 Hard Disk : 500 GB
2. ซอฟต์แวร์ ได้แก่
 - 2.1 ชุดโปรแกรม Microsoft Office 2010 (Access, Excel, Word, Power Point)
 - 2.2 โปรแกรม WEKA เวอร์ชัน 3.6.2
 - 2.3 ชุดโปรแกรม Appserv (Apache2.2, MySQL, PHP5)
 - 2.4 โปรแกรมอรรถประโยชน์อื่นๆ ได้แก่ EditPlus3, Sublime Text และ SnagIt

3.3 ระยะเวลาในการดำเนินการวิจัย

ตารางที่ 3.1 ระยะเวลาในการดำเนินการวิจัย ตั้งแต่เดือนตุลาคม 2556 – สิงหาคม 2557

ช่วงเวลา (เดือน) ขั้นตอน	ปี											
	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.	
1. กำหนดประเด็นปัญหาในการทำวิจัย	↔											
2. เก็บรวบรวมข้อมูล		↔										
3. ศึกษาและทำความเข้าใจกับข้อมูล		↔										

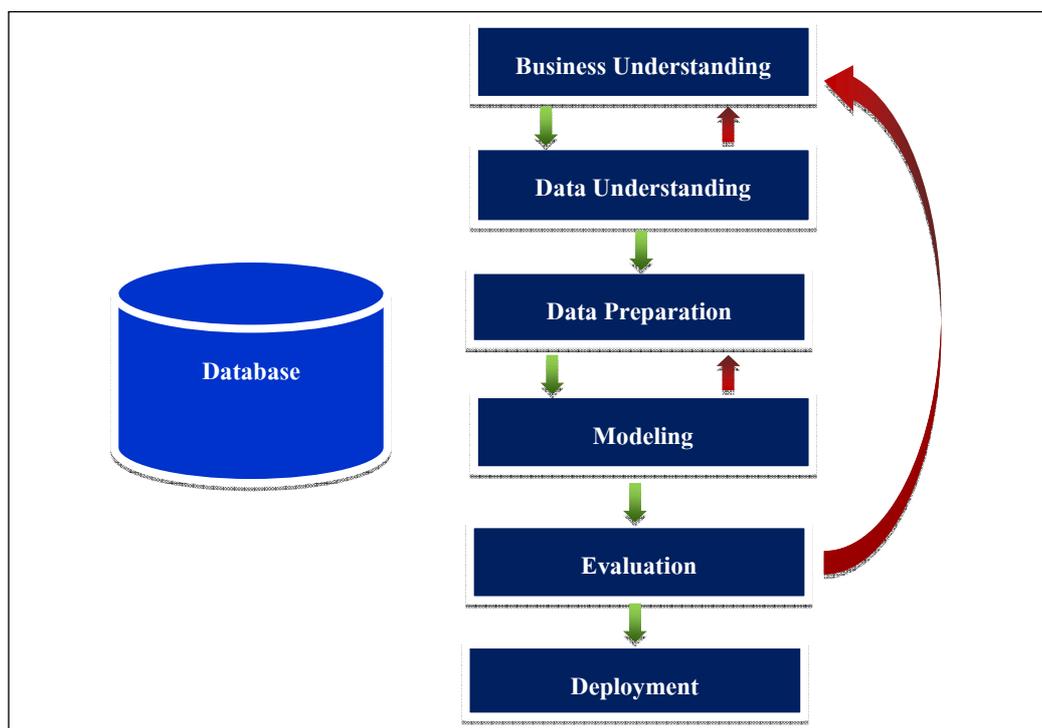
ตารางที่ 3.1 (ต่อ)

ช่วงเวลา (เดือน) ขั้นตอน	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.
	4. ศึกษารูปแบบ และแนว ทางการทำวิจัย										
5. ศึกษาหา ความรู้เพิ่มเติม											
6. จัดเตรียม ข้อมูล											
7. ประมวลผล และคัดเลือก โมเดล											
8. พัฒนา โปรแกรม พยากรณ์											
9. ทดสอบ โปรแกรม พยากรณ์											
10. ประเมินผล โปรแกรม พยากรณ์											
11. สรุป ผลการวิจัย											

ตารางที่ 3.1 (ต่อ)

ช่วงเวลา (เดือน) ขั้นตอน	ช่วงเวลา (เดือน)											
	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.	
12. เก็บรวบรวม เนื้อหา นำเสนอ และ จัดพิมพ์	←—————→											

3.4 การดำเนินการวิจัย



ภาพที่ 3.1 กระบวนการมาตรฐานในการวิเคราะห์ข้อมูล ตามมาตรฐาน CRISP-DM

งานวิจัยนี้ดำเนินการวิจัยตามขั้นตอนมาตรฐาน CRISP-DM ประกอบด้วย 6 ขั้นตอน ดังนี้

1. ขั้นตอนการทำความเข้าใจปัญหาทางธุรกิจ (Business Understanding)

เนื่องจากกระบวนการทำงานเกี่ยวกับงานทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีเดิม อยู่ในความควบคุมดูแลของเจ้าหน้าที่ธนาคารเพียงหนึ่งคน ซึ่งทำให้เจ้าหน้าที่สินเชื่อท่านอื่นไม่ได้รับความสะดวกในการสอบถามข้อมูลว่าลูกค้าที่อยู่ในความดูแลนั้นอยู่ในเกณฑ์หรือมีโอกาที่จะได้รับการทบทวนวงเงินหรือไม่ ดังนั้น เพื่ออำนวยความสะดวกแก่เจ้าหน้าที่สินเชื่อให้สามารถทำการสอบถามข้อมูลด้วยตนเองว่า บัญชีนั้นๆ จะได้รับการทบทวนวงเงินหรือไม่และมีความเป็นไปได้มากน้อยเพียงไร มีปัจจัยอะไรบ้างที่จะนำมาใช้ประกอบการพิจารณาทบทวนวงเงิน และมีหลักเกณฑ์เงื่อนไขอย่างไร ผู้วิจัยจึงมีแนวคิดที่จะนำความรู้ใหม่ที่ได้จากการทำเหมืองข้อมูลมาประยุกต์ใช้กับงานทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีดังกล่าว และพัฒนาเป็น โปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

2. ขั้นตอนการเก็บรวบรวม การตรวจสอบ และวิเคราะห์ข้อมูล (Data Understanding)

จากการทำความเข้าใจปัญหาทางธุรกิจข้างต้น ผู้วิจัยจึงได้ทำการศึกษาและเก็บรวบรวมข้อมูลบัญชีเงินกู้เบิกเกินบัญชีของลูกค้าที่อยู่ในความดูแลของฝ่ายพิจารณาสินเชื่อธุรกิจ SME ของธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) และเคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้า ตั้งแต่เดือน มกราคม 2556 ถึงเดือนกันยายน 2557 จำนวนทั้งสิ้น 13,765 เร็คคอร์ด มาทำการวิเคราะห์และคัดเลือกแอตทริบิวต์ที่เป็นข้อมูลปัจจัยสำคัญ ที่ใช้ในการพิจารณาตามหลักเกณฑ์การทบทวนวงเงิน จำนวน 4 แอตทริบิวต์ และข้อมูลที่เป็นผลการพิจารณาว่าบัญชีนั้นๆ ผ่านเกณฑ์การพิจารณาให้ทบทวนวงเงินหรือไม่ จำนวน 1 แอตทริบิวต์ รวมเป็น 5 แอตทริบิวต์ ได้แก่ Utilization, LTV, Movement, Payment และ Class

3. ขั้นตอนการเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูล ผู้วิจัยได้ทำการแบ่งข้อมูลบัญชีเงินกู้เบิกเกินบัญชี จำนวน 13,765 เร็คคอร์ด ออกเป็น 2 ชุด โดยใช้เดือนของข้อมูลเป็นเกณฑ์ในการแบ่ง ดังนี้

3.1 ข้อมูลชุดแรกเป็นข้อมูลการทบทวนวงเงิน ตั้งแต่เดือนมกราคม 2556 ถึงเดือนสิงหาคม 2557 จำนวน 12,893 เร็คคอร์ด ชุดนี้ใช้สำหรับเรียนรู้และทดสอบโมเดลด้วยโปรแกรม WEKA

3.2 ข้อมูลชุดที่สองเป็นข้อมูลการทบทวนวงเงิน ของเดือนกันยายน 2557 จำนวน 872 เร็คคอร์ด ชุดนี้ใช้สำหรับประเมินประสิทธิภาพโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

การแปลงข้อมูล (Data Transformation) ตามทฤษฎีการทำเหมืองข้อมูลด้วยวิธีการจำแนกข้อมูลนั้น ต้องทำการแปลงค่าข้อมูลในที่เก็บอยู่ในรูปแบบตัวเลข ให้เป็นรูปแบบนาม (Nominal) ทั้งนี้ เพื่อลดการกระจายข้อมูลที่มากเกินไป วิธีการ คือ ให้แบ่งค่าข้อมูลออกเป็นช่วงหรือเป็นกลุ่ม (Cluster) การแบ่งช่วงข้อมูลที่เหมาะสม ถูกต้องและสอดคล้องกับความเป็นจริงของงาน จะมีผลต่อค่าความถูกต้องของการพยากรณ์ ซึ่งค่าความถูกต้องที่สูงหรือต่ำนี้ จะเป็นตัวบ่งบอกว่าผลลัพธ์จากการพยากรณ์นั้น มีความน่าเชื่อถือหรือไม่ ซึ่งจะเห็นผลชัดเจนมากถ้าใช้วิธีการสร้างรูปแบบโมเดลจากชุดข้อมูลเรียนรู้แบบมีครู (Supervised Learning)

แอตทริบิวต์ที่ไม่ต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบ Nominal คือ แอตทริบิวต์ Class เนื่องจากข้อมูลถูกเก็บอยู่ในรูปแบบ Nominal เรียบร้อยแล้ว โดยมีค่าข้อมูลเป็น No และ Pass ส่วนที่ต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบ Nominal มีจำนวน 4 แอตทริบิวต์ ได้แก่ Utilization, LTV, Movement และ PaymentHistory รายละเอียดการแปลงข้อมูล ในตารางที่ 3.2

ตารางที่ 3.2 การแปลงข้อมูลของแต่ละแอตทริบิวต์ให้เป็น Nominal

Cluster No. Attribute	1	2	3	4	5	6	7
LTV	<=75	>75-80	>80-100	>100	Deposit	CleanLoan	NoData
PaymentHistory	Bad	Good	NoData				
Utilization	<=80	>80-90	>90-95	>95-100	>100	NoData	
Movement	<10	10-14.99	15-19.99	>=20	NoData		

เงื่อนไขการแปลงข้อมูลของแอตทริบิวต์ LTV

--Query1

```
UPDATE IS_Src SET IS_Src.LTV_txt = ">100" WHERE (((IS_Src.LTV)>1));
```

--Query2

```
UPDATE IS_Src SET IS_Src.LTV_txt = ">80-100" WHERE (((IS_Src.LTV)>0.8 And (IS_Src.LTV)<=1));
```

--Query3

```
UPDATE IS_Src SET IS_Src.LTV_txt = ">75-80" WHERE (((IS_Src.LTV)>0.75 And (IS_Src.LTV)<=0.8));
```

--Query4

UPDATE IS_Src SET IS_Src.LTV_txt = "<=75" WHERE (((IS_Src.LTV)<=0.75));

--Query5

UPDATE IS_Src SET IS_Src.LTV_txt = "CleanLoan" WHERE (((IS_Src.LTV2)="Clean"));

--Query6

UPDATE IS_Src SET IS_Src.LTV_txt = "Deposit" WHERE (((IS_Src.LTV1)="เงินฝาก"));

--Query7

UPDATE IS_Src SET IS_Src.LTV_txt = "NoData" WHERE (((IS_Src.LTV2)="#N/A" Or
(IS_Src.LTV2) Is Null));

เนื่องจากการแปลงข้อมูลของแอตทริบิวต์ PaymentHistory

--Query8

UPDATE IS_Src SET IS_Src.History_24M_txt = "Good" WHERE
(((IS_Src.History_24M)="Yes"));

--Query9

UPDATE IS_Src SET IS_Src.History_24M_txt = "Bad" WHERE
(((IS_Src.History_24M)="No"));

--Query10

UPDATE IS_Src SET IS_Src.History_24M_txt = "NoData" WHERE (((IS_Src.History_24M) Is
Null Or (IS_Src.History_24M) ="#N/A"));

เนื่องจากการแปลงข้อมูลของแอตทริบิวต์ Utilization

--Query11

UPDATE IS_Src SET IS_Src.Utilization_txt = ">100" WHERE (((IS_Src.Utilization)>1));

--Query12

UPDATE IS_Src SET IS_Src.Utilization_txt = ">95-100" WHERE (((IS_Src.Utilization)>0.95
And (IS_Src.Utilization)<=1));

--Query13

UPDATE IS_Src SET IS_Src.Utilization_txt = ">90-95" WHERE (((IS_Src.Utilization)>0.9 And
(IS_Src.Utilization)<=0.95));

--Query14

```
UPDATE IS_Src SET IS_Src.Utilization_txt = ">80-90" WHERE (((IS_Src.Utilization)>0.8 And
(IS_Src.Utilization)<=0.9));
```

--Query15

```
UPDATE IS_Src SET IS_Src.Utilization_txt = "<=80" WHERE (((IS_Src.Utilization)<=0.8));
```

--Query16

```
UPDATE IS_Src SET IS_Src.Utilization_txt = "NoData" WHERE (((IS_Src.Utilization) Is
Null));
```

เงื่อนไขการแปลงข้อมูลของแอตทริบิวต์ Movement

--Query17

```
UPDATE IS_Src SET IS_Src.Movement_txt = "<10" WHERE (((IS_Src.Movement)<0.1));
```

--Query18

```
UPDATE IS_Src SET IS_Src.Movement_txt = "10-14.99" WHERE (((IS_Src.Movement)>=0.1
And (IS_Src.Movement)<0.15));
```

--Query19

```
UPDATE IS_Src SET IS_Src.Movement_txt = "15-19.99" WHERE (((IS_Src.Movement)>=0.15
And (IS_Src.Movement)<0.2));
```

--Query20

```
UPDATE IS_Src SET IS_Src.Movement_txt = ">=20" WHERE (((IS_Src.Movement)>=0.2));
```

--Query21

```
UPDATE IS_Src SET IS_Src.Movement_txt = "NoData" WHERE (((IS_Src.Movement) Is
Null));
```

เงื่อนไขการเลือกข้อมูลเพื่อนำไปประมวลผลกับโปรแกรม WEKA

--Query22

```
SELECT IS_Src.ID AS ID_Hint, IS_Src.Utilization_txt AS Utilization, IS_Src.LTV_txt AS
LTV, IS_Src.Movement_txt AS Movement, IS_Src.History_24M_txt AS
PaymentHistory,IS_Src.Auto_Review_Class AS Class
FROM IS_Src
WHERE (((IS_Src.Utilization_txt)<>"NoData") AND ((IS_Src.LTV_txt)<>"NoData") AND
((IS_Src.Movement_txt)<>"NoData") AND ((IS_Src.History_24M_txt)<>"NoData"))
```

ORDER BY IS_Src.ID;

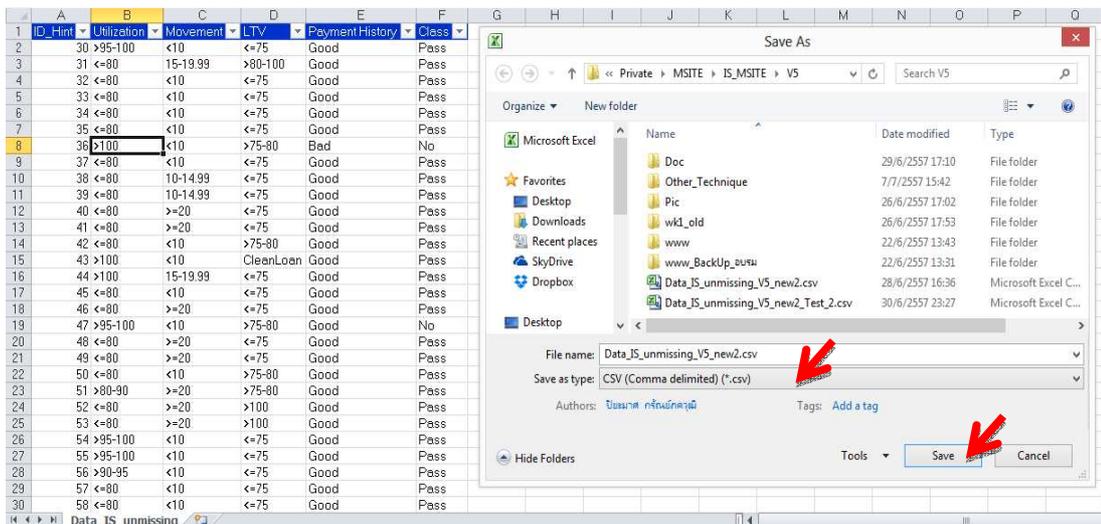
4. ขั้นตอนการวิเคราะห์และออกแบบ (Modeling)

4.1 นำข้อมูลที่ได้เก็บรวบรวมไว้ในขั้นตอนการจัดเตรียมข้อมูล มาเข้าสู่กระบวนการแปลงข้อมูลให้อยู่ในรูป Nominal โดยโปรแกรม Ms. Access ตามเงื่อนไขที่กำหนดไว้ในเงื่อนไขการแปลงข้อมูล ดังภาพที่ 3.2

ID_Hint	Utilization	Movement	LTV	PaymentHistc	Class
12850	>100	<10	<=75	Good	No
12851	<=80	<10	<=75	Good	Pass
12852	>90-95	>=20	CleanLoan	Good	Pass
12853	<=80	<10	Deposit	Good	Pass
12854	<=80	<10	>75-80	Good	Pass
12855	>80-90	10-14.99	<=75	Good	No
12856	<=80	<10	CleanLoan	Good	Pass
12857	>90-95	<10	<=75	Good	Pass
12858	<=80	<10	<=75	Good	Pass
12859	<=80	<10	<=75	Good	Pass
12860	<=80	<10	<=75	Good	Pass
12861	<=80	<10	<=75	Good	Pass
12862	>95-100	<10	CleanLoan	Good	Pass
12863	>95-100	>=20	<=75	Good	Pass
12864	>95-100	>=20	CleanLoan	Good	Pass
12865	<=80	<10	<=75	Good	Pass
12866	>95-100	<10	<=75	Good	Pass
12867	>90-95	<10	>80-100	Good	Pass
12868	<=80	>=20	CleanLoan	Good	Pass
12869	<=80	10-14.99	<=75	Good	Pass
12870	<=80	<10	<=75	Good	Pass
12871	<=80	<10	<=75	Good	Pass
12872	<=80	>=20	>80-100	Good	Pass
12873	<=80	>=20	<=75	Good	Pass
12874	>100	<10	<=75	Good	No
12875	<=80	>=20	<=75	Good	Pass
12876	>95-100	<10	>100	Good	No
12877	<=80	>=20	Deposit	Good	Pass
12878	>95-100	<10	>100	Good	No
12879	<=80	>=20	<=75	Good	Pass

ภาพที่ 3.2 การประมวลผลโดยโปรแกรม Ms. Access

4.2 Export ข้อมูลที่ได้ทำการแปลงข้อมูลจากโปรแกรม Ms. Access (.accdb) เรียบร้อยแล้ว ไปเป็นไฟล์ Excel (.xlsx) จากนั้นเปิดโปรแกรม Ms. Excel เปิดไฟล์ที่ Export มาจาก Ms. Access นั้น แล้ว Save as เป็นไฟล์ CSV (.csv) ในขั้นตอนนี้ผู้วิจัยเลือกเฉพาะเร็คคอร์ดที่มีค่าข้อมูลครบทุกแอตทริบิวต์เท่านั้น ทั้งนี้ เพื่อให้ค่าความถูกต้องของโมเดลการตัดสินใจมีความถูกต้องและมีความน่าเชื่อถือ ดังภาพที่ 3.3



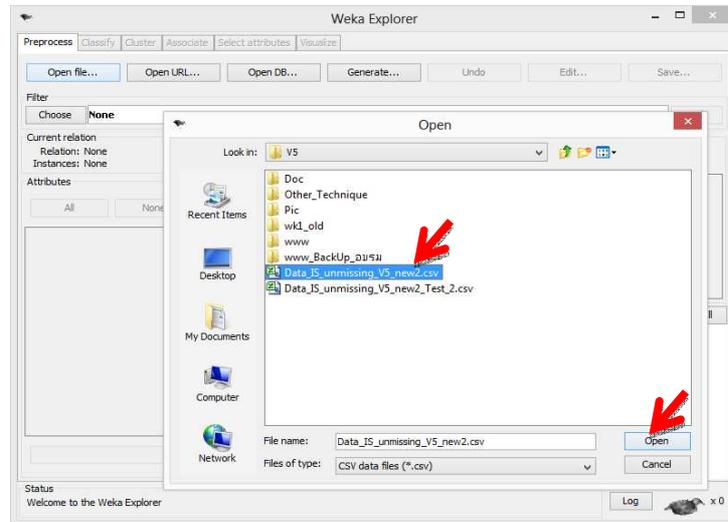
ภาพที่ 3.3 การบันทึกไฟล์ Excel เป็นไฟล์ .csv

4.3 ทำการประมวลผลข้อมูลเพื่อหารูปแบบโมเดลการตัดสินใจตามกระบวนการทำเหมืองข้อมูล โดยใช้โปรแกรม WEKA และมีลำดับขั้นตอนดังนี้

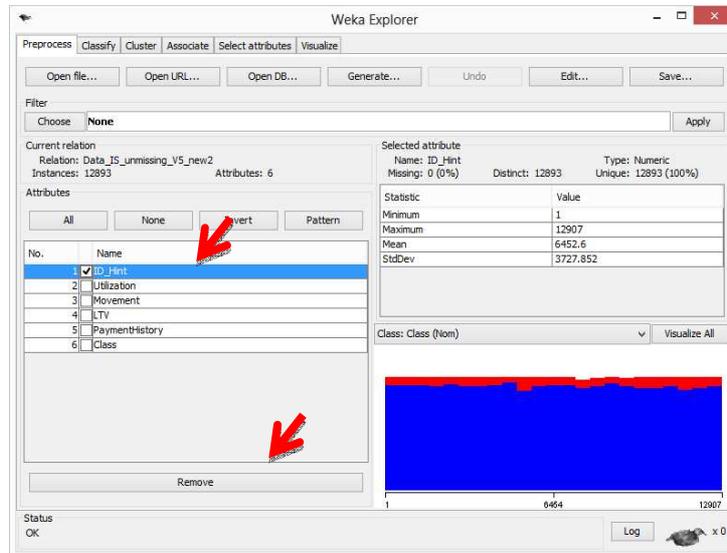
1. เปิดโปรแกรม WEKA > เลือก explorer > เลือก open file > เลือกไฟล์ .arff หรือ .csv ที่ต้องการใช้เป็นชุดข้อมูลเรียนรู้ ดังภาพที่ 3.4 และภาพที่ 3.5
2. ลบแอตทริบิวท์ที่ไม่เกี่ยวข้องออก > คลิกเลือก > Remove ดังภาพที่ 3.6
3. เลือกแอตทริบิวท์ที่ต้องการ รวมถึงแอตทริบิวท์ที่จะใช้เป็นตัวจัดคลาสด้วย (กรณีใช้วิธีจำแนกข้อมูล) กรณีนำเข้าข้อมูลที่เป็นไฟล์ .csv ให้ save เป็นไฟล์ .arff ดังภาพที่ 3.7



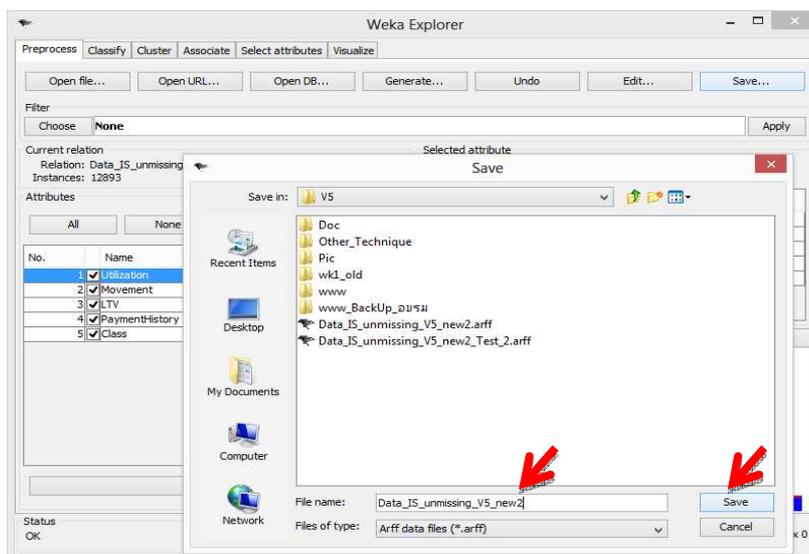
ภาพที่ 3.4 โปรแกรม WEKA เวอร์ชัน 3.6.2



ภาพที่ 3.5 ตัวอย่างการนำเข้าข้อมูล .csv



ภาพที่ 3.6 ตัวอย่างการลบแอตทริบิวท์ที่ไม่เกี่ยวข้อง



ภาพที่ 3.7 ตัวอย่างการบันทึกไฟล์เป็น .arff

4.4 เลือกวิธีวิเคราะห์ข้อมูลด้วยวิธีจำแนกข้อมูล (Classify) งานวิจัยนี้ใช้เทคนิค J48 และ ID3 และเลือกใช้วิธีทดสอบประสิทธิภาพโมเดลกับทั้ง 2 เทคนิค ดังนี้

4.4.1 คลิก Tab classify > เลือกเทคนิคที่ต้องการใช้จำแนกข้อมูล ที่ Classifier ดังภาพที่ 3.8

4.4.2 เลือกวิธีการวัดประสิทธิภาพโมเดลที่กำลังจะถูกสร้างขึ้นที่ Test Options > เลือกวิธีวัดประสิทธิภาพโมเดลแบบ Cross-validation 10 Folds

ลักษณะการทำงานของ Test Options แต่ละแบบของโปรแกรม WEKA มี 4 รูปแบบ ดังนี้

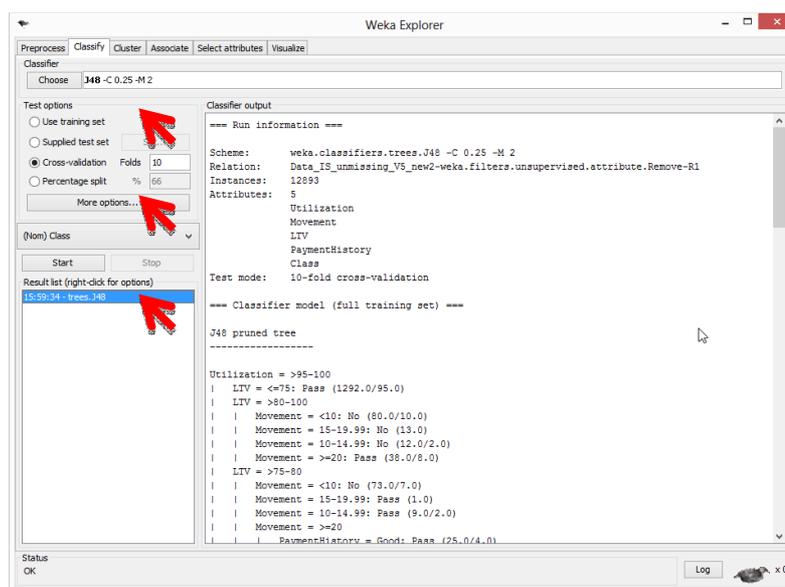
1. Use training set คือ เป็นการกำหนดให้โปรแกรม WEKA ใช้ชุดข้อมูลเรียนรู้นั้นเป็นตัวทดสอบโมเดลที่สร้างขึ้น

2. Supplied test set คือ เป็นการนำชุดข้อมูลทดสอบคนละชุดกับข้อมูลเรียนรู้ ซึ่งถูกจัดเตรียมไว้เรียบร้อยแล้ว และต้องเป็นไฟล์ .arff ที่มีโครงสร้างและคุณสมบัติของข้อมูลเหมือนกับชุดข้อมูลเรียนรู้

3. Cross-validation คือ เป็นการนำความสามารถพิเศษที่มีอยู่ในโปรแกรม WEKA ให้ทำการแบ่งชุดข้อมูลเรียนรู้ออกเป็นชุดๆ ละเท่าๆ กัน ตามจำนวน Folds ที่กำหนดโดยทุกๆ รอบของการเรียนรู้รูปแบบข้อมูลเพื่อที่จะสร้างโมเดลนั้น โปรแกรม WEKA จะเหลือข้อมูลไว้ 1 ชุด เพื่อ

ใช้สำหรับทดสอบและวัดประสิทธิภาพของโมเดล และจะทำเช่นนี้ไปเรื่อยๆ จนกว่าจะครบทุกชุด ดังนั้น ค่าความน่าจะเป็นที่ได้จึงเป็นค่าความน่าจะเป็นเฉลี่ย

4. Percentage split คือ เป็นการใช้ความสามารถพิเศษที่มีอยู่ในโปรแกรม WEKA ให้จัดการแบ่งชุดข้อมูลเรียนรู้่ออกเท่ากับ % split ที่ถูกกำหนดและส่วนที่เหลือนำไปใช้สำหรับทดสอบและวัดประสิทธิภาพของโมเดล



ภาพที่ 3.8 การเลือกวิธีการจำแนกข้อมูลและวิธีการใช้ข้อมูลทดสอบ

4.5 หากต้องการ Set Option เพิ่มเติม ให้เลือกที่ More Option

4.6 คลิก Start เพื่อรันโปรแกรม

4.7 ผลจากการประมวลผลข้อมูลจะได้รูปแบบโมเดลการตัดสินใจของเทคนิค J48 และ ID3 ดังที่แสดงผลรูปแบบการตัดสินใจอย่างย่อ ในภาพที่ 3.9 และภาพที่ 3.10

```

Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:   12893
Attributes:  5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
-----

Utilization = >95-100
| LTV = <=75: Pass (1292.0/95.0)
| LTV = >80-100
| | Movement = <10: No (80.0/10.0)
|
.
.
Utilization = <=80: Pass (8202.0/58.0)
Utilization = >100
| LTV = <=75
| | Movement = <10: No (293.0/4.0)
| | Movement = 15-19.99: Pass (2.0/1.0)
|
.
.
Utilization = >80-90
| LTV = <=75: Pass (858.0/16.0)
| LTV = >80-100
| | Movement = <10: No (42.0/3.0)
|
.
.
Utilization = >90-95
| LTV = <=75: Pass (656.0/33.0)
| LTV = >80-100
| | Movement = <10: No (25.0/4.0)
|
.
.
| LTV = Deposit: Pass (50.0/2.0)

Number of Leaves :    55

Size of the tree :    72

Time taken to build model: 0.06 seconds

```

ภาพที่ 3.9 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค J48

```

Scheme:      weka.classifiers.trees.Id3
Relation:    Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:   12893
Attributes:  5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

Id3

Utilization = >95-100
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = <=80
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = >100
| LTV = <=75
| | Movement = <10
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
.
.
Utilization = >80-90
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = >90-95
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
| | Movement = >=20: Pass

Time taken to build model: 0.03 seconds

```

ภาพที่ 3.10 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค ID3

จากการวิเคราะห์ผลการประมวลผลข้อมูลของโปรแกรม WEKA ในขั้นตอนนี้พบว่าเทคนิค J48 มีลำดับชั้นของการจำแนกข้อมูล 72 ชั้น (layers) และได้รูปแบบการตัดสินใจ 55 รูปแบบ และพบว่าเทคนิค ID3 มีลำดับชั้นของการจำแนกข้อมูล 181 ชั้น (layers) และได้รูปแบบการตัดสินใจ 132 รูปแบบ นอกจากนี้ยังพบว่าเทคนิค ID3 มีรูปแบบการตัดสินใจที่ไม่สามารถพยากรณ์ข้อมูลได้ จำนวน 4 รูปแบบ คือแสดงผลการพยากรณ์เป็น “null”

5. ขั้นตอนการวัดประสิทธิภาพของรูปแบบหรือโมเดล (Evaluation)

จากการวัดประสิทธิภาพของโมเดลการตัดสินใจที่ได้จากเทคนิค J48 และ ID3 ของโปรแกรม WEKA ได้ผลลัพธ์ ดังนี้

```

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

Correctly Classified Instances   12495      96.9131 %
Incorrectly Classified Instances    398      3.0869 %
Kappa statistic                   0.7673
Mean absolute error                0.0545
Root mean squared error            0.1671
Relative absolute error            37.2147 %
Root relative squared error        61.7539 %
Total Number of Instances         12893

=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.992   0.297   0.975     0.992   0.983     0.92     Pass
                0.703   0.008   0.885     0.703   0.784     0.92     No
Weighted Avg.   0.969   0.274   0.968     0.969   0.968     0.92

=== Confusion Matrix ===
      a  b  <-- classified as
11774  94 |  a = Pass
 304   721 |  b = No
  
```

ภาพที่ 3.11 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค J48

```

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

Correctly Classified Instances      12489      96.8665 %
Incorrectly Classified Instances    400        3.1025 %
Kappa statistic                     0.7661
Mean absolute error                 0.0531
Root mean squared error             0.1659
Relative absolute error             36.2797 %
Root relative squared error         61.3118 %
UnClassified Instances              4          0.031 %
Total Number of Instances          12893

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.992    0.298    0.975     0.992    0.983     0.938    Pass
      0.702    0.008    0.883     0.702    0.783     0.939    No
Weighted Avg.  0.969    0.275    0.967     0.969    0.967     0.938

=== Confusion Matrix ===

  a    b  <-- classified as
11769  95 |  a = Pass
 305   720 |  b = No

```

ภาพที่ 3.12 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค ID3

6. ขั้นตอนการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ให้เกิดประโยชน์ต่อองค์กร (Deployment)

จากการวิเคราะห์เปรียบเทียบรูปแบบโมเดลและค่าดัชนีที่ใช้วัดประสิทธิภาพของรูปแบบโมเดลการตัดสินใจที่ได้จากทั้งเทคนิค J48 และ ID3 ผลปรากฏว่า เทคนิค J48 มีประสิทธิภาพในการจำแนกหรือพยากรณ์ข้อมูลได้ดีกว่าเทคนิค ID3 ผู้วิจัยจึงได้เลือกใช้โมเดลการตัดสินใจที่ได้จากเทคนิค J48 ไปใช้กับโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษณะการใช้งานแบบ Web Application โดยใช้ภาษา PHP และเรียกใช้รูปแบบโมเดลการตัดสินใจจากโปรแกรม WEKA เวอร์ชัน 3.6.2 โดยตรง ดังรายละเอียดเพิ่มเติมในบทที่ 4