



การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในกระบวนการอนุมัติสินเชื่อ
APPLYING DATA MINING TECHNIQUE IN LOAN APPROVAL PROCESS

เสถียร วิหาเรือง

งานค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553

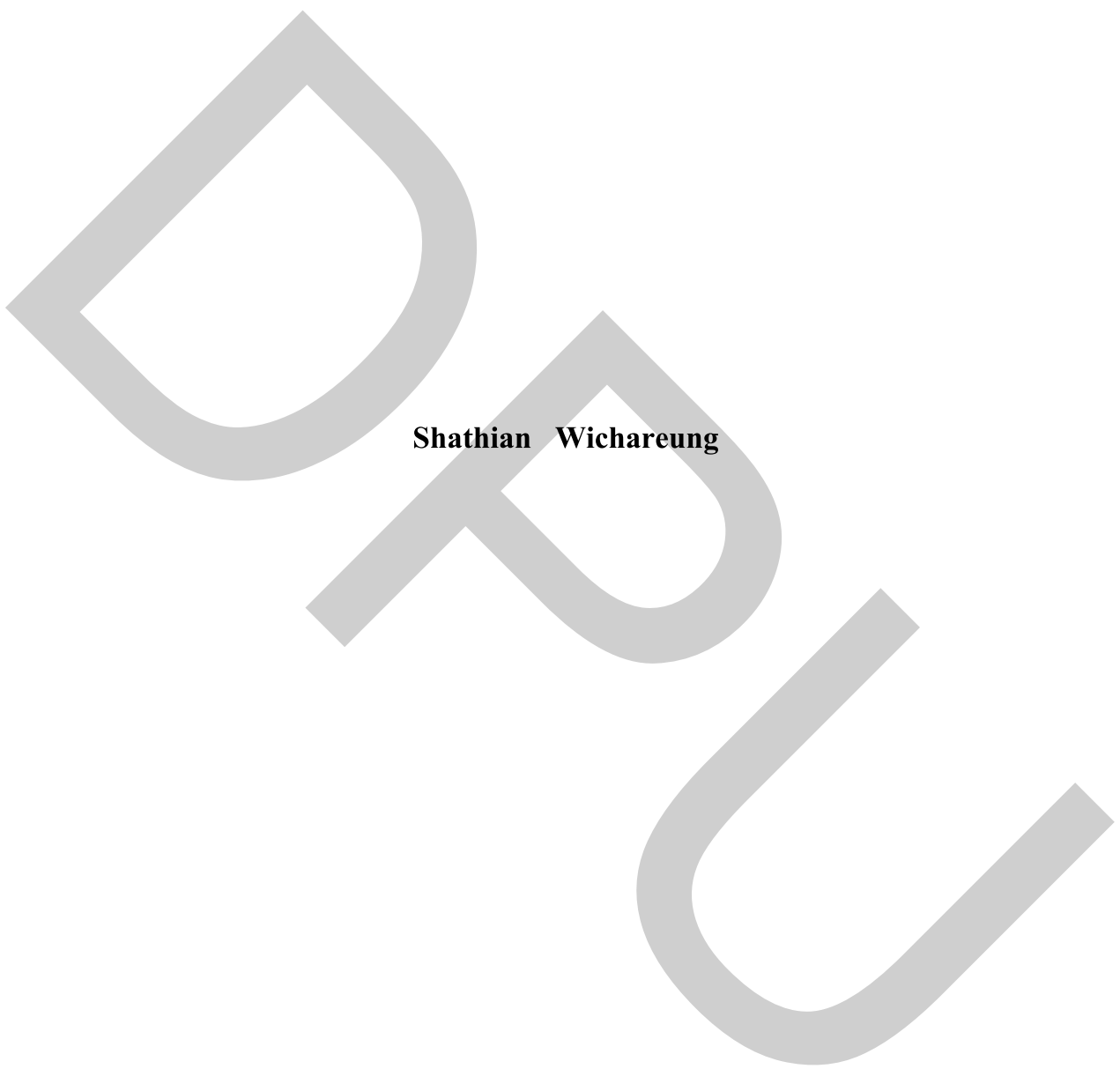
การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในกระบวนการอนุมัติสินเชื่อ

เสถียร วิชาเรือง

งานค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553

Applying Data Mining Technique in Loan Approval Process



Shathian Wichareung

**An Independent Study Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Computer and Communication Technology)**

Department of Computer and Communication Technology

Graduate School, Dhurakij Pundit University

2010

กิตติกรรมประกาศ

งานค้นคว้าอิสระฉบับนี้สำเร็จลุล่วงได้ด้วยดี ด้วยความอนุเคราะห์และเสียสละเวลาอันมีค่าของอาจารย์ที่ปรึกษางานค้นคว้าอิสระ อาจารย์ ดร.อรรณ อัมสมบัติ ที่มีความเมตตากรุณาให้คำปรึกษา แนะนำ ช่วยเหลือ ตรวจสอบ ด้วยความเอาใจใส่เป็นกันเองอย่างดียิ่ง ซึ่งผู้จัดทำมีความซาบซึ้งและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณผู้ช่วยศาสตราจารย์ ดร.ประณต บุญไชยอภิสิทธิ์ ผู้อำนวยการหลักสูตรฯ สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณาจารย์ทุกท่านตลอดจนเจ้าหน้าที่มหาวิทยาลัยธุรกิจบัณฑิต ที่ให้คำแนะนำ สนับสนุนและช่วยเหลือในการจัดทำงานค้นคว้าอิสระนี้

ขอขอบคุณมหาวิทยาลัยธุรกิจบัณฑิตที่ให้การสนับสนุนทุนการศึกษา และ ขอบพระคุณ บริษัทเอกชนกรหลวงลิซซิ่ง จำกัด คุณศราธร อินทรประสาท ที่มีความเมตตากรุณาให้ข้อมูลในการศึกษา แนะนำ ช่วยเหลือ ตลอดจนพนักงานของบริษัทเอกชนกรหลวงลิซซิ่ง จำกัด หลายๆท่านที่ไม่ได้เอ่ยนามที่ได้ให้ความร่วมมือในการทำงานค้นคว้าอิสระนี้

ขอกราบขอบพระคุณบิดามารดา ญาติพี่น้องทุกคน ที่เป็นกำลังใจ พร้อมทั้ง พี่น้อง และเพื่อนๆร่วมคณะที่ให้คำปรึกษา ในการทำงานค้นคว้าอิสระนี้

ผู้วิจัยหวังเป็นอย่างยิ่งว่า งานค้นคว้าอิสระฉบับนี้ จะเป็นประโยชน์กับผู้ที่ต้องการศึกษาด้านการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในกระบวนการอนุมัตินิเชื่อ และหากมีข้อผิดพลาดประการใดในงานค้นคว้าอิสระฉบับนี้ ผู้วิจัยต้องกราบขออภัยมา ณ ที่นี้ด้วย

เสถียร วิหาเรือง

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	๗
บทคัดย่อภาษาอังกฤษ.....	๘
กิตติกรรมประกาศ.....	๙
สารบัญตาราง.....	๑๐
สารบัญภาพ.....	๑๑
บทที่	
1. บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์.....	2
1.3 ขอบเขตของการวิจัย	3
1.4 ผลที่คาดว่าจะได้รับ.....	3
1.5 นิยามศัพท์เฉพาะ.....	3
2. แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง.....	5
2.1 เหมืองข้อมูล (Data Mining).....	5
2.2 เทคนิคการจำแนกประเภทข้อมูล(Data Classification).....	18
2.3 โปรแกรม WEKA.....	18
2.4 ความรู้ทั่วไปเกี่ยวกับระบบฐานข้อมูล.....	21
2.5 เอสคิวแอล (SQL)	26
2.6 งานวิจัยที่เกี่ยวข้อง.....	30
3. ระเบียบวิธีวิจัย.....	34
3.1 ขั้นตอนการดำเนินการวิจัย.....	34
3.2 เครื่องมือ (Tool) ที่ใช้ในการดำเนินการ.....	34
3.3 ระยะเวลาในการดำเนินการวิจัย.....	35
3.4 สรุป.....	37

สารบัญ (ต่อ)

	หน้า
4. ผลดำเนินการวิจัย.....	39
4.1 การศึกษาและรวบรวมข้อมูล.....	39
4.2 การเตรียมข้อมูล (Data Pre-processing)	40
4.3 การศึกษาทดสอบและเลือกเทคนิคที่เหมาะสม.....	51
4.4 การวิเคราะห์และสรุปผล.....	56
4.5 การพัฒนาระบบสนับสนุนการตัดสินใจการอนุมัติสินเชื่อ.....	65
5. สรุปผลการศึกษาและข้อเสนอแนะการวิจัย.....	71
5.1 สรุปผลการวิจัย.....	71
5.2 ปัญหาและแนวทางแก้ไข.....	72
5.3 ข้อเสนอแนะ.....	73
บรรณานุกรม.....	74
ภาคผนวก.....	77
ประวัติผู้เขียน.....	98

สารบัญตาราง

ตารางที่	หน้า
2.1 แสดงตารางข้อมูลเกี่ยวกับลูกค้า.....	10
2.2 แสดงตารางพนักงาน.....	24
3.1 ระยะเวลาในการดำเนินการวิจัย.....	35
4.1 โครงสร้างข้อมูลของแฟ้มข้อมูลลูกค้า(บางส่วน).....	39
4.2 โครงสร้างข้อมูลที่ใช้ในการวิเคราะห์การพิจารณาอนุมัติสินเชื่อ.....	46
4.3 การแปลงค่าของข้อมูลสถานะภาพ	47
4.4 ผลการวิเคราะห์ค่าของข้อมูลประเภทอาชีพ.....	49
4.5 ผลการวิเคราะห์ค่าของข้อมูลอายุ.....	49
4.6 ผลการวิเคราะห์ค่าของข้อมูลกลุ่มสินค้า.....	49
4.7 ผลการวิเคราะห์ค่าของข้อมูลยี่ห้อสินค้า.....	50
4.8 ผลการวิเคราะห์ค่าของข้อมูลปีที่ผลิตสินค้า.....	50
4.9 ผลการวิเคราะห์ค่าของข้อมูลจำนวนงวดที่ผ่อนชำระ.....	50
4.10 ผลการวิเคราะห์ค่าของข้อมูลราคาสินค้า.....	50
4.11 ผลการวิเคราะห์ค่าของข้อมูลยอดกู้.....	51
4.12 ผลการวิเคราะห์ค่าของข้อมูลค่างวด.....	51
4.13 ผลการวิเคราะห์ค่าของข้อมูลเงินคาวน.....	51
4.14 ผลการวิเคราะห์ค่าของข้อมูลสถานภาพลูกค้า.....	51
4.15 การหาค่ากฎที่หน้าเชื่อถือ.....	65

สารบัญภาพ

ภาพที่	หน้า
2.1 กระบวนการของ Data Mining (KDD: Knowledge Discovery in Database).....	7
2.2 Architecture of a Typical Data Mining System.....	9
2.3 ตัวอย่างของ Neural Network.....	13
2.4 Data Mining: Confluence of Multiple Disciplines.....	16
2.5 แสดงหน้าจอการทำงานโปรแกรม Weka.....	19
3.1 แผนผังขั้นตอนการดำเนินการวิจัย.....	36
4.1 ข้อมูลในระบบปฏิบัติการ DOS	39
4.2 การนำออกข้อมูล.....	40
4.3 การเปิดโปรแกรม DBF Manager เพื่อศึกษาโครงสร้างข้อมูล.....	41
4.4 หน้าจอโปรแกรม DBF Manager.....	41
4.5 การตรวจสอบโครงสร้างของข้อมูล (Data Structure).....	42
4.6 การบันทึกข้อมูล.....	42
4.7 การเปิดโปรแกรม SQL Server 2008.....	43
4.8 การนำเข้าไฟล์ DBF ให้มาอยู่ใน SQL Server 2008.....	43
4.9 การสืบค้นข้อมูลในฐานข้อมูลใหม่ที่ตั้งสร้างขึ้น.....	44
4.10 แสดงการสืบค้นหาแอทริบิวต์ที่ต้องการ.....	44
4.11 ข้อมูลที่นำออกมาอยู่ในไฟล์ Excel.....	45
4.12 ข้อมูลที่ได้จากการลดขนาดของข้อมูล.....	46
4.13 ผลลัพธ์การแปลงค่าของข้อมูลสถานะภาพ.....	47
4.14 การบันทึกข้อมูลในรูปแบบนามสกุล .csv.....	48
4.15 แสดงไฟล์ข้อมูลในรูปแบบ.csv.....	48
4.16 การเปิดโปรแกรม Weka3.5.7.....	53
4.17 การนำเข้าข้อมูลในโปรแกรม Weka.....	53
4.18 หน้าจอการวิเคราะห์ข้อมูล.....	54
4.19 การเลือกเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูล.....	54
4.20 หน้าจอแสดงผลลัพธ์การจำแนกกลุ่มข้อมูลด้วยเทคนิค Decision Rule: Part.....	55
4.21 หน้าจอแสดงผลลัพธ์ความถูกต้องของโมเดล.....	55

สารบัญภาพ (ต่อ)

ภาพที่	หน้า
4.22 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Tree C4.5.....	56
4.23 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Tree: Partial Rules.....	57
4.24 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Table.....	58
4.25 การสร้างฐานข้อมูล.....	66
4.26 การสร้างตารางในฐานข้อมูล.....	66
4.27 หน้าจอหลักของระบบวิเคราะห์สินเชื่อออนไลน์.....	67
4.28 หน้าจอการรับข้อมูลลูกค้าที่ต้องการขอสินเชื่อ.....	67
4.29 หน้าจอแสดงผลการวิเคราะห์ข้อมูลลูกค้าที่ดี.....	68
4.30 หน้าจอแสดงผลการวิเคราะห์ข้อมูลลูกค้าที่ไม่ดี.....	68
4.31 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนแรก.....	69
4.32 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนที่เป็นเงื่อนไข.....	69
4.33 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนสุดท้าย.....	70
4.34 หน้าจอแสดงรายงานข้อมูลลูกค้า.....	70

หัวข้องานค้นคว้าอิสระ

การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในกระบวนการ
การอนุมัติสินเชื่อ

ชื่อผู้เขียน

เสถียร วิชาเรือง

อาจารย์ที่ปรึกษางานค้นคว้าอิสระ

อาจารย์ ดร.อรรณณ อัมสมบัติ

สาขาวิชา

เทคโนโลยีคอมพิวเตอร์และการสื่อสาร

ปีการศึกษา

2552

บทคัดย่อ

ระบบเศรษฐกิจในปัจจุบันมีการแข่งขันกันมากขึ้น ธุรกิจต่างๆจึงต้องมีการประยุกต์ใช้เครื่องมือทางเทคโนโลยีสารสนเทศเพื่อให้งานบรรลุเป้าหมายขององค์กร และชนะคู่แข่งซึ่งมีอยู่เป็นจำนวนมาก การทำเหมืองข้อมูลเป็นวิธีการหนึ่งที่หลายๆองค์กรได้นำมาใช้เป็นเครื่องมือในการค้นหาความรู้หรือกฎที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่เพื่อที่จะนำความรู้ที่ได้ไปเป็นแนวทางในการปฏิบัติงาน

การศึกษาครั้งนี้ได้ประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อนำความรู้ที่ได้จากการทำเหมืองข้อมูลมาเป็นแนวทางสนับสนุนการตัดสินใจการอนุมัติสินเชื่อ โดยข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลการเช่าซื้อรถจักรยานยนต์ของบริษัทเอกชนแห่งหนึ่ง ในการศึกษาจะเริ่มตั้งแต่การศึกษาและวิเคราะห์โครงสร้างของข้อมูลที่มีอยู่ การโอนย้ายข้อมูล การลดขนาดของข้อมูล และการทำความสะอาดข้อมูลเพื่อให้ได้ข้อมูลที่สมบูรณ์เหมาะสำหรับการทำเหมืองข้อมูล ซึ่งจะได้ข้อมูลจำนวน 17,964 สัญญา หลังจากนั้นจึงได้นำมาทดสอบด้วยเทคนิคการจำแนกกลุ่ม (Classification) เพื่อหากฎที่ใช้ในการจำแนกลูกค้ากลุ่มดีและกลุ่มไม่ดี ในการศึกษาครั้งนี้ได้ทดลองจำแนกกลุ่มข้อมูลด้วยเทคนิคการจำแนกกลุ่ม 3 เทคนิคที่ให้ผลลัพธ์เป็นกฎ คือ Decision Tree: C4.5, Decision Tree: Partial Rules และ Decision Table ผลการทดลองพบว่า การจำแนกกลุ่มแบบ Decision Tree: Partial Rules ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด คือ ร้อยละ 72.7 และได้กฎจำนวน 346 กฎ กฎที่ได้ทั้งหมดนี้จะถูกนำมาวิเคราะห์ด้วยผู้เชี่ยวชาญอีกครั้งเพื่อเลือกเฉพาะกฎที่มีความเหมาะสม ซึ่งคงเหลือเพียง 188 กฎ นอกจากนี้ผู้วิจัยยังนำกฎที่ได้มาพัฒนาเป็นระบบสนับสนุนการอนุมัติสินเชื่อเพื่อใช้ในการพิจารณาอนุมัติสินเชื่อให้กับลูกค้ารายใหม่ต่อไปอีกด้วย

Independent Study Title	Applying Data Mining Technique in Loan Approval Process
Author	Shathian Wichareung
Independent Study Advisor	Dr. Aurawan Imsombut
Department	Computer and Communication Technology
Academic Year	2009

ABSTRACT

The current economic system is more competitive. Businesses must apply to the information technology tools to make the work of corporate goals and win the competition, which contained a lot of Data mining is one of several organizations have used as tools to find knowledge or hidden rules in a large database to apply the knowledge to guide.

This time education applies data mining techniques to apply knowledge from data mining as the guidelines decision support credit approval. By the information that is used in the study as hire purchase motorcycle company one of the private sector. In education begins to study and analyze the structure of the existing data. To transfer data to reduce the size of the data and data cleaning to get complete information suitable for data mining, which is the number of 17,964 contracts then brought test with classification group (Classification) to find the rules that are used to classify the customer, good and bad group. To study this experiment with harmonized data group classification group 3 technical result rule is Decision Tree: C 4.5, Decision Tree: Partial Rules and Decision Table trial results found that classification groups Decision Tree: Partial Rules, the result is maximum is 72.7 percent and rule number 346 rule. The rule that all this will be analyzed by experts again to select only the rules that are appropriate only 188 rules hand from research to apply rules that have developed a system support credit approval to credit approval, with subsequent new customers.

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันคอมพิวเตอร์มีบทบาทอย่างมากในการด้านธุรกิจ เครื่องมือที่สำคัญที่ช่วยในการวิเคราะห์ข้อมูล ซึ่งเราเรียกระบบคอมพิวเตอร์ที่ใช้เพื่อการนี้ว่า ระบบสารสนเทศ (Information system) หน้าที่หลักของระบบสารสนเทศคือการนำข้อมูล (Data) ที่มีอยู่ในองค์กรมาประมวลผลเพื่อให้เกิดสารสนเทศ (Information) ที่สามารถนำไปใช้ประโยชน์ได้ในการบริหารองค์กร (กิตติพงษ์ กลมกล่อม 2548: 2) การวิเคราะห์สินเชื่อก็เป็นกิจกรรมที่สำคัญอย่างหนึ่งขององค์กรธุรกิจทางการเงิน เนื่องจากการให้สินเชื่อปล่อยให้ผู้กู้ทำธุรกรรมทางการเงินซึ่งดอกเบี้ยที่เป็นรายได้หล่อเลี้ยงองค์กรที่ให้สินเชื่อที่ดำเนินการเกี่ยวกับการเงินธนาคาร บริษัทเงินทุนหลักทรัพย์ และสถาบันการเงินต่างๆ ให้สามารถดำเนินกิจการต่อไปได้ ดังนั้นการวิเคราะห์สินเชื่อจึงเป็นขั้นตอนที่สำคัญที่ใช้ในการบริหารองค์กร ในการวิเคราะห์การอนุมัติสินเชื่อโดยทั่วไปนั้นจะเป็นการนำเสนอในรูปแบบรายงานรายละเอียดของลูกค้า เช่นรายงานการใช้จ่ายเครดิต การผ่อนชำระ รายงานพฤติกรรมการใช้บัตรเครดิต รายงานยอดค้างชำระของลูกค้า รายงานเพื่อช่วยในการตัดสินใจของผู้บริหาร ข้อมูลดังกล่าวนี้ต้องถูกต้องทันต่อเหตุการณ์ที่เปลี่ยนแปลงตลอดเวลา ปัจจุบันข้อมูลที่เกิดจากการทำธุรกรรม (Transaction) มีจำนวนมากมาย การนำข้อมูลดังกล่าวมาใช้ได้อย่างเหมาะสม รวมถึงความสามารถของบุคคลในการวิเคราะห์ข้อมูลจึงมีความสำคัญเป็นอย่างมากโดยเฉพาะอย่างยิ่งในปัจจุบันที่มีการแข่งขันที่รุนแรงมากดังนั้นวิธีที่นิยมของผู้บริหารในหลายองค์กรคือการนำข้อมูลจำนวนมากที่มีอยู่มาวิเคราะห์โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นการวิเคราะห์ข้อมูลในฐานข้อมูลขนาดใหญ่ เพื่อช่วยในการตัดสินใจของผู้บริหาร ซึ่งเหมืองข้อมูล จะเป็นการค้นหา วิเคราะห์ หรือสร้างองค์ความรู้ใหม่จากข้อมูลขนาดใหญ่ซึ่งอาจจะเป็นการค้นหารูปแบบหรือกฎ โดยการใช้เทคนิคทางสถิติ ทางคณิตศาสตร์หรือเทคนิคทางวิทยาการคอมพิวเตอร์ ซึ่งการทำเหมืองข้อมูลนั้นมักถูกกล่าวถึงในหัวข้อเกี่ยวกับ Knowledge discovery หรือ KDD (Knowledge Discovery in Database) เนื่องจากเหมืองข้อมูลถือเป็นขั้นตอนที่สำคัญในการทำ KDD โดยในการทำเหมืองข้อมูลจะต้องอาศัยการทำคลังข้อมูล (Data warehousing) ซึ่งเป็นการเก็บรวบรวมข้อมูลจำนวนมากเป็นฐานข้อมูลขนาดใหญ่ และเราสามารถนำข้อมูลเหล่านั้นมาวิเคราะห์ข้อมูลแบบหลายมิติ (Multidimensional data analysis) ได้ หรือสามารถทำการประมวลผลในเชิงวิเคราะห์แบบ

ออนไลน์ (On-Line-Analytic Processing หรือ OLAP)ได้ อีกทั้งยังสามารถเรียกใช้ข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ เพื่อเป็นประโยชน์ต่อการตัดสินใจของผู้บริหาร นอกจากนี้การทำการคลังข้อมูลนั้นยังลดปัญหาที่เกิดขึ้นจากการใช้ข้อมูลจากฐานข้อมูลปฏิบัติงาน (Operational database) เช่น ปัญหาความไม่ยืดหยุ่นหรือไม่สามารถเปลี่ยนแปลงรูปแบบของสารสนเทศตามความต้องการผู้ใช้งานได้ ปัญหาการไม่สามารถวิเคราะห์

ข้อมูลที่ซับซ้อนได้ ปัญหาการไม่มีการเก็บข้อมูลย้อนหลัง เป็นต้น ดังนั้นขั้นตอนที่สำคัญยิ่งในการทำเหมืองข้อมูลคือขั้นตอนการสร้างคลังข้อมูล โดยจะต้องมีการรวบรวมข้อมูลที่กระจายตามหน่วยงานต่างๆ การทำความสะอาดข้อมูล และการปรับรูปแบบของข้อมูลเหล่านั้นให้อยู่ในมาตรฐานเดียวกัน ถัดไปคือการพิจารณาเลือกใช้เทคนิคการทำเหมืองข้อมูลที่เหมาะสม โดยวิธีการหนึ่งที่สำคัญในการทำเหมืองข้อมูลคือ การรวมกลุ่ม (Clustering) โดยการรวมกลุ่มมีหลายเทคนิคด้วยกัน เช่น เทคนิคการรวมกลุ่มแบบมีลำดับชั้น (Hierarchical clustering technique) และเทคนิค K-Means ซึ่งการรวมกลุ่มนี้เป็นเทคนิคที่เหมาะสมในการนำมาใช้ในการแบ่งกลุ่มลูกค้าเพื่อวิเคราะห์สินเชื่อ โดยจะทำให้บริษัทสามารถแบ่งกลุ่มลูกค้าได้อย่างเหมาะสม

งานวิจัยนี้จึงนำเสนอการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการวิเคราะห์แบ่งกลุ่มลูกค้าเพื่อพิจารณาการอนุมัติสินเชื่อ โดยเป็นกรณีศึกษาจากบริษัทเอกชนหลวงลิสซึ่งจำกัด โดยใช้ข้อมูลลูกค้า ในปี 2546 – 2550 ที่ปิดบัญชีไปแล้ว จำนวน 20,000 บัญชี ซึ่งการวิเคราะห์การอนุมัติสินเชื่อนี้มีประโยชน์ในการตัดสินใจของผู้บริหารเกี่ยวกับนโยบายส่งเสริมการปล่อยกู้ต่าง ๆ ที่จำเป็นต้องทำอย่างรวดเร็ว ถูกต้อง และทันต่อเหตุการณ์ คลังข้อมูลและเทคนิคการทำเหมืองข้อมูลจะช่วยทำให้การวิเคราะห์การอนุมัติสินเชื่อมีความรวดเร็วและมีประสิทธิภาพมากขึ้น อีกทั้งยังสามารถดึงสารสนเทศที่ซ่อนอยู่ในข้อมูลออกมา เพื่อช่วยในการตัดสินใจในการบริหารงานให้ทันต่อการเปลี่ยนแปลงในทางธุรกิจ เช่น การรวมกลุ่มลูกค้าที่มีลักษณะคล้ายคลึงกัน จะสามารถทำให้เรากำหนดกิจกรรมต่างๆที่จะกระทำต่อลูกค้ามีประสิทธิภาพ ถูกกลุ่มเป้าหมายเพิ่มมากขึ้น เพื่อเพิ่มจำนวนลูกค้าชั้นดีให้กับองค์กร

1.2 วัตถุประสงค์ของการวิจัย

วัตถุประสงค์ของการวิจัยมีดังต่อไปนี้

1. เพื่อศึกษาเทคนิคการทำการคลังข้อมูลและการทำเหมืองข้อมูล
2. เพื่อศึกษาโครงสร้างข้อมูลลูกค้าที่ใช้ในการวิเคราะห์สินเชื่อ
3. เพื่อพัฒนาซอฟต์แวร์สนับสนุนการอนุมัติสินเชื่อ โดยใช้ผลจากการทำเหมืองข้อมูล

1.3 ขอบเขตของการวิจัย

1. ศึกษารวบรวมโอนย้ายข้อมูลและวิเคราะห์โครงสร้างของข้อมูลลูกค้าที่เช่าซื้อรถกับทางบริษัทเอกชนนครหลวงลิซซิ่ง จำกัด ในช่วงปี 2546-2550
2. พัฒนาล้างข้อมูลลูกค้าเพื่อใช้ในการวิเคราะห์การอนุมัติสินเชื่อซึ่งประกอบด้วยขั้นตอนการทำความสะอาดข้อมูล การลดขนาดของข้อมูลให้อยู่ในมาตรฐานเดียวกัน
3. ศึกษาการทำงานของเครื่องมือการทำเหมืองข้อมูล (Data mining) เช่น Program Weka3.5.7 ศึกษาและเลือกใช้เทคนิคการทำเหมืองข้อมูล (Data mining) ที่เหมาะสมในการวิเคราะห์การอนุมัติสินเชื่อ
4. วิเคราะห์ผลการวิจัย
5. สรุปผลการวิจัย

1.4 ผลที่คาดว่าจะได้รับ

1. ได้ซอฟต์แวร์สนับสนุนการตัดสินใจการอนุมัติสินเชื่อ
2. ผู้บริหารมีเครื่องมือช่วยในการอนุมัติสินเชื่อ
3. ลดปัญหาที่อาจเกิดขึ้นเนื่องจากความผิดพลาดในการอนุมัติสินเชื่อ
4. ช่วยให้การตัดสินใจอนุมัติสินเชื่อได้สะดวกและรวดเร็วขึ้น
5. มีหลักเกณฑ์และแนวทางการตัดสินใจอนุมัติสินเชื่อ
6. ได้พัฒนาซอฟต์แวร์ช่วยสนับสนุนการตัดสินใจอนุมัติสินเชื่อที่มีหลักเกณฑ์และแนวทางการตัดสินใจอนุมัติสินเชื่อที่ทันสมัยมากขึ้น

1.5 นิยามศัพท์เฉพาะ

- ลูกค้า หมายถึง กลุ่มลูกค้าที่เช่าซื้อรถจักรยานยนต์กับบริษัท
- บริษัท หมายถึง บริษัทเอกชนนครหลวงลิซซิ่ง จำกัด
- สัญญา หมายถึง จำนวนบัญชีที่ลูกค้าทำสัญญากับบริษัท
- โปรแกรม หมายถึง โปรแกรมที่ใช้ในการทำเหมืองข้อมูล (Weka3.5.7)
- ฐานข้อมูล หมายถึง ฐานข้อมูลลูกค้าบริษัทเอกชนนครหลวงลิซซิ่ง จำกัด
- กฎ หมายถึง ผลที่ได้จากการทำเหมืองข้อมูล
- ค่าเป็นจริง หมายถึง ค่าของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นไปตามเงื่อนไข

ค่าไม่เป็นจริง หมายถึง ค่าของข้อมูลที่มีเงื่อนไขและผลลัพธ์ไม่สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นไปตามเงื่อนไข

ความถูกต้องของตัวแบบ หมายถึง การนำกลุ่มข้อมูลสำหรับตรวจสอบ (Validation Set) มาหาความสัมพันธ์และเปรียบเทียบค่าความเชื่อมั่นกับตัวแบบที่สร้างได้ที่เงื่อนไขเดียวกัน ถ้าค่าความเชื่อมั่นของข้อมูลตรวจสอบเท่ากับหรือมากกว่าแสดงว่าตัวแบบมีความถูกต้อง

บทที่ 2

แนวคิด ทฤษฎี และ ผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้องในการศึกษางานค้นคว้าอิสระนี้ ซึ่งได้แก่ การทำเหมืองข้อมูล เทคนิคการจำแนกข้อมูล โปรแกรม WEKA และงานวิจัยที่เกี่ยวข้อง

2.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ การค้นหาความสัมพันธ์และรูปแบบ (Pattern) ทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก การทำเหมืองข้อมูล จะทำการสำรวจและวิเคราะห์อย่างอัตโนมัติหรือกึ่งอัตโนมัติ ในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ (Rule) โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่าง ๆ ที่มีประโยชน์ในฐานข้อมูล การทำเหมืองข้อมูล จะเป็นการสังเคราะห์ข้อมูลอย่างละเอียดจากฐานข้อมูลขนาดใหญ่ หรืออาจวิเคราะห์มาจากรายการ Transaction โดยเรียนรู้ข้อมูลจากอดีต หรือปัจจุบันผลลัพธ์ที่ได้จากการสังเคราะห์ของ การทำเหมืองข้อมูลอาจจะเป็นข้อมูลแบบ Unknowns , Valid, หรือ Actionable ซึ่งความหมายของข้อมูลทั้ง 3 ประเภทนี้ มีดังนี้

1. ข้อมูลแบบ Unknowns เป็นข้อมูลที่ผู้ใช้งานไม่เคยรู้มาก่อนไม่ชัดเจน ไม่สามารถตั้งสมมติฐานล่วงหน้าว่าจะเป็นแบบใด เช่น ห้างสรรพสินค้าแห่งหนึ่งค้นพบพฤติกรรมของผู้บริโภคที่พ่อบ้านมักซื้อเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น ดังนั้นเป็นสัญญาณให้เจ้าของกิจการควรจะเตรียมสินค้าไว้เพื่อจำหน่าย ในขณะที่ห้างคู่แข่งอาจจะไม่รู้ข้อมูลเหล่านี้

2. ข้อมูลแบบ Valid เมื่อผู้ใช้เริ่มใช้เทคนิคของ Data Mining จะค้นพบสิ่งที่น่าสนใจตลอดเวลา แต่จะต้องพิจารณาด้วยว่าสิ่งนั้นถูกต้อง (Valid) หรือไม่ เช่น มักจะพบว่ามีความสัมพันธ์ของการซื้อสินค้า 2 อย่างเสมอ เมื่อจำนวนความหลากหลายของสินค้ามากขึ้น แต่ไม่ได้หมายความว่าต้องให้ห้างสรรพสินค้าเก็บสินค้าในคลังมากขึ้น เพราะข้อมูลที่ได้อาจเกิดความคลาดเคลื่อน เพราะฉะนั้นจะต้องทำการตรวจสอบความถูกต้อง (Validation and Checking) ของข้อมูลและวิเคราะห์ความถูกต้องอีกครั้ง

3. ข้อมูลแบบ Actionable: ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจ เพื่อสร้างความได้เปรียบในเชิงธุรกิจ บางครั้งข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้ว (เราช้าไป)

หรืออาจผิดกฎหมาย ซึ่งจะต้องมีวิจารณ์ในการใช้ด้วย บางทีข้อมูลดังกล่าวอาจไม่มีประโยชน์อะไร

2.1.1 วิวัฒนาการของ Data Mining

1. ปี ค.ศ 1960 : Data Collection มีการนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่นำเชื่อถือ เพื่อป้องกันการสูญหายได้เป็นอย่างดี
2. ปี 1980: Data Access มีการนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ระหว่างกัน เพื่อนำไปวิเคราะห์ และตัดสินใจอย่างมีประสิทธิภาพ
3. ปี 1990: Data Warehouse and Decision Support มีการนำข้อมูลมาเก็บลงในฐานข้อมูลขนาดใหญ่ ครอบคลุมการใช้งานทั้งหมดขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ
4. ปี 2000 : Data Mining นำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล โดยสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

2.1.2 วัตถุประสงค์ในการใช้ Data Mining

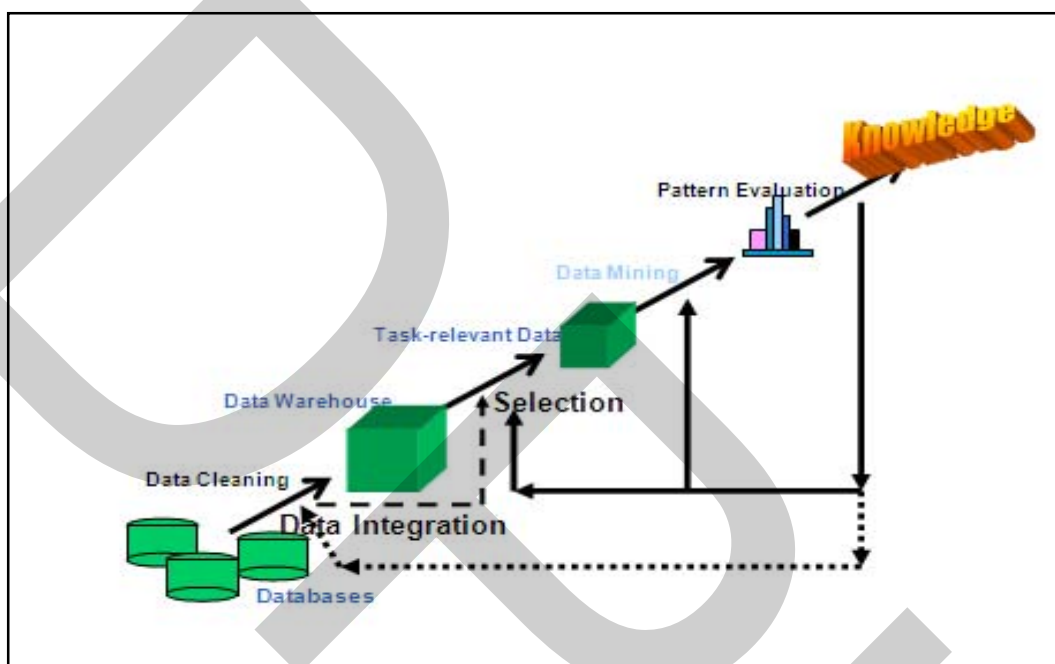
1. เพื่อการค้นพบองค์ความรู้ใหม่ในฐานข้อมูล (Knowledge discovery in databases)
2. เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่ (Knowledge extraction)
3. เพื่อจัดการกับข้อมูลในอดีต (Data archeology)
4. เพื่อสำรวจข้อมูล (Data exploration)
5. เพื่อค้นหา Pattern ของข้อมูลที่ซ่อนอยู่ (Data pattern processing)
6. เพื่อใช้ขุดเจาะข้อมูล (Data dredging)
7. เพื่อเก็บเกี่ยวผลประโยชน์ให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

2.1.3 เป้าหมายหลักของ Data Mining

คุณลักษณะและเป้าหมายหลักของ Data Mining คือ ใช้สกัดหรือค้นหา Pattern ของข้อมูลที่ฝังลึกและซ่อนเร้นอยู่ภายในฐานข้อมูลขนาดใหญ่ โดยใช้สถาปัตยกรรม Client-Server (Client/server architecture) ใช้เครื่องมือสมัยใหม่ที่สามารถแสดงผลแบบกราฟฟิก ผู้ใช้สามารถดูข้อมูลแบบเจาะลึก (data drills) และสามารถใช้อุปกรณ์ในการสอบถามข้อมูลได้อย่างง่ายดายโดยไม่ต้องอาศัยความชำนาญของ programmer บ่อยครั้งเราอาจค้นพบผลลัพธ์ที่เราไม่คาดหวังมาก่อน เครื่องมือจะทำให้เราใช้งานได้ง่าย ซึ่งเครื่องมือนอกจากจะแสดงผลกราฟิกได้แล้วยังรวม Spreadsheets เอาไว้ด้วย

2.1.4 กระบวนการของ Data Mining (A KDD Process)

เป็นกระบวนการในการค้นหาลักษณะแฝงของข้อมูล (Pattern) ที่ซ่อนอยู่ในฐานข้อมูล



ภาพที่ 2.1 กระบวนการของ Data Mining (KDD: Knowledge Discovery in Database)

ที่มา : <http://www.no-poor.com/dssandos/Chapter5-dss.htm>

2.1.5 ขั้นตอนของการค้นหาคำรู้ใหม่ (Steps of a KDD Process)

1. เรียนรู้และศึกษาเกี่ยวกับโปรแกรมที่จะใช้ (Learning the application domain)
2. คัดเลือกข้อมูล (data selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาทำ mining รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น
3. การกรองข้อมูลและประมวลผล (Data cleaning and preprocessing) ข้อมูลที่เก็บรวบรวมมามีจำนวนมากจะต้องนำมากรอง เพื่อเลือกข้อมูลที่ตรงประเด็น เพราะบางข้อมูลอาจจะไม่เป็นประโยชน์กับเรา ในขั้นตอนนี้จะได้มาซึ่งคุณภาพของข้อมูล ที่จะนำไปวิเคราะห์
4. การแปลงรูปแบบข้อมูล (Data reduction and transformation) ลดรูปและจัดข้อมูลให้อยู่ในรูปแบบ เดียวกัน มีรูปแบบ (Format) ที่เป็นมาตรฐาน และเหมาะสมที่จะนำไปใช้กับ Algorithm และแบบจำลองที่ใช้ทำ Data Mining

5. เลือก Functions ของ data mining เช่น summarization, classification, regression, association และ clustering เป็นต้น
6. เลือก Algorithm ของ data mining เป็นเทคนิคสำหรับการ Mine ข้อมูล
7. ทำการค้นหา Patterns ที่เราสนใจ
8. ประเมินผล Pattern และนำเสนอองค์ความรู้ ในขั้นตอนนี้จะเป็นการวิเคราะห์ผลลัพธ์ที่ได้ และแปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่ และนำเสนอ
9. ใช้องค์ความรู้ที่ค้นพบ (Use of discovered knowledge)

2.1.6 ชนิดขององค์ความรู้ที่ค้นพบ (Types of knowledge to be mined)

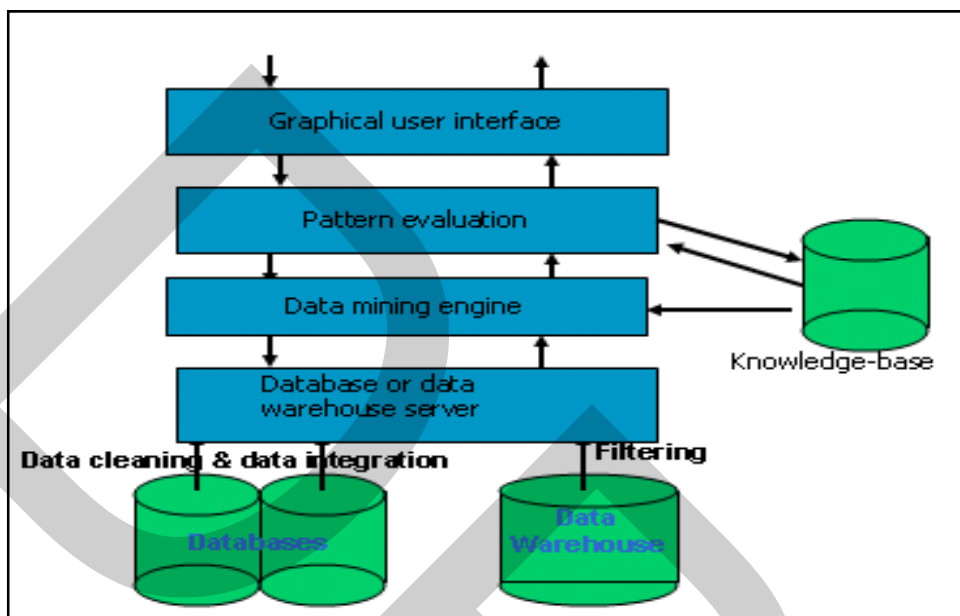
1. องค์ความรู้เกี่ยวกับคุณลักษณะของข้อมูล (Characterization) เช่น รู้ว่าคนที่สามารถเรียนต่อในระดับปริญญาเอกได้จะพิจารณาได้จากคุณลักษณะใด
2. องค์ความรู้เกี่ยวกับการจำแนกข้อมูล (Discrimination)
3. องค์ความรู้เกี่ยวกับความสัมพันธ์ของข้อมูล (Association) เช่น มีความสัมพันธ์ของการซื้อสินค้าพบว่า ถ้าลูกค้าซื้อบอว์น จะต้องซื้อเป๊ปซี่ตามมา
4. องค์ความรู้เกี่ยวกับการแยกประเภทข้อมูลและการพยากรณ์ (Classification/prediction)
5. องค์ความรู้เกี่ยวกับการจัดกลุ่มข้อมูล (Clustering)
6. องค์ความรู้เกี่ยวกับการวิเคราะห์ข้อมูลจากภายนอก (Outlier analysis)
7. องค์ความรู้เกี่ยวกับข้อมูลอื่น ๆ ในงานที่ค้นพบ (Other data mining tasks)

2.1.7 Data Mining และ Business Intelligence

Data Mining เป็นระดับการนำข้อมูลไปใช้ที่สูงกว่า Data Warehouse และ Data Mart นำเอาข้อมูลมาใช้ในการวิเคราะห์ให้เกิดประโยชน์สูงสุด เพื่อช่วยสนับสนุนการตัดสินใจแก่ฝ่ายบริหาร โดยอาศัยกฎเกณฑ์ต่าง ๆ ในการทำงาน

BI (Business Intelligence) คือ ข้อมูลสรุปที่สามารถนำมาช่วยในการตัดสินใจ หรือตอบคำถามในเชิงธุรกิจให้กับผู้บริหารได้ ดังนั้นระบบ BI ที่ดีจะต้องสามารถ นำเสนอข้อมูลสารสนเทศในเชิงภาพรวมของธุรกิจทั้งหมดขององค์กรได้ เพื่อให้เข้าใจความสามารถในการวิเคราะห์ข้อมูลสารสนเทศดี เนื่องจากสามารถวิเคราะห์และตอบคำถามของทั้งระบบธุรกิจได้ (<http://www.g-able.com/thai/solutions/g-biz/bis.htm>)

2.1.8 สถาปัตยกรรมของ Data Mining (Architecture of a Typical Data Mining System)



ภาพที่ 2.2 Architecture of a Typical Data Mining System

ประเภทของข้อมูลที่จะใช้ใน Data Mining

1. ข้อมูลที่มาจากฐานข้อมูลเชิงสัมพันธ์ (Relational databases)
2. ข้อมูลจากคลังข้อมูล (Data warehouses)
3. ข้อมูลจากฐานข้อมูลรายการปรับปรุง (Transactional databases)
4. ข้อมูลจากฐานข้อมูลพิเศษหรือที่เก็บข่าวสารพิเศษ ซึ่งได้แก่
 - ฐานข้อมูลเชิงวัตถุ
 - ข้อมูลเกี่ยวกับเวลา
 - ฐานข้อมูลข้อความ (Text databases) และฐานข้อมูลมัลติมีเดีย
 - ฐานข้อมูลแบบเก่าในอดีตหรือข้อมูลที่มาจากต่างฐานข้อมูลกัน
 - ข้อมูลจากแหล่ง WWW

2.1.9 Data Mining Functionalities (Data Mining Task)

งานของ Data Mining สามารถทำงานในการขุดค้นข้อมูล ดังนี้

1. การวิเคราะห์คุณสมบัติและการแยกแยะข้อมูล (Characterization and discrimination)

2. การหาความสัมพันธ์ของข้อมูล (Association)

3. การจัดหมวดหมู่และการวิเคราะห์การถดถอย (Classification/ Regression)

การจัดหมวดหมู่ (Classification) ตัวอย่างของการจัดหมวดหมู่ ที่นำมาใช้กับงานด้านธุรกิจ เช่น มีนักวิเคราะห์ขององค์กรแห่งหนึ่งต้องการรู้เหตุผลว่า “ทำไมลูกค้าบางกลุ่มถึงยังคงซื้อผลิตภัณฑ์ยี่ห้อดีต่อยี่ห้อสินค้า (Brand Loyalty) ขององค์กร และขณะเดียวกันก็มีลูกค้าอีกกลุ่มที่เปลี่ยนใจไปหาคู่แข่ง “ ในการหาคำตอบนี้ นักวิเคราะห์ต้องทำนายลักษณะนิสัยของลูกค้าที่องค์กรอาจต้องเสียไปให้กับคู่แข่ง ดังนั้นเมื่อมีเป้าหมายคือ “อยากทราบเหตุผล” นักวิเคราะห์สามารถนำข้อมูลการซื้อสินค้า ของลูกค้าในอดีตมาทดลองกับแบบจำลองเพื่อวิเคราะห์ผลว่าทำไมลูกค้าบางกลุ่มซื้อผลิตภัณฑ์ บางกลุ่มไม่ซื้อผลิตภัณฑ์

จุดประสงค์คือ : ต้องการศึกษา “ข้อมูลลูกค้า” โดยสมมติให้ข้อมูลลูกค้ามีคุณสมบัติที่เกี่ยวข้อง ดังนี้

ตารางที่ 2.1 แสดงตารางข้อมูลเกี่ยวกับลูกค้า

Field	Data Type	Value	Description
Cus_id	Int	Unique	รหัสลูกค้า
Time	Int	Integer	ระยะเวลาที่ลูกค้าอยู่กับองค์กร
Trend	Text	เพิ่มขึ้น, คงที่, ลดลง	ตัวบ่งชี้แนวโน้มการใช้สินค้า 6 เดือนล่าสุด
Status	Text	สูง, กลาง, ต่ำ, ไม่ทราบ	การสำรวจผลความพอใจของลูกค้า
Cus_type	Text	ซื้อผลิตภัณฑ์, ไม่ซื้อผลิตภัณฑ์	ลูกค้ายังคงอยู่กับองค์กรหรือเสียไปให้คู่แข่ง

คำตอบที่เราต้องการ (Output) คือรหัสลูกค้า (Cus_type) ถือเป็นตัวแปรตาม(Dependent vairable) ซึ่งผลของตัวแปรตามจะขึ้นอยู่กับตัวแปรอิสระ ((Independent vairable) ในที่นี้คือฟิลด์ Time, Trend และ Status มีหลายเทคนิคของ Data Mining ที่ใช้ในการแก้ปัญหาแบบ Classification แต่ละเทคนิคก็จะมีหลาย Algorithm ให้เลือกและแต่ละ Algorithm จะให้ผลลัพธ์ที่ต่างกัน ซึ่งปัญหาประเภทนี้จะให้ผลลัพธ์เป็นค่าที่แน่นอน เช่น อาจจะได้คำตอบเป็น (Yes, No) หรือ (High, Medium, Low) เป็นต้น

เทคนิคของ Data Mining ที่ใช้ ในการแก้ปัญหาแบบ Classification ได้แก่

- Decision Tree
- Neural Networks
- Naïve-Bayes
- K-nearest neighbor (K-NN)

การวิเคราะห์การถดถอย (Regression)

ปัญหาแบบ Regression จะเหมือนกับแบบ Classification ต่างกันตรงที่ผลลัพธ์ที่ได้จาก Regression เป็นค่าแน่นอน ที่ไม่จำกัด จะเป็นค่าอะไรก็ได้ เช่น แบบจำลองทำนายว่า นาย B จะตอบรับข้อเสนอของบริษัท ถ้านาย B ได้รับผลกำไร 1,000 บาท (1,000 เป็นคำตอบเฉพาะที่แน่นอน แต่ไม่จำกัด ซึ่งตัวเลขอาจจะเป็นค่าอื่นไปได้เรื่อย ๆ ต่างจากคำตอบแบบ Yes, No)

4. การวิเคราะห์การรวมกลุ่ม หรือ การแบ่งแยกข้อมูล (Cluster analysis/ Segmentation) การวิเคราะห์การรวมกลุ่ม (Clustering) เป็นการรวมกลุ่มข้อมูลที่มีลักษณะเหมือนกัน รูปแบบหรือแนวโน้มที่จะเหมือนกัน การใช้เทคนิค Clustering จะไม่มีผลลัพธ์ (Output) ไม่มีตัวแปรอิสระ (Independent Variable) ไม่มีการจัดโครงสร้างของวัตถุ เราจะเรียกเทคนิคของ Clustering ว่าเป็นแบบเรียนรู้ข้อมูลโดยไม่ต้องอาศัยครูสอน (Unsupervised Learning) การทำ Clustering จะทำบนพื้นฐานของข้อมูลในอดีต

Ex : องค์กรต้องการทราบความเหมือนที่มีในกลุ่มของลูกค้าของตน เพื่อที่จะให้เข้าใจลักษณะเฉพาะของลูกค้ากลุ่มเป้าหมาย และสร้างกลุ่มของลูกค้าเพื่อที่องค์กรจะได้สามารถขายสินค้าได้ในอนาคต องค์กรจะทำการแยกกลุ่มของข้อมูลลูกค้าออกเป็นกลุ่ม ๆ (หาส่วนที่เป็น Intersection และ Union)

เทคนิคของ Data Mining เพื่อแก้ปัญหาแบบ Clustering คือวิธี Demographic Clustering กับ Neural Clustering

5. การประเมินและการพยากรณ์ (Estimation/Prediction) การประเมิน (Estimation) เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติที่ชัดเจนได้ ใช้จัดการกับค่าที่มีผลแบบต่อเนื่อง เช่น ใช้ประเมินรายได้ของครอบครัว ประเมินความสูงของบุคคลในครอบครัว ประเมินจำนวนเด็ก ๆ ในครอบครัว

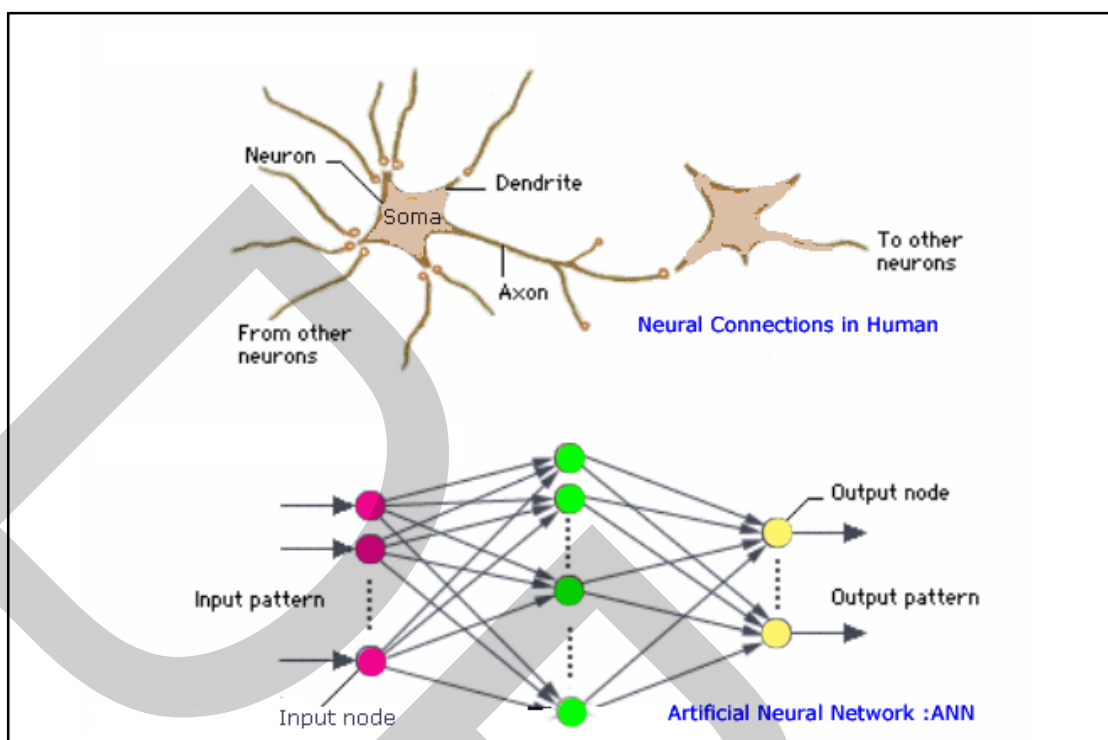
การพยากรณ์ (Prediction) จะเหมือนกับ Classification และ Estimation ต่างกันตรงที่ Record ถูกแยกจัดลำดับในการทำนายค่าในอนาคต และนำข้อมูลในอดีตมาสร้างเป็นแบบจำลอง ใช้ทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น การทำนายว่าลูกค้ากลุ่มใด ที่องค์กรจะสูญเสียไปในอีก 6 เดือนข้างหน้า หรือการทำนายยอดซื้อของลูกค้าจะเป็นเท่าใด ถ้าบริษัทลดราคาสินค้าลง 10%

6. การบรรยายและการแสดงภาพของข้อมูล (Description / Visualization) การบรรยาย (Description) เป็นการหาคำอธิบายถึงสิ่งที่จะเกิดขึ้น โดยอาศัยข้อมูลจากฐานข้อมูล เช่น กลุ่มคนที่มีการศึกษาหรือรายได้ต่ำ จะเลือกนักการเมืองที่มีนโยบายทุนนิยม มากกว่ากลุ่มคนชั้นกลาง การแสดงภาพของข้อมูล (Visualization) เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิค หรืออาจนำเสนอในรูปแบบ 2 มิติ สร้างรายละเอียดในการนำเสนอให้เข้าใจมากยิ่งขึ้น เช่น องค์กรต้องการหาสถานที่ในขยายสาขาใหม่ที่อยู่ในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้แผนที่ Plot ที่ตั้งขององค์กรคู่แข่งที่มีสาขาอยู่ในเขตนั้น เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด

2.1.10 เครื่องมือและเทคโนโลยีของ Data Mining (Data Mining Tools and Technologies)

1. Neural Network เป็นแนวคิดให้คอมพิวเตอร์ทำงานสมองของมนุษย์ เปลี่ยนตัวเองจากการประมวลผลตามลำดับ (Sequential Processing) ให้เป็นการประมวลผลแบบคู่ขนานได้ (Parallel Processing) มีลักษณะการทำงานโดยแต่ละ Process จะรับ Input เข้าไปคำนวณ และสร้าง Output ออกมาในลักษณะที่ไม่ใช่การทำงานแบบเชิงเส้นตรง เพราะ Input แต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงกันนี้ จะถูกนำมาเปรียบเทียบกับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็จะนำไปสู่การปรับค่าหรือน้ำหนัก (weight) ของค่าที่ใส่ไว้ให้แต่ละ Input

Neural Network เป็นการสร้างแบบจำลอง ที่เลียนแบบการทำงานของสมองมนุษย์ มีโครงสร้างเป็นกลุ่มของ Node ที่เชื่อมโยงถึงกันในแต่ละ Layer คือ Input layer, Hidden layer, output layer



ภาพที่ 2.3 ตัวอย่างของ Neural Network

2. Decision Trees เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ (Decision Trees) ซึ่ง Decision Trees จะมีการทำงานแบบ Supervised Learning (คือการเรียนรู้แบบมีครูสอน) สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ก่อนล่วงหน้า เรียกว่า Training set ได้อัตโนมัติ และ พยากรณ์ กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ ได้ด้วยรูปแบบของ Tree โครงสร้างประกอบด้วย Root Node, Child และ Leaf Node

3. Memory Based Reasoning (MBR) เปรียบเหมือนกับประสบการณ์ การเรียนรู้ของมนุษย์ ซึ่งอาศัยการสังเกตที่เกิดขึ้น แล้วสร้างรูปแบบของสิ่งนั้นขึ้นมา เราใช้ MBR เพื่อวิเคราะห์ฐานข้อมูลที่มีอยู่ และกำหนดลักษณะพิเศษของข้อมูลที่อยู่ therein ซึ่งข้อมูลจะต้องมีลักษณะที่สมบูรณ์ การสังเกตจึงจะสมบูรณ์และทำนายผลได้แม่นยำยิ่งขึ้น แบบจำลองจะถูกบอกคำตอบที่ถูกต้อง มีการเก็บคำตอบสำหรับแก้ปัญหาไว้ก่อนล่วงหน้าแล้ว (Supervised Learning)

4. Cluster Detection คือจะแบ่งฐานข้อมูลออกเป็นส่วน ๆ เรียกว่า Segment (กลุ่ม Record ที่มีลักษณะคล้ายกัน) ส่วน Record ที่ต่างกันก็จะอยู่นอก Segment, Cluster Detection ถูกใช้เพื่อค้นหากลุ่มย่อย (Sub Group) ที่เหมือน ๆ กันในฐานข้อมูล เพื่อที่จะเพิ่มความถูกต้องในการวิเคราะห์ และสามารถมุ่งไปยังกลุ่มเป้าหมายได้ถูกต้อง

5. Link Analysis มุ่งเน้นทำงานบน Record ที่มีความสัมพันธ์กัน หรือเรียกว่า Association เทคนิคนี้จะมุ่งไปที่รูปแบบการซื้อหรือเหตุการณ์ที่เกิดขึ้นเป็นลำดับ มีอยู่ 3 เทคนิค คือ

- Association Discovery ใช้วิเคราะห์การซื้อขายสินค้าในรายการเดียวกัน ศึกษาความสัมพันธ์อย่างใกล้ชิดที่ถูกปิดซ่อนอยู่ของสินค้า ซึ่งสินค้าเหล่านั้นอาจมีแนวโน้มที่จะถูกซื้อควบคู่กันไป การวิเคราะห์แบบนี้เรียกว่า Market Basket Analysis คือ รายการทั้งหมดที่ถูกค้าซื้อต่อครั้งที่ Super market การวิเคราะห์นี้สามารถนำมาใช้ประโยชน์ในการตัดสินใจ เช่น การเตรียมสินค้าคงเหลือ การวางแผนจัดชั้นวางสินค้า การทำ Mailing list สำหรับ Direct Mail การวางแผนเพื่อจัด Promotion สนับสนุนการขาย ตัวอย่างของ Association เช่น 75% ของผู้ซื้อน้ำอัดลมจะซื้อข้าวโพดคั่วด้วย

- Sequential Pattern Discovery ถูกใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าของลูกค้า มีจุดหมายที่จะเข้าใจพฤติกรรมของการซื้อสินค้าของลูกค้าในลักษณะ logn term เช่น ผู้ขายอาจพบว่าลูกค้าที่ซื้อ TV มีแนวโน้มที่จะซื้อ VDO ในเวลาต่อมา

- Similar Time Sequence Discovery ค้นหาความเกี่ยวเนื่องกันระหว่างข้อมูล 2 กลุ่ม ซึ่งขึ้นต่อกันทางด้านเวลาโดยมีรูปแบบการเคลื่อนที่เหมือนกัน ผู้ขายสินค้ามักใช้เพื่อดูแนวโน้มเพื่อเตรียม Stock เช่น เมื่อไรก็ตามที่ยอดขายสินค้าน้ำอัดลมสูงขึ้น ยอดขายมันฝรั่งจะสูงขึ้นตาม

- Genetic Algorithm (GA) เปรียบเสมือนเป็นการสร้างพันธุ์กรรมที่ดีที่สุด บนขั้นตอนของวิวัฒนาการทางชีวภาพ แนวคิดหลักคือ เมื่อเวลาผ่านไป วิวัฒนาการของเซลล์ชีวิตจะเลือกสายพันธุ์ที่ดีที่สุด “Fittest Species” GA มีความสามารถในการทำงานแบบรวมกลุ่มเข้าด้วยกัน เช่น มีการแบ่งกลุ่มและจัดรวมกลุ่มข้อมูลเป็น 3 ชุด ขั้นตอนการทำงานของ GA เริ่มจาก

- จับกลุ่มข้อมูลเป็นกลุ่ม ๆ ด้วยการสุ่มเดา เปรียบเหมือนกลุ่ม 3 กลุ่มนี้เป็นเซลล์ของสิ่งมีชีวิต GA จะมี Fittest Function ที่จะบอกว่ากลุ่มข้อมูลใดเหมาะกับกลุ่ม ๆ ใดโดย Fittest Function จะเป็นตัวบ่งชี้ว่าข้อมูลเหมาะกับกลุ่มมากกว่าข้อมูลอื่น ๆ

- GA จะมี Operator ซึ่งยอมให้มีการเลียนแบบและแก้ไขลักษณะของกลุ่มข้อมูล Operator จะจำลองหน้าที่ของชีวิตที่ถูกพบในธรรมชาติ คือ มีการแพร่พันธุ์ จับคู่ผสมพันธุ์ และเปลี่ยนรูปร่างตามต้นแบบของพันธุ์กรรม เปรียบกับข้อมูลถ้ามีข้อมูลใดในกลุ่มถูกพบว่าตรงกับคุณสมบัติของ Fittest function แล้ว มันจะคงอยู่และถูกถ่ายเข้าไปในกลุ่มนั้น แต่ถ้าไม่ตรงกับคุณสมบัติ ก็ยังมีโอกาสที่จะถ่ายข้ามไปยังกลุ่มอื่นได้

- Rule Induction ดึงเอาชุดกฎเกณฑ์ต่าง ๆ มาสร้างเป็นเงื่อนไขหรือกรณี วิธีการของ Rule Induction จะสร้างชุดของกฎที่เป็นอิสระ ซึ่งไม่จำเป็นต้องอยู่ในรูปแบบของโครงสร้างต้นไม้

- K-nearest neighbor (K-NN) จะใช้วิธีในการจัดแบ่งคลาส โดยจะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวนของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไขหรือกรณีต่าง ๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกับมันมากที่สุด

K-NN ก่อนข้างใช้ปริมาณงานในการคำนวณสูงมากบนคอมพิวเตอร์ เพราะเวลาสำหรับการคำนวณจะเพิ่มขึ้นแบบแฟกทอเรียล ตามจำนวนจุดทั้งหมด เทคนิคของ K-NN จะมีการคำนวณเกิดขึ้นทุกครั้งที่มีกรณีใหม่ ๆ เกิดขึ้น ดังนั้นถ้าจะให้เทคนิคแบบ K-NN ทำงานได้เร็ว ข้อมูลที่ใช้อยู่ควรเก็บอยู่ใน MBR (Memory-Based Reasoning)

- Association and Sequence Detection

Association ใช้หาความสัมพันธ์ที่เกิดขึ้นระหว่างกลุ่มของข้อมูล (Item) ต่าง ๆ ใช้ใน Market

-basket analysis อาจใช้เพื่อวิเคราะห์การสั่งซื้อสินค้า Sequence Detection เหมือนกับ Association แต่จะนำเหตุการณ์ที่เกิดขึ้น และเพิ่มตัวแปรด้านเวลาเข้ามาเกี่ยวข้องด้วย เพื่อใช้วิเคราะห์พฤติกรรมของข้อมูลการเขียนความสัมพันธ์ (Association) เช่น $A \rightarrow B$ หมายถึง A เป็นเหตุการณ์ที่เกิดขึ้นก่อน (Antecedent) หรือ LHS (Left-Hand Side) B เป็นผลของเหตุการณ์ (Consequent) หรือ RHS (Right- Hand Side) เช่น ในกฎของความสัมพันธ์ “ถ้าซื้อค้อน แล้วจะซื้อตะปู” เหตุการณ์ที่เกิดขึ้นก่อนคือ ค้อน เกิดหลังคือ ตะปู

- Logic Regression เป็นการวิเคราะห์ความถดถอยแบบเส้นตรงทั่วไป ใช้ในการพยากรณ์ผลลัพธ์ของ 2 ตัวแปร เช่น Yes/No , 0/1 แต่เนื่องจากตัวแปรตาม (Dependent Variable) มีค่าเพียง 2 อย่างเท่านั้น จึงไม่สามารถสร้างแบบจำลอง (Model) ได้สำหรับการวิเคราะห์แบบ Logic Regression ดังนั้นแทนที่จะทำการพยากรณ์โดยอาศัยเพียงค่าของตัวแปรตามที่ได้ เราจะสร้าง Model โดยอาศัย Algorithm ของความน่าจะเป็นของการเกิดเหตุการณ์ เราเรียก Algorithm นี้ว่า Log Odds หรือ Logic Transtromation

อัตราส่วนความน่าจะเป็น :
$$\frac{\text{ความน่าจะเป็นที่จะเกิดเหตุการณ์}}{\text{ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์}}$$

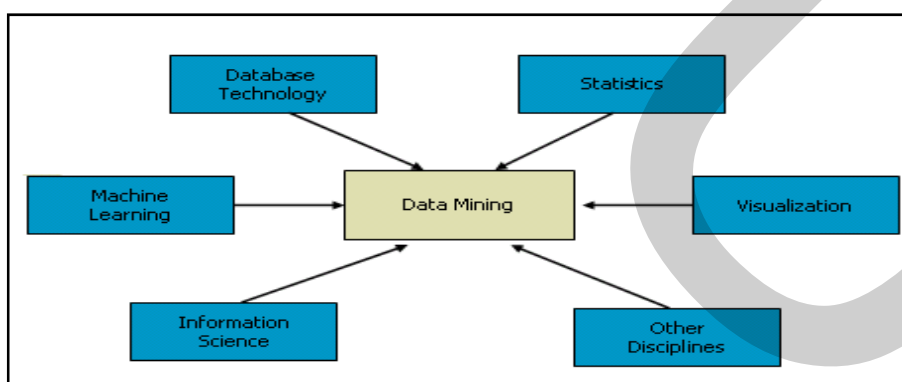
- Discriminant Analysis : เป็นวิธีการทางคณิตศาสตร์ที่เก่าแก่วิธีหนึ่งซึ่งใช้ในการจำแนกและวิเคราะห์ วิธีนี้ได้รับการเผยแพร่ครั้งแรกในปี 1936 โดย R.A Fisher เพื่อแยกต้น Iris ออกเป็น 3 พันธุ์ วิธีการนี้ทำให้ค้นพบต้นไม้ประเภทอื่น ๆ อีกมาก ผลลัพธ์ที่ได้จากแบบจำลอง

ชนิดนี้ง่ายต่อการทำความเข้าใจ เพราะผู้ใช้งานทั่ว ๆ ไปก็สามารถพิจารณาได้ว่าผลลัพธ์จะอยู่ทางด้านใดของเส้นทางในแบบจำลอง การเรียนรู้สามารถทำได้ง่าย วิธีการที่ใช้มีความไวต่อรูปแบบของข้อมูล วิธีนี้ถูกนำมาใช้ในทางการแพทย์ สังคมวิทยา และชีววิทยา แต่ไม่เป็นที่นิยมในการทำ Data Mining

- Generalized Additive Models (GAM) : พัฒนามาจาก Linear Regression และ Logistic Regression มีการตั้งสมมติฐานว่า Model สามารถเขียนออกมาได้ในรูปของผลรวมของ Possibly Non-Linear Function GAM สามารถใช้ได้กับปัญหาแบบ Regression และ Classification

GAM จะใช้ความสามารถของคอมพิวเตอร์ในการค้นหารูปแบบของ Function ที่ให้ Curve ที่เหมาะสม ทำการรวมค่าความสัมพันธ์ต่าง ๆ เข้าด้วยกัน แทนที่จะใช้ Parameter จำนวนมาก เหมือนที่ Neural Network ใช้ แต่ GAM จะก้าวไปเหนือกว่านั้นอีกขั้นหนึ่ง GAM จะประเมินค่าของ Output ในแต่ละ Input เช่นเดียวกับ Neural Network GAM จะสร้างเส้นโค้งขึ้นมาอย่างอัตโนมัติ โดยอาศัยข้อมูลที่มีอยู่

- Multivariate Adaptive Regression Splines (MARS) : ถูกคิดค้นเมื่อกลางทศวรรษที่ 80 โดย Jerome H. Friedman หนึ่งในผู้คิดค้น CART MARS สามารถที่จะค้นหาและแสดงรายการตัวแปรอิสระที่มีความสำคัญสูงสุดเช่นเดียวกับปฏิสัมพันธ์ระหว่างตัวแปรอิสระ และ MARS สามารถ Plot จุดแสดงความเป็นอิสระของแต่ละตัวแปรอิสระ ออกมาได้ ผลลัพธ์ที่ได้ก็คือ Non-Linear Step-wise regression tools ความสามารถที่หลากหลายของ Data Mining



ภาพที่ 2.4 Data Mining: Confluence of Multiple Disciplines

2.1.11 การประยุกต์ใช้ Data Mining กับงานด้านธุรกิจ

สามารถนำเทคนิคของ Data Mining ไปวิเคราะห์ข้อมูลในฐานข้อมูล เพื่อนำข้อมูลที่ได้ไปใช้ประโยชน์ในงานด้านต่าง ๆ ดังต่อไปนี้

1. งานด้านการตลาด (Marketing) เช่น การทำ Promotion ส่งเสริมการขาย
2. งานด้านธนาคารและการเงิน (Banking / Financial Analysis) เช่น ใช้ในการวิเคราะห์ การให้สินเชื่อแก่ลูกค้า การจัดทำ Package ในการกู้ยืม การทำนายอัตราดอกเบี้ย การแบ่งกลุ่มลูกค้าเพื่อหาเป้าหมายทางการตลาด (ลูกค้าชั้นดี)
3. งานด้านการขายปลีก (Retailing and sales) เป็นงานที่มีการเก็บข้อมูลจำนวนมาก ประยุกต์ใช้เพื่อหากลยุทธ์ ทำให้เกิดการได้เปรียบคู่แข่งทางการค้าในการหาลักษณะการซื้อของลูกค้า ความสัมพันธ์ของการซื้อกับช่วงเวลา ความสัมพันธ์ระหว่างตัวสินค้า และการวิเคราะห์ประสิทธิภาพของการโฆษณา เป็นต้น ช่วยให้สามารถหาวิธีการตอบสนองความต้องการของลูกค้าได้มากที่สุด และอาจหมายถึงส่วนแบ่งทางการตลาดที่เพิ่มขึ้นนั่นเอง
4. งานด้านการวางแผนในการผลิตสินค้า (Manufacturing and production) เช่น การพยากรณ์ยอดจำนวนการผลิตสินค้าเพื่อให้ได้กำไรมากที่สุด
5. งานด้านนายหน้าและความปลอดภัยด้านการค้า (Brokerage and securities trading) เช่น การพัฒนาวิธีการเพื่อสร้างความเชื่อมั่นในเรื่องความปลอดภัยของข้อมูล ในขณะที่มีการพัฒนาวิธีการเข้าถึงข้อมูล และการ Mining ให้สะดวกต่อการใช้งานมากขึ้น
6. งานด้านชีวการแพทย์และวิเคราะห์ DNA (Biomedical and DNA Analysis) เช่น การวิเคราะห์รูปแบบการจัดเรียงตัวของหน่วยพันธุกรรม เพื่อหาสาเหตุความผิดปกติที่ทำให้เกิดโรครวมไปถึงด้านการวินิจฉัยโรค การป้องกัน และการรักษา

2.1.12 Intelligent Data Mining

ใช้ Intelligent Data Mining เพื่อการค้นพบข้อมูลและข่าวสารภายในคลังข้อมูล (Data warehouses) ที่ซึ่งการสอบถามและการออกรายงาน (Reports) นั้นจะไม่แสดงผลออกมา เช่น การค้นหา Patterns ในข้อมูลและลงความเห็นตามกฎ ที่เราได้กำหนดไว้ การใช้ Patterns และ Rules ในการแนะนำทางการตัดสินใจและการทำนาย ซึ่งข้อมูลที่จัดเป็น Intelligent Data Mining ได้แก่ ข้อมูลข่าวสาร 5 ประเภท ต่อไปนี้

- 1) ข้อมูลความสัมพันธ์ (association)
- 2) ข้อมูลการจัดลำดับ (sequences)
- 3) ข้อมูลการแยกประเภทหรือจัดหมวดหมู่ (classifications)
- 4) ข้อมูลการแบ่งกลุ่มหรือจัดกลุ่ม (clusters)
- 5) ข้อมูลการทำนายหรือพยากรณ์ (forecasting)

2.1.13 เครื่องมือหลักที่ใช้ใน Intelligent Data Mining

สามารถใช้เครื่องมือหรือเทคนิคต่อไปนี้ในการขุดค้นข้อมูลใน Intelligent data mining

1. Case-based Reasoning
2. Neural Computing
3. Intelligent Agents
4. เครื่องมือชนิดอื่น ๆ (Other Tools) ได้แก่
 - Decision trees
 - Rule induction
 - Data visualization

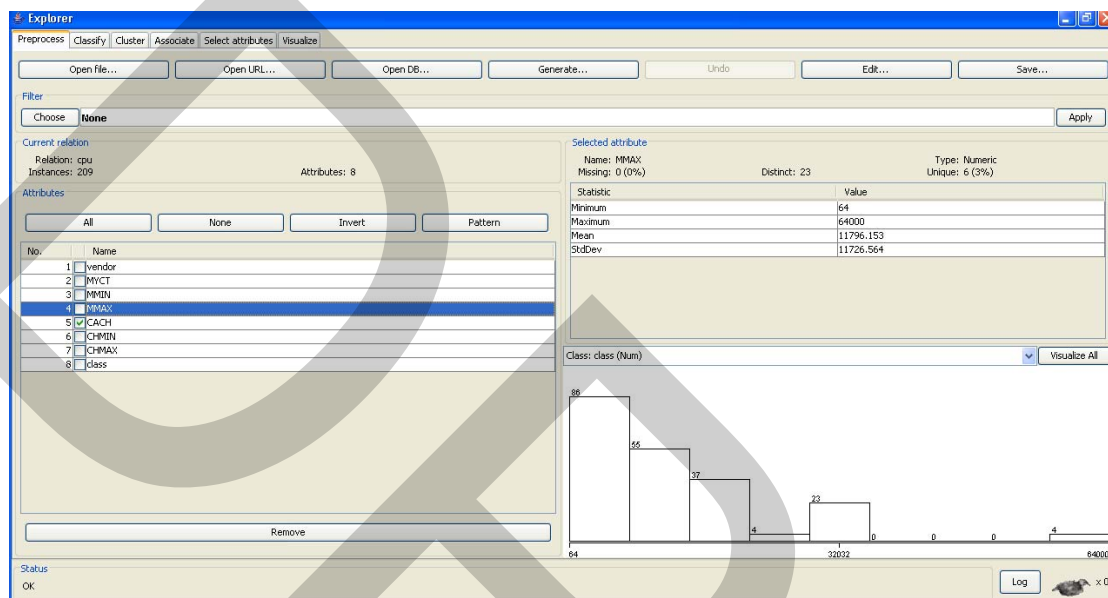
2.2 เทคนิคการจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูล (Data Classification) ฤชณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวิทย์ รักธรรมานนท์ (2544:135) กล่าวว่า การจำแนกประเภทของข้อมูลเป็นกระบวนการสร้าง โมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (Training Data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้คือ โมเดลจำแนกประเภทข้อมูล (Classifier Model) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลสอนระบบ มาเป็นข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมา เปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้องและปรับปรุงโมเดลจนกว่าจะได้ ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามาจะนำข้อมูลผ่าน โมเดลเพื่อ ทำนายกลุ่มของข้อมูล ซึ่งโมเดลที่ได้อาจแสดงในรูปของ โมเดลการพยากรณ์ข้อมูล (Prediction Model) ซึ่งเป็นกระบวนการสร้างโมเดลเพื่อทำนายค่าที่ต้องการจากข้อมูลที่มีอยู่ โดยมีกระบวนการ สร้างโมเดลคล้ายกับการจำแนกประเภทข้อมูล แต่ต่างกันตรงที่การพยากรณ์ข้อมูลนี้เป็นการ พยากรณ์ค่าที่ต้องการออกมาเป็นตัวเลข ตัวอย่างเช่น หายอดขายของเดือนถัดไป จากข้อมูลการ ขายทั้งหมดที่ผ่านมา (Berry, Michel J.A. and Linnoff 2000:7) หรือ ทำนายเกรดเฉลี่ยของนักเรียน ในปีการศึกษาหน้า จากข้อมูลการลงทะเบียนของนิสิตทั้งหมด (Han, Jiawei and Kamber 2000: 19) เป็นต้น

2.3 โปรแกรม WEKA

Weka ย่อมาจาก Waikato Environment for Knowledge Analysis เป็นซอฟต์แวร์ฟรีที่ แจกจ่ายภายใต้ GPL license โดยทั้งหมดนี้สร้างด้วยภาษาจาวาโปรแกรมจาวาซึ่งประกอบไปด้วย

อุปกรณ์การวิเคราะห์ข้อมูลและเพื่อนำไปใช้โมเดลข้อมูลได้ เน้นการเรียนรู้ด้วยเครื่อง (Machine learning) กับ การทำเหมืองข้อมูล (Data mining) มีโมดูลย่อยสำหรับจัดการข้อมูล และแสดงผล ใช้ GUI และคำสั่ง ในการสั่งให้ซอฟต์แวร์ประมวลผล



ภาพที่ 2.5 แสดงหน้าจอการทำงานของโปรแกรม Weka

โปรแกรมหลักของซอฟต์แวร์ Weka

- Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งการทำงานผ่านกาพิมพ์
- Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
- Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
- Knowledge Flow เป็นโปรแกรมออกแบบผังการไหลของความรู้
- Arff Viewer เป็นโปรแกรมที่ใช้สำหรับแก้ไขแฟ้มประเภท Arff
- Log เป็นโปรแกรมที่ใช้อ่านข้อความบันทึกเก็บระหว่างการทำงาน

ในการวิจัยนี้จะใช้เฉพาะโปรแกรม Explorer ในการศึกษาวิเคราะห์และทำเหมืองข้อมูลซึ่งมีหน้าจการทำงานดังรูปที่ 2.5

เมนูหลักของ Explorer

- Preprocess การเตรียมข้อมูล
- Classify รวมโมดูลการทำเหมืองข้อมูลแบบจำแนกกลุ่ม

- Cluster รวมโมดูลการทำเหมืองข้อมูลแบบการจัดกลุ่ม
- Associate รวมโมดูลการทำเหมืองข้อมูลแบบกฎเชื่อมโยง
- Select attributes รวมโมดูลสำหรับการวิเคราะห์ความเกี่ยวพันของลักษณะประจำ
- Visualize นำเสนอข้อมูลด้วยภาพนามธรรมสองมิติ

ประเภทของแฟ้มข้อมูลที่น่าเข้าข้อมูล ได้แก่

- แฟ้มข้อมูลในรูปแบบ arff, csv, C45
- ในกรณีแฟ้มข้อมูลอยู่ในเครือข่ายผู้ใช้สามารถเรียกใช้โดยอาศัย URL
- หรืออาจใช้ข้อมูลที่อยู่ในฐานข้อมูลที่เชื่อมโยงผ่าน JDBC

WEKA เป็นโปรแกรมที่ได้พัฒนาขึ้นบนพื้นฐานของภาษาจาวา (Java) สามารถรัน (run) ได้หลายระบบปฏิบัติการ และสามารถพัฒนาต่อยอดโปรแกรมได้ เป็นเครื่องมือที่ใช้ทำงานในด้านการทำค้ำไม้มนั้่ง ที่รวบรวมแนวคิดอัลกอริทึมมากมาย ซึ่งอัลกอริทึมสามารถเลือกใช้งานโดยตรงได้ 2 ทางคือ จากชุดเครื่องมือที่มีอัลกอริทึมมาให้ หรือเลือกใช้อัลกอริทึมที่ได้เขียนเป็นโปรแกรมลงไปเป็นชุดเครื่องมือเพิ่มเติม และ ชุดเครื่องมือมีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูล ได้แก่ Pre-Processing, Classification, Regression, Clustering, Association rules, Selection และ Visualization ข้อดีของโปรแกรม Weka คือ มีอัลกอริทึมที่รู้จักกันดีของการทำค้ำไม้มนั้่งให้เลือกใช้อย่างครบถ้วน และสามารถเขียนฟังก์ชันเพิ่มเข้าไปในโปรแกรมได้ เทคนิคที่ใช้ในการทดสอบในงานวิจัยครั้งนี้ ได้แก่

2.3.1 Decision Tree C4.5 เป็นการจัดกลุ่มในลักษณะของต้นไม้การตัดสินใจ ที่มีกิ่งต้นไม้จำนวนมาก โดยค่าของข้อมูลของแอทริบิวต์ที่มีค่าต่อเนื่องจะใช้เครื่องหมาย “<”, “<=”, “>” และ “>” ในการพิจารณา ส่วนข้อมูลที่เป็นค่าไม่ต่อเนื่องจะใช้เครื่องหมาย “=” และใช้เงื่อนไข “AND” ในการเชื่อมโยงระหว่างแอทริบิวต์ จะเชื่อมโยงไปจนถึงแอทริบิวต์สุดท้ายที่จะแบ่งกลุ่มได้

2.3.2 Decision Tree: Partial Rules เป็นการจัดกลุ่มในลักษณะของกฎที่นำเอาข้อมูลในแต่ละแอทริบิวต์มาเป็นเงื่อนไขในการตัดสินใจลักษณะของข้อมูลที่ใกล้เคียงกันให้มาอยู่ในกลุ่มเดียวกันโดยค่าของข้อมูลของแอทริบิวต์ที่มีค่าต่อเนื่องจะใช้เครื่องหมาย “<”, “<=”, “>” และ “>” ในการพิจารณา ส่วนข้อมูลที่เป็นค่าไม่ต่อเนื่องจะใช้เครื่องหมาย “=” และใช้เงื่อนไข “AND” ในการเชื่อมโยงระหว่างแอทริบิวต์ จะเชื่อมโยงไปจนถึงแอทริบิวต์สุดท้ายที่จะแบ่งกลุ่มได้

2.3.3 Decision Table เป็นการจัดกลุ่มในลักษณะตารางการตัดสินใจโดยการนำผลที่ได้จากการเรียนรู้จากข้อมูลที่น่าเข้ามาที่มีลักษณะของข้อมูลที่ใกล้เคียงกันให้มาอยู่ในกลุ่มเดียวกันโดยแอทริบิวต์สุดท้ายของตารางจะใช้กำหนดกลุ่มที่ต้องการแบ่ง

2.4 ความรู้ทั่วไปเกี่ยวกับระบบฐานข้อมูล

ฐานข้อมูลเป็นการจัดเก็บข้อมูลอย่างเป็นระบบ ทำให้ผู้ใช้สามารถใช้ข้อมูลที่เกี่ยวข้องในระบบงานต่าง ๆ ร่วมกันได้ โดยที่จะไม่เกิดความซ้ำซ้อนของข้อมูล และยังสามารถหลีกเลี่ยงความขัดแย้งของข้อมูลด้วย อีกทั้งข้อมูลในระบบก็จะต้องเชื่อถือได้ และเป็นมาตรฐานเดียวกัน โดยจะมีการกำหนดระบบความปลอดภัยของข้อมูลขึ้นนับได้ว่าปัจจุบันเป็นยุคของสารสนเทศ เป็นที่ยอมรับกันว่า สารสนเทศเป็นข้อมูลที่ผ่านการกลั่นกรองอย่างเหมาะสม สามารถนำมาใช้ประโยชน์อย่างมากมาย ไม่ว่าจะเป็นการนำมาใช้งานด้านธุรกิจ การบริหาร และกิจการอื่น ๆ องค์กรที่มีข้อมูลปริมาณมาก ๆ จะพบความยุ่งยากลำบากในการจัดเก็บข้อมูล ตลอดจนการนำข้อมูลที่ต้องการออกมาใช้ให้ทันต่อเหตุการณ์ ดังนั้นคอมพิวเตอร์จึงถูกนำมาใช้เป็นเครื่องมือช่วยในการจัดเก็บข้อมูล การประมวลผลข้อมูล ซึ่งทำให้ระบบการจัดเก็บข้อมูลเป็นไปได้อย่างสะดวก ทั้งนี้โปรแกรมแต่ละโปรแกรมจะต้องสร้างวิธีควบคุมและจัดการกับข้อมูลขึ้นเอง ฐานข้อมูลจึงเข้ามามีบทบาทสำคัญอย่างมาก โดยเฉพาะระบบงานต่าง ๆ ที่ใช้คอมพิวเตอร์ การออกแบบและพัฒนาระบบฐานข้อมูลจึงต้องคำนึงถึงการควบคุมและการจัดการความถูกต้อง ตลอดจนประสิทธิภาพในการเรียกใช้ข้อมูลด้วย

2.4.1 ความรู้พื้นฐานเกี่ยวกับระบบฐานข้อมูล

ระบบฐานข้อมูล (Database System) หมายถึง โครงสร้างสารสนเทศที่ประกอบด้วยรายละเอียดของข้อมูลที่เกี่ยวข้องกันที่จะนำมาใช้ในระบบต่าง ๆ ร่วมกันระบบฐานข้อมูล จึงนับว่าเป็นการจัดเก็บข้อมูลอย่างเป็นระบบ ซึ่งผู้ใช้สามารถจัดการกับข้อมูลได้ในลักษณะต่าง ๆ ทั้งการเพิ่ม การแก้ไข การลบ ตลอดจนการเรียกดูข้อมูล ซึ่งส่วนใหญ่จะเป็นการประยุกต์นำเอาระบบคอมพิวเตอร์เข้ามาช่วยในการจัดการฐานข้อมูลนิยามและคำศัพท์พื้นฐานเกี่ยวกับระบบฐานข้อมูล

บิต (Bit) หมายถึง หน่วยของข้อมูลที่มีขนาดเล็กที่สุดไบต์ (Byte) หมายถึง หน่วยของข้อมูลที่เกิดจากการนำบิตมารวมกันเป็นตัวอักษร (Character) เขตข้อมูล (Field) หมายถึง หน่วยของข้อมูลที่เกิดจากการนำบิตมารวมกันเป็นตัวอักษรตั้งแต่หนึ่งตัวขึ้นไปมารวมกันแล้วได้ความหมายของสิ่งใดสิ่งหนึ่ง เช่น ชื่อ ที่อยู่ เป็นต้นระเบียน (Record) หมายถึง หน่วยของข้อมูลที่เกิดจากการนำเขตข้อมูลหลาย ๆ เขตข้อมูลมารวมกัน เพื่อเกิดเป็นข้อมูลเรื่องใดเรื่องหนึ่ง เช่น ข้อมูลของนักศึกษา 1 ระเบียน (1 คน) จะประกอบด้วย รหัสประจำตัวนักศึกษา 1 เขตข้อมูลชื่อนักศึกษา 1 เขตข้อมูลที่อยู่ 1 เขตข้อมูลแฟ้มข้อมูล (File) หมายถึง หน่วยของข้อมูลที่เกิดจากการนำข้อมูลหลาย ๆ ระเบียนที่เป็นเรื่องเดียวกันมารวมกัน เช่น แฟ้มข้อมูลนักศึกษา แฟ้มข้อมูลลูกค้า แฟ้มข้อมูลพนักงาน

2.4.2 ส่วนในระบบฐานข้อมูล มีคำศัพท์ต่าง ๆ ที่เกี่ยวข้องดังนี้

เอนทิตี (Entity) หมายถึง ชื่อของสิ่งใดสิ่งหนึ่ง ได้แก่ คน สถานที่ สิ่งของ การกระทำ ซึ่งต้องการจัดเก็บข้อมูลไว้ เช่น เอนทิตีลูกค้า เอนทิตีพนักงาน เอนทิตีชนิดอ่อนแอ (Weak Entity) เป็นเอนทิตีที่ไม่มีความหมาย หากขาดเอนทิตีอื่นในฐานข้อมูลแอททริบิวต์ (Attribute) หมายถึง รายละเอียดข้อมูลที่แสดงลักษณะและคุณสมบัติของเอนทิตีหนึ่ง ๆ เช่น เอนทิตีนักศึกษา ประกอบด้วย แอททริบิวต์รหัสนักศึกษา

- แอททริบิวต์ชื่อนักศึกษา

- แอททริบิวต์ที่อยู่นักศึกษา

ความสัมพันธ์ (Relationships) หมายถึง ความสัมพันธ์ระหว่างเอนทิตี เช่น ความสัมพันธ์ระหว่างเอนทิตีนักศึกษาและเอนทิตีคณะวิชา เป็นลักษณะว่า นักศึกษาแต่ละคนเรียน อยู่คณะวิชาใดคณะวิชาหนึ่งในการแสดงความสัมพันธ์ระหว่างเอนทิตี เราจะใช้หัวลูกศรเพื่อแสดงความสัมพันธ์

ความสัมพันธ์ระหว่างเอนทิตี แบ่งออกเป็น 3 ประเภท คือ

1. ความสัมพันธ์แบบหนึ่งต่อหนึ่ง (One-to-one Relationships) เป็นการแสดงความสัมพันธ์ของข้อมูลในเอนทิตีหนึ่งที่มีความสัมพันธ์กับข้อมูลในอีกเอนทิตีหนึ่ง ในลักษณะหนึ่งต่อหนึ่ง (1 : 1)

2. ความสัมพันธ์แบบหนึ่งต่อกลุ่ม (One-to-many Relationships) เป็นการแสดงความสัมพันธ์ของข้อมูลในเอนทิตีหนึ่ง ที่มีความสัมพันธ์กับข้อมูลหลาย ๆ ข้อมูลในอีกเอนทิตีหนึ่ง ในลักษณะ (1:m)

3. ความสัมพันธ์แบบกลุ่มต่อกลุ่ม (Many-to-many Relationships) เป็นการแสดงความสัมพันธ์ของข้อมูลสองเอนทิตีในลักษณะกลุ่มต่อกลุ่ม (m:n)

เอนทิตีใบสั่งซื้อแต่ละใบจะสามารถสั่งซื้อสินค้าได้มากกว่าหนึ่งชนิด ความสัมพันธ์ของข้อมูลจากเอนทิตีใบสั่งซื้อไปยังเอนทิตีสินค้า จึงเป็นแบบหนึ่งต่อกลุ่ม (1:m) ในขณะที่สินค้าแต่ละชนิด จะถูกสั่งซื้อในใบสั่งซื้อหลายใบ ความสัมพันธ์ของข้อมูลจากเอนทิตีสินค้าไปยังเอนทิตีใบสั่งซื้อ จึงเป็นแบบหนึ่งต่อกลุ่ม (1:m) ดังนั้นความสัมพันธ์ของเอนทิตีทั้งสอง จึงเป็นแบบกลุ่มต่อกลุ่ม

(m:n) จากคำศัพท์ต่าง ๆ ที่เกี่ยวข้องกับระบบฐานข้อมูลที่ได้กล่าวมาแล้วข้างต้น จึงอาจให้นิยามของฐานข้อมูลในอีกลักษณะได้ว่า “ฐานข้อมูล” อาจหมายถึง โครงสร้างสารสนเทศ ที่ประกอบด้วยหลาย ๆ เอนทิตีที่มีความสัมพันธ์กัน

ความสำคัญของการประมวลผลแบบระบบฐานข้อมูล

จากการจัดเก็บข้อมูลรวมเป็นฐานข้อมูลจะก่อให้เกิดประโยชน์ดังนี้

1. สามารถลดความซ้ำซ้อนของข้อมูลได้ การเก็บข้อมูลชนิดเดียวกันไว้หลาย ๆ ที่ทำให้เกิดความซ้ำซ้อน (Redundancy) ดังนั้นการนำข้อมูลมารวมเก็บไว้ในฐานข้อมูล จะช่วยลดปัญหาการเกิดความซ้ำซ้อนของข้อมูลได้ โดยระบบจัดการฐานข้อมูล (Database Management System : DBMS) จะช่วยควบคุมความซ้ำซ้อนได้ เนื่องจากระบบจัดการฐานข้อมูลจะทราบได้ตลอดเวลาว่ามีข้อมูลซ้ำซ้อนกันอยู่ที่ใดบ้าง
2. หลีกเลี่ยงความขัดแย้งของข้อมูลได้หากมีการเก็บข้อมูลชนิดเดียวกันไว้หลาย ๆ ที่และมีการปรับปรุงข้อมูลเดียวกันนี้ แต่ปรับปรุงไม่ครบทุกที่ที่มีข้อมูลเก็บอยู่ก็จะทำให้เกิดปัญหาข้อมูลชนิดเดียวกัน อาจมีค่าไม่เหมือนกันในแต่ละที่ที่เก็บข้อมูลอยู่ จึงก่อให้เกิดความขัดแย้งของข้อมูลขึ้น (Inconsistency)
3. สามารถใช้ข้อมูลร่วมกันได้ฐานข้อมูลจะเป็นการจัดเก็บข้อมูลรวมไว้ด้วยกัน ดังนั้นหากผู้ต้องการใช้ข้อมูลในฐานข้อมูลที่มาจากแฟ้มข้อมูลต่างๆ ก็จะทำให้ทำได้โดยง่าย
4. สามารถรักษาความถูกต้องเชื่อถือได้ของข้อมูลบางครั้งพบว่าการจัดเก็บข้อมูลในฐานข้อมูลอาจมีข้อผิดพลาดเกิดขึ้น เช่น จากการที่ผู้ป้อนข้อมูลป้อนข้อมูลผิดพลาดคือป้อนจากตัวเลขหนึ่งไปเป็นอีกตัวเลขหนึ่ง โดยเฉพาะกรณีมีผู้ใช้หลายคนต้องใช้ข้อมูลจากฐานข้อมูลร่วมกัน หากผู้ใช้คนใดคนหนึ่งแก้ไขข้อมูลผิดพลาดก็ทำให้ผู้อื่นได้รับผลกระทบตามไปด้วย ในระบบจัดการฐานข้อมูล (DBMS) จะสามารถใส่กฎเกณฑ์เพื่อควบคุมความผิดพลาดที่เกิดขึ้น
5. สามารถกำหนดความเป็นมาตรฐานเดียวกันของข้อมูลได้การเก็บข้อมูลรวมกันไว้ในฐานข้อมูลจะทำให้สามารถกำหนดมาตรฐานของข้อมูลได้รวมทั้งมาตรฐานต่าง ๆ ในการจัดเก็บข้อมูลให้เป็นไปในลักษณะเดียวกันได้ เช่นการกำหนดรูปแบบการเขียนวันที่ ในลักษณะ วัน/เดือน/ปี หรือ ปี/เดือน/วัน ทั้งนี้จะมีผู้ที่คอยบริหารฐานข้อมูลที่เราเรียกว่า ผู้บริหารฐานข้อมูล (Database Administrator : DBA) เป็นผู้กำหนดมาตรฐานต่างๆ

6. สามารถกำหนดระบบความปลอดภัยของข้อมูลได้ระบบความปลอดภัยในที่นี้ เป็นการป้องกันไม่ให้ผู้ใช้ที่ไม่มีสิทธิมาใช้ หรือมาเห็นข้อมูลบางอย่างในระบบ ผู้บริหารฐานข้อมูลจะสามารถกำหนดระดับการเรียกใช้ข้อมูลของผู้ใช้แต่ละคนได้ตามความเหมาะสม

7. เกิดความเป็นอิสระของข้อมูล ในระบบฐานข้อมูลจะมีตัวจัดการฐานข้อมูลที่ทำหน้าที่เป็นตัวเชื่อมโยงกับฐานข้อมูล โปรแกรมต่าง ๆ อาจไม่จำเป็นต้องมีโครงสร้างข้อมูลทุกครั้ง ดังนั้นการแก้ไขข้อมูลบางครั้ง จึงอาจกระทำเฉพาะกับโปรแกรมที่เรียกใช้ข้อมูลที่เปลี่ยนแปลงเท่านั้น ส่วนโปรแกรมที่ไม่ได้เรียกใช้ข้อมูลดังกล่าว ก็จะเป็นอิสระจากการเปลี่ยนแปลง

2.4.2 รูปแบบของระบบฐานข้อมูล รูปแบบของระบบฐานข้อมูล มีอยู่ด้วยกัน 3 ประเภทคือ

1. ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) เป็นการเก็บข้อมูลในรูปแบบที่เป็นตาราง (Table) หรือเรียกว่า รีเลชัน (Relation) มีลักษณะเป็น 2 มิติ คือเป็นแถว (row) และเป็นคอลัมน์ (column) การเชื่อมโยงข้อมูลระหว่างตาราง จะเชื่อมโยงโดยใช้แอททริบิวต์ (attribute) หรือคอลัมน์ที่เหมือนกันทั้งสองตารางเป็นตัวเชื่อมโยงข้อมูล ฐานข้อมูลเชิงสัมพันธ์นี้จะเป็นรูปแบบของฐานข้อมูลที่นิยมใช้ในปัจจุบัน ดังตัวอย่าง

ตารางที่ 2.2 แสดงตารางพนักงาน

รหัสพนักงาน	ชื่อพนักงาน	ที่อยู่	เงินเดือน	รหัสแผนก
12501535	นายสมพงศ์	กรุงเทพ	12000	VO
12534568	นายมนตรี	นครปฐม	12500	VN
12503452	นายเอก	กรุงเทพ	13500	VO
12356892	นายบรรทัด	นนทบุรี	11500	VD

2. ฐานข้อมูลแบบเครือข่าย (Network Database) ฐานข้อมูลแบบเครือข่ายจะเป็นการรวมระเบียบต่าง ๆ และความสัมพันธ์ระหว่างระเบียบแต่ละระเบียบต่างกับฐานข้อมูลเชิงสัมพันธ์ คือ ในฐานข้อมูลเชิงสัมพันธ์จะแฝงความสัมพันธ์เอาไว้ โดยระเบียบที่มีความสัมพันธ์กันจะต้องมีค่าของ

ข้อมูลในแอททริบิวต์ใดแอททริบิวต์หนึ่งเหมือนกัน แต่ฐานข้อมูลแบบเครือข่าย จะแสดงความสัมพันธ์อย่างชัดเจน ตัวอย่างเช่น

3. ฐานข้อมูลแบบลำดับชั้น (Hierarchical Database) ฐานข้อมูลแบบลำดับชั้น เป็นโครงสร้างที่จัดเก็บข้อมูลในลักษณะความสัมพันธ์แบบพ่อ-ลูก (Parent-Child Relationship Type : PCR Type) หรือเป็นโครงสร้างรูปแบบต้นไม้ (Tree) ข้อมูลที่จัดเก็บในที่นี้ คือ ระเบียบ (Record) ซึ่งประกอบด้วยค่าของเขตข้อมูล (Field) ของเอนทิตีหนึ่ง ๆ ฐานข้อมูลแบบลำดับชั้นนี้คล้ายคลึงกับฐานข้อมูลแบบเครือข่าย แต่ต่างกันที่ฐานข้อมูลแบบลำดับชั้น มีกฎเพิ่มขึ้นมาอีกหนึ่งประการ คือในแต่ละกรอบจะมีลูกศรวิ่งเข้าหาได้ไม่เกิน 1 หัวลูกศร

2.4.5 โปรแกรมฐานข้อมูลที่นิยมใช้

โปรแกรมฐานข้อมูล เป็นโปรแกรมหรือซอฟต์แวร์ที่ช่วยจัดการข้อมูลหรือรายการต่าง ๆ ที่อยู่ในฐานข้อมูล ไม่ว่าจะเป็นการจัดเก็บ การเรียกใช้ การปรับปรุงข้อมูล โปรแกรมฐานข้อมูล จะช่วยให้ผู้ใช้สามารถค้นหาข้อมูลได้อย่างรวดเร็ว ซึ่งโปรแกรมฐานข้อมูลที่นิยมใช้มีอยู่ด้วยกันหลายตัว เช่น Access, FoxPro, Clipper, dBase, Fox Base, Oracle, SQL เป็นต้น โดยแต่ละโปรแกรมจะมีความสามารถต่างกัน บางโปรแกรมใช้ง่ายแต่จะจำกัดขอบเขตการใช้งาน บ้างโปรแกรมใช้งานยากกว่า แต่จะมีความสามารถในการทำงานมากกว่า

- **โปรแกรม Access** นับเป็นโปรแกรมที่นิยมใช้กันมากในขณะนี้ โดยเฉพาะในระบบฐานข้อมูลขนาดใหญ่ สามารถสร้างแบบฟอร์มที่ต้องการจะเรียกดูข้อมูลในฐานข้อมูล หลังจากบันทึกข้อมูลในฐานข้อมูลเรียบร้อยแล้ว จะสามารถค้นหาหรือเรียกดูข้อมูลจากเขตข้อมูลใดก็ได้ นอกจากนี้ Access ยังมีระบบรักษาความปลอดภัยของข้อมูล โดยการกำหนดรหัสผ่านเพื่อป้องกันความปลอดภัยของข้อมูลในระบบได้ด้วย

- **โปรแกรม FoxPro** เป็นโปรแกรมฐานข้อมูลที่มีผู้ใช้งานมากที่สุด เนื่องจากใช้ง่าย ทั้งวิธีการเรียกจากเมนูของ FoxPro และประยุกต์โปรแกรมอื่นใช้งาน โปรแกรมที่เขียนด้วย FoxPro จะสามารถใช้กลับ dBase คำสั่งและฟังก์ชันต่าง ๆ ใน dBase จะสามารถใช้งานบน FoxPro ได้นอกจากนี้ใน FoxPro ยังมีเครื่องมือช่วยในการเขียนโปรแกรม เช่น การสร้างรายงาน

- **โปรแกรม dBase** เป็นโปรแกรมฐานข้อมูลชนิดหนึ่ง การใช้งานจะคล้ายกับโปรแกรม FoxPro ข้อมูลรายงานที่อยู่ในไฟล์บน dBase จะสามารถส่งไปประมวลผลในโปรแกรม

Word Processor ได้ และแม้แต่ Excel ก็สามารถอ่านไฟล์ .DBF ที่สร้างขึ้นโดยโปรแกรม dBase ได้ด้วย

- **โปรแกรม SQL** เป็นโปรแกรมฐานข้อมูลที่มีโครงสร้างของภาษาที่เข้าใจง่าย ไม่ซับซ้อน มีประสิทธิภาพการทำงานสูง สามารถทำงานที่ซับซ้อนได้โดยใช้คำสั่งเพียงไม่กี่คำสั่ง โปรแกรม SQL จึงเหมาะที่จะใช้กับระบบฐานข้อมูลเชิงสัมพันธ์ และเป็นภาษาหนึ่งที่มีผู้นิยมใช้กันมาก โดยทั่วไปโปรแกรมฐานข้อมูลของบริษัทต่าง ๆ ที่มีใช้อยู่ในปัจจุบัน เช่น Oracle, DB2 ก็มักจะมีคำสั่ง SQL ที่ต่างจากมาตรฐานไปบ้างเพื่อให้เป็นจุดเด่นของแต่ละโปรแกรมไป

2.5 เอสคิวแอล (SQL)

เอสคิวแอล (SQL) คือ ภาษาสอบถามข้อมูล หรือภาษาจัดการข้อมูลอย่างมีโครงสร้าง มีการพัฒนาภาษาคอมพิวเตอร์ และโปรแกรมฐานข้อมูลทีรื่องรับมากมาย เพราะจัดการข้อมูลได้ง่าย เช่น MySQL, MsSQL, PostgreSQL หรือ MS Access เป็นต้น สำหรับโปรแกรมฐานข้อมูลที่ได้รับความนิยมคือ MySQL เป็น Open Source ที่ใช้งานได้ทั้งใน Linux และ Windows SQL เป็นภาษาที่ใช้ในการเขียนโปรแกรม เพื่อจัดการกับฐานข้อมูลโดยเฉพาะ เราสามารถแบ่งการทำงานได้เป็น 4 ประเภท ดังนี้

1. Select query ใช้สำหรับดึงข้อมูลที่ต้องการ
2. Update query ใช้สำหรับแก้ไขข้อมูล
3. Insert query ใช้สำหรับการเพิ่มข้อมูล
4. Delete query ใช้สำหรับลบข้อมูลออกไป

Select query

ใช้ในการดึงข้อมูลในฐานข้อมูล จะมีการค้นหารายการจากตารางในฐานข้อมูล ตั้งแต่หนึ่งตารางขึ้นไป ตามเงื่อนไขที่สั่ง ผลลัพธ์ที่ได้จะเป็นเซตของข้อมูลที่สามารถสร้าง เป็นตารางใหม่ หรือหรือใช้แสดงออกมาทางจอภาพเท่านั้น โดยมีรูปแบบดังนี้ Select รายละเอียดที่เลือก From ตารางแหล่งที่มา Where กำหนดเงื่อนไขฐานข้อมูล que เลือก Group by ชื่อคอลัมน์

ตัวอย่างการใช้งาน

1. Select fname,lname From stdinfo

หมายถึง ให้แสดงเฉพาะคอลัมน์ fname คือ ชื่อ และคอลัมน์ lname คือ นามสกุล จากตาราง stdinfo

2. Select fname,lname From stdinfo Where programe="สังคมศึกษา"

หมายถึง ให้แสดงชื่อ และนามสกุลจากตาราง stdinfo ซึ่งมีโปรแกรมวิชาเป็นสังคมศึกษา

3. Select fname From stdinfo Where fname Like 'ส%'

หมายถึง ให้เลือกรายชื่อ นักศึกษาที่มีอักษรนำหน้าเป็น "ส" ขึ้นมาแสดงทั้งหมด

4. Select id,fname,lname From stdinfo Where id="001" AND id="005"

หมายถึง ให้แสดง รหัสประจำตัวนักศึกษา, ชื่อ และ นามสกุล ที่มีรหัสเป็น 001 และ 005

ข้อสังเกต

1. ประโยคย่อย WHERE เราสามารถระบุเงื่อนไขได้โดยใช้โอเปอเรเตอร์ ทั่วไป เช่น NOT < > =
กรณีทีคอลัมน์เป็นตัวเลข เราก็สามารถระบุเงื่อนไขที่เป็นการคำนวณได้เช่น +,-,*,/

2. คำว่า Like ใช้กับค่าในคอลัมน์ประเภทตัวอักษรว่าตรงกับประโยคที่ต้องการหรือไม่ เราสามารถใช้เครื่องหมาย wildcard เช่น *,?%,% ในประโยคได้ ตามตัวอย่างข้างต้น

3. ในการคำนวณนั้นมีฟังก์ชัน COUNT,SUM,AVG,MIN,MAX ซึ่งสามารถนำมาใช้ได้ เช่น

Select Count(id) From stdinfo

หมายถึง ให้แสดงจำนวนรายการทั้งหมดในตาราง

4. ในกรณีที่ตารางสองตารางมีความสัมพันธ์กัน เราก็สามารถดูข้อมูลทั้งสองตารางพร้อมกันได้ เช่น ตารางที่ 1 ข้อมูลนักศึกษาชื่อ stdinfo

กำหนดให้มีคอลัมน์รหัสประจำตัว (id) ,ชื่อ (fname), นามสกุล (lname)

ตารางที่ 2 ข้อมูลเกี่ยวกับวิชาที่เรียน ชื่อ substd

กำหนดให้มีคอลัมน์ชื่อวิชา (subject), รหัสประจำตัวอ้างอิง (rid) ,อาจารย์ผู้สอน (teacher)

เราต้องการดูข้อมูลรหัสประจำตัว ชื่อ นามสกุล และชื่อวิชาที่เรียน เราจะใช้คำสั่งดังนี้

Select stdinfo.id, stdinfo.fname,stdinfo.lname,substd.subject From stdinfo,substd Where
stdinfo.id=substd.rid

Update query

ใช้สำหรับการแก้ไขข้อมูลในตาราง โดยแก้ไขในคอลัมน์ที่มีค่าตรงตามเงื่อนไข มีรูปแบบดังนี้
Update ชื่อตาราง Set [ชื่อคอลัมน์=ค่าที่จะใส่เข้าไปในคอลัมน์นั้น ๆ] Where เงื่อนไขเช่น จาก
ตารางแสดงรายชื่อนักศึกษากรณีที่นักศึกษาชื่อ สมบัติ มัคน้อย ย้ายโปรแกรมวิชา จาก สังคมศึกษา
ไปเป็นภาษาไทย เราใช้คำสั่งดังนี้

```
Select stdinfo Set programe='ภาษาไทย' Where Fname='สมบัติ' and Lname='มัคน้อย'
```

Insert query

ใช้ในการเพิ่มเติมข้อมูลใหม่ ๆ เข้าไปในฐานข้อมูล มีรูปแบบดังนี้
Insert Into ชื่อตาราง [=ชื่อคอลัมน์1,2...] Values [ค่าที่จะใส่ลงในคอลัมน์ 1,2...]
เช่น ต้องการเพิ่มรายชื่อนักศึกษา ที่มีรหัสประจำตัวเป็น 007 ชื่อ กมลวรรณ ศิริกุล โปรแกรมวิชา
วิทยาศาสตร์ เราสามารถใช้คำสั่งดังนี้
Insert into stdinfo (id,fname,lname,programe) Values ('007','กมลวรรณ','ศิริกุล','วิทยาศาสตร์')

Delete query

ใช้ลบข้อมูลออกจากตาราง มีรูปแบบดังนี้
Delete From ชื่อตาราง Where เงื่อนไข
เช่น ต้องการลบรหัสประจำตัวนักศึกษา 005 ออกจากฐานข้อมูล เราใช้คำสั่งดังนี้
Delete From stdinfo Where id='005'

MySQL (มายเอสคิวแอล)

MySQL (มายเอสคิวแอล) เป็นระบบจัดการฐานข้อมูลโดยใช้ภาษา SQL. แม้ว่า MySQL เป็นซอฟต์แวร์โอเพนซอร์ส แต่แตกต่างจากซอฟต์แวร์โอเพนซอร์สทั่วไป โดยมีการพัฒนาภายใต้บริษัท MySQL AB ในประเทศสวีเดน โดยจัดการ MySQL ทั้งในแบบที่ให้ใช้ฟรี และแบบที่ใช้ในเชิงธุรกิจ MySQL สร้างขึ้นโดยชาวสวีเดน 2 คน และชาวฟินแลนด์ ชื่อ David Axmark, Allan Larsson และ Michael "Monty" Widenius. ปัจจุบันบริษัทซันไมโครซิสเต็มส์ (Sun Microsystems, Inc.) เข้าซื้อกิจการของ MySQL AB เรียบร้อยแล้ว ฉะนั้นผลิตภัณฑ์ภายใต้ MySQL AB ทั้งหมดจะตกเป็นของซัน

ชื่อ “MySQL” อ่านออกเสียงว่า “มายเอสคิวเอล” หรือ “มายเอสคิวแอล” (ในการอ่านอักษร L ในภาษาไทย) ซึ่งทางซอฟต์แวร์ไม่ได้อ่าน มายซีเควล หรือ มายซีควล เหมือนกับซอฟต์แวร์จัดการฐานข้อมูลตัวอื่น

การใช้งาน

MySQL เป็นที่นิยมใช้กันมากสำหรับฐานข้อมูลสำหรับเว็บไซต์ เช่น มีเดียวิกิ และ phpBB และนิยมใช้งานร่วมกับภาษาโปรแกรม PHP ซึ่งมักจะได้อัปเดตเป็นประจำ จะเห็นได้จากคู่มือคอมพิวเตอร์ต่างๆ ที่จะสอนการใช้งาน MySQL และ PHP ควบคู่กันไป นอกจากนี้ หลายภาษาโปรแกรมที่สามารถทำงานร่วมกับฐานข้อมูล MySQL ซึ่งรวมถึง ภาษาซี ซีพลัสพลัส ปาสคาล ซีชาร์ป ภาษาจาวา ภาษาเพิร์ล พีเอชพี ไพทอน รูบี และภาษาอื่น ใช้งานผ่าน API สำหรับโปรแกรมที่ติดต่อผ่าน ODBC หรือ ส่วนเชื่อมต่อกับภาษาอื่น (database connector) เช่น เอเอสพี สามารถเรียกใช้ MySQL ผ่านทาง MyODBC, ADO, ADO.NET เป็นต้น

2.6 งานวิจัยที่เกี่ยวข้อง

นัจรูนิ เจริญเจริญดิบวร (2549:49) ใช้เทคนิคเหมืองข้อมูลเพื่อช่วยในการแนะแนวการศึกษาต่อระดับอุดมศึกษาโดยการนำเอาข้อมูลนักเรียนปีการศึกษา 2546- 2548 ได้แก่ ปีการศึกษา, รหัสนักเรียน, ภาคเรียน, วิชา, ผลการเรียน มาแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปศึกษาและพัฒนาตัวแบบโดยใช้เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Discovery) เพื่อเป็นแนวทางในการศึกษาต่อของนักเรียนในระดับอุดมศึกษาได้ศึกษาตัวแปรที่จะนำมาใช้สร้างเป็นเงื่อนไขและผลลัพธ์ ซึ่งในการวิจัยได้ใช้ผลการเรียนจำนวน 7 วิชาในระดับมัธยมศึกษา สร้างเป็นเงื่อนไข โดยแบ่งระดับผลการเรียนเป็น 3 ระดับคือ High หมายถึง ระดับเกรด 4, 3.5, 3 จะแทนด้วย “H” Medium หมายถึง ระดับเกรด 2.5, 2 จะแทนด้วย M และ Low หมายถึง ระดับเกรด 1.5, 1, 0 จะแทนด้วย L ส่วนผลลัพธ์คือคณะ 7 คณะ และมหาวิทยาลัยจำนวน 5 มหาวิทยาลัย ซึ่งการสร้างตัวแบบนี้จะต้องหากฎความสัมพันธ์ระหว่างผลการเรียน, คณะ และมหาวิทยาลัย โดยที่กฎความสัมพันธ์ที่ได้จะต้องคำนวณค่าความเชื่อมั่นและค่าสนับสนุนการพัฒนาตัวแบบโดยนำข้อมูลผลการสอบเข้าที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วมาสร้างตัวแบบตามที่ได้ศึกษามา โดยแบ่งข้อมูลจำนวน 158 คน ออกเป็น 2 ชุด คือข้อมูลเรียนรู้ และข้อมูลทดสอบ ในการแบ่งข้อมูลได้ใช้วิธีการเลือกตัวอย่างแบบมีระบบวงกลม และนำข้อมูลแต่ละชุดมาสร้างตัวแบบเป็นแบบข้อมูลเรียนรู้ (Training data) และแบบข้อมูลทดสอบแบบ (Testing data) โดยจะนำแบบข้อมูลเรียนรู้และแบบข้อมูลทดสอบมาเปรียบเทียบเพื่อหาความถูกต้องของตัวแบบ และ

คำนวณหาร้อยละความถูกต้องของตัวแบบ ผู้วิจัยได้สร้างตัวแบบข้อมูลการเรียนรู้ และสร้างตัวแบบข้อมูลตรวจสอบโดยข้อมูลที่แบ่งได้ตามสัดส่วนข้อมูล 60:40, 65:35, 70:30, 75:25 ผลที่ได้พบว่าตัวแบบมีความถูกต้อง คือ 82.32%, 81.87%, 91.78%, 92.37% ตามลำดับ

บวร น้อยแสง (2549: 69) ได้ศึกษาเรื่องคลังข้อมูลและเทคนิคเหมืองข้อมูลสำหรับการวิเคราะห์ โดยใช้กรณีศึกษาของบริษัทแฟรนไชส์ที่ได้แบ่งการดำเนินการออกเป็นสองส่วนคือส่วนของการทำคลังข้อมูลจากระบบการขาย และการทำเหมืองข้อมูลโดยใช้ข้อมูลที่ได้จากคลังข้อมูลในการทำคลังข้อมูลจากระบบการขาย เพื่อนำมาช่วยในการวิเคราะห์การขายพัฒนาในลักษณะของ Star Schema โดยมี Fact Table ที่มีตัววัดคือ จำนวนรวมของสินค้าที่ขาย ราคาสินค้าเฉลี่ย ต้นทุนเฉลี่ยของสินค้า มูลค่ารวมในการขายสินค้า และมีมิติ (Dimension) อยู่ 4 มิติคือ มิติของเวลา มิติของกลุ่มสินค้า มิติของลูกค้า มิติของพนักงานขาย โดยนำเสนอสารสนเทศของการวิเคราะห์การขายอยู่ในรูปของระบบ OLAP (On-Line Analytic Processing) โดยใช้โปรแกรม Microsoft SQL Server 2005 ซึ่งเป็นโปรแกรมจัดการฐานข้อมูลในการดำเนินงาน การใช้งานระบบ OLAP ช่วยให้การวิเคราะห์การขายมีความรวดเร็วในการเรียกดูข้อมูลหรือสารสนเทศและยังสามารถเรียกดูสารสนเทศในเชิงมิติต่างๆ ที่มีความซับซ้อนได้ เพื่อช่วยสนับสนุนในการตัดสินใจในการบริหารงานด้านการขาย ส่วนการทำเหมืองข้อมูลได้ใช้ข้อมูลที่ได้จากคลังข้อมูลซึ่งมีการกรองข้อมูลให้มีความถูกต้องของข้อมูลมากขึ้น เนื่องจากเทคนิคในการทำเหมืองข้อมูลนั้นมีมากมายไม่ว่าจะเป็นการใช้เทคนิคทางคณิตศาสตร์ ทางสถิติ หรือทางคอมพิวเตอร์เข้ามาช่วยจัดการข้อมูลที่มีขนาดใหญ่ เพื่อให้ได้สารสนเทศหรือสามารถแก้ปัญหาที่ต้องการได้ ใช้เทคนิคการรวมกลุ่มแบบ K-Means เพื่อรวมกลุ่มลูกค้าเพื่อช่วยให้สามารถดำเนินนโยบายทางการขาย การทำการส่งเสริมการขายการสร้างความสัมพันธ์กับลูกค้า หรือการบริหารงานด้านการขายอื่นๆ ได้ถูกกลุ่มเป้าหมายมากขึ้น ในการทำเหมืองข้อมูลผู้วิจัยได้ใช้ตัวแบบสำหรับการดำเนินแบบ CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งมีการดำเนินการ 6 ระยะ และวิเคราะห์การรวมกลุ่มด้วยเทคนิค K-Means ในการศึกษาได้ใช้โปรแกรม SPSS ซึ่งเป็นโปรแกรมวิเคราะห์ข้อมูลทางสถิติในการวิเคราะห์ ผลจากการดำเนินงานสามารถแบ่งกลุ่มลูกค้าได้เป็น 5 กลุ่มลูกค้า จากลูกค้าจำนวน 760 ราย โดยกลุ่มที่ 1 มีลูกค้าที่ถูกรวมเข้ากลุ่มนี้ 41 ราย เป็นกลุ่มที่มีค่าเฉลี่ยของวงเงินสินเชื่อในการขายสินค้ามากที่สุด และเป็นกลุ่มที่ทำรายได้ให้กับบริษัทมากที่สุด กลุ่มที่ 2 และ 3 เป็นกลุ่มลูกค้าที่เป็นบริษัทในเครือที่เปิดเป็นร้านค้าปลีกซึ่งมีจำนวนสมาชิกในกลุ่มกลุ่มละ 2 ราย แต่มูลค่ารวมที่ขายสินค้าทั้งหมดในกลุ่มที่ 3 จะมากกว่ากลุ่มที่ 2 กลุ่มที่ 4 มีจำนวนลูกค้าในกลุ่มจำนวน 48 รายเป็นกลุ่มลูกค้าที่จะซื้อสินค้าที่มีราคาเฉลี่ยของสินค้าสูงที่สุด กลุ่มที่ 5 เป็นกลุ่มที่มีขนาดใหญ่

ที่สุดคือมีลูกคำที่อยู่ในกลุ่มนี้จำนวน 667 ราย และเป็นกลุ่มที่มีจำนวนสินค้าเฉลี่ยที่ขายและมูลค่ารวมโดยเฉลี่ยของแต่ละบริษัทน้อยที่สุด

ศุวณีย์ กุลกรนิษฐรม (2549: 64) ได้ศึกษาและพัฒนาตัวแบบสำหรับการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED ในกลุ่มสาขาวิชาวิทยาศาสตร์ โดยใช้เทคนิคการทำเหมืองข้อมูล และพัฒนาระบบการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED โดยใช้ตัวแบบที่สร้างขึ้น ซึ่งได้เสนอขั้นตอนวิธีการศึกษาตัวแบบและพัฒนาตัวแบบ เพื่อนำตัวแบบที่สร้างขึ้น ไปพัฒนาระบบการจัดกลุ่มหลักสูตร ตามกลุ่มสาขาวิชา ISCED โดยมีขั้นตอนวิธีที่นำเสนอ 5 ส่วน ส่วนที่ 1 เป็นการเตรียมข้อมูลหลักสูตรสำหรับสร้างตัวแบบในงานวิจัย เลือกหลักสูตรสาขาวิชาที่มีชื่อตรงกับชื่อกลุ่มสาขาวิชา ISCED เพื่อความถูกต้องในการสร้างตัวแบบในการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED ส่วนที่ 2 เป็นการศึกษาตัวแบบเพื่อใช้ในการจัดกลุ่มหลักสูตร ตามกลุ่มสาขาวิชา ISCED ในกลุ่มสาขาวิชาวิทยาศาสตร์ โดยใช้เทคนิคการจำแนกประเภทข้อมูลแบบเบย์ ได้ศึกษาคุณสมบัติของตัวแปรที่จะนำมาใช้สร้างตัวแบบ และทฤษฎีการจำแนกประเภทข้อมูลแบบเบย์เพื่อนำมาใช้กับข้อมูลหลักสูตรที่ใช้ในการวิจัย ซึ่งในการวิจัยได้ใช้คำอธิบายรายวิชาให้หลักสูตรเป็นตัวแปร ส่วนผลลัพธ์คือกลุ่มสาขาวิชา ISCED ในการสร้างตัวแบบจะต้องหาค่าความน่าจะเป็นของแต่ละกลุ่มสาขาวิชา โดยนำค่าที่ได้จากคำอธิบายรายวิชาของหลักสูตรมาคำนวณ และยึดค่าความน่าจะเป็นของกลุ่มสาขาวิชาที่มีค่ามากที่สุดเป็นคำตอบของหลักสูตรส่วนที่ 3 เป็นการพัฒนาตัวแบบโดยนำข้อมูลหลักสูตรในกลุ่มสาขาวิชาวิทยาศาสตร์ที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วมาสร้างตัวแบบตามที่ได้ศึกษา โดยแบ่งข้อมูลออกเป็น 2 ชุดคือข้อมูลเรียนรู้ และข้อมูลทดสอบ โดยนำข้อมูลแต่ละชุดมาสร้างตัวแบบเป็นตัวแบบข้อมูลเรียนรู้และตัวแบบทดสอบข้อมูล ส่วนที่ 4 เป็นการทดสอบตัวแบบโดยจะนำตัวแบบข้อมูลเรียนรู้และตัวแบบทดสอบข้อมูลมาเปรียบเทียบเพื่อหาความถูกต้องของตัวแบบ และคำนวณหาร้อยละความถูกต้องของตัวแบบเพื่อพิจารณาว่าตัวแบบที่ศึกษาและพัฒนามานั้นสามารถนำไปใช้ในการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED ได้หรือไม่ ส่วนที่ 5 เป็นการพัฒนากระบวนการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED หลังจากทดสอบตัวแบบแล้วว่าตัวแบบที่ศึกษาและสร้างขึ้นนั้นสามารถนำไปใช้ได้ ดังนั้นในขั้นตอนนี้จะนำข้อมูลหลักสูตรในกลุ่มสาขาวิชาวิทยาศาสตร์มาสร้างตัวแบบ โดยข้อมูลหลักสูตรที่นำมาสร้างตัวแบบเป็นข้อมูลหลักสูตรในกลุ่มสาขาวิชาวิทยาศาสตร์จำนวน 7 กลุ่มสาขาวิชา ของมหาวิทยาลัยของรัฐจำนวน 8 แห่งด้วยกันจากผลการวิจัยแสดงให้เห็นว่าสามารถสร้างตัวแบบในการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED ได้จริงและทดสอบความถูกต้องของตัวแบบตามสัดส่วนของข้อมูลเรียนรู้และข้อมูลทดสอบเป็น 3 ช่วง คือ 60:40 , 70:30 และ 80:20 ช่วงละ 3 ครั้ง และคิดเป็นค่าเฉลี่ยร้อยละความถูกต้องของตัวแบบได้ 78.57, 85.71 และ 90.48 ตามลำดับ

นางสาวบุญมา เฟ่งชวน (2548: 102) ได้เสนอขั้นตอนวิธีการศึกษาตัวแบบและพัฒนาตัวแบบ เพื่อนำตัวแบบที่ได้นำไปประยุกต์ใช้ในการสนับสนุนการตัดสินใจด้านการผลิตบัณฑิตระดับปริญญาตรี โดยขั้นตอนวิธีที่นำเสนอแบ่งเป็น 5 ส่วน ส่วนแรกเป็นการเตรียมข้อมูลสถานะการมีงานทำของบัณฑิตโดยนำข้อมูลสถานะการมีงานทำของบัณฑิตมาแปลงข้อมูลให้อยู่ในรูปที่เหมาะสมสำหรับการนำไปศึกษาและพัฒนาตัวแบบโดยการแปลงข้อมูลนั้นต้องทำ 3 ขั้นตอนคือการ Mapping สาขาวิชา การ Mapping รหัสอาชีพตามมาตรฐานอาชีพแบบอัตโนมัติ และการ Mapping รหัสอาชีพตามมาตรฐานอาชีพแบบผู้ใช้เป็นผู้กำหนดเอง ซึ่งการ Mapping รหัสอาชีพนั้นจะใช้ฐานความรู้ที่สร้างขึ้นโดยพิจารณาจากตำแหน่งงานอาชีพต่างๆ ว่าสามารถจัดเข้ากับอาชีพตามมาตรฐานอาชีพใดได้ ส่วนที่ 2 เป็นการศึกษาตัวแบบเพื่อใช้ในการทำนายแนวโน้มการเลือกอาชีพแรกหลังสำเร็จการศึกษาของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการค้นหากฎความสัมพันธ์ ผู้วิจัยได้ศึกษาตัวแปรที่จะนำมาใช้สร้างเป็นเงื่อนไขและผลลัพธ์ ซึ่งในการวิจัยนี้ได้ใช้สาขาวิชาและเพศเป็นเงื่อนไขส่วนผลลัพธ์คืออาชีพ นั่นคือการสร้างตัวแบบนี้จะต้องหากฎความสัมพันธ์ระหว่าง สาขาวิชา เพศและอาชีพ โดยที่กฎความสัมพันธ์ที่ได้จะต้องคำนวณค่าความเชื่อมั่นและค่าสนับสนุน ส่วนที่ 3 เป็นการพัฒนาตัวแบบโดยนำข้อมูลสถานะการมีงานทำของบัณฑิตที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วมาสร้างตัวแบบตามที่ได้ศึกษามา โดยแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลเรียนรู้ และข้อมูลตรวจสอบ ส่วนที่ 4 เป็นการทดสอบตัวแบบโดยจะนำตัวแบบข้อมูลเรียนรู้และตัวแบบข้อมูลตรวจสอบมาเปรียบเทียบเพื่อหาความถูกต้องของตัวแบบ และคำนวณหาร้อยละความถูกต้องของตัวแบบ เพื่อพิจารณาว่าตัวแบบที่ศึกษาและพัฒนามานั้นสามารถนำไปใช้ในการสนับสนุนการตัดสินใจด้านการผลิตบัณฑิตระดับปริญญาตรีได้หรือไม่ ส่วนที่ 5 เป็นการพัฒนาระบบสนับสนุนการตัดสินใจ หลังจากทดสอบตัวแบบแล้วว่าตัวแบบที่ศึกษาและสร้างขึ้นนั้นสามารถนำไปใช้ได้ดังนั้นในขั้นตอนนี้จะนำข้อมูลสถานะการมีงานทำของบัณฑิตทั้งหมดมาสร้างตัวแบบ โดยข้อมูลที่นำมาสร้างตัวแบบจะเลือกเฉพาะรายการข้อมูลที่ค่าของรหัสอาชีพไม่เป็นค่าว่าง โดยตัวแบบที่ได้นั้นมีกฎความสัมพันธ์เกิดขึ้นจำนวนมากดังนั้นเพื่อเป็นการลดจำนวนกฎความสัมพันธ์จึงกำหนดค่าความเชื่อมั่นต่ำสุด และค่าสนับสนุนต่ำสุด นั่นหมายความว่าเฉพาะกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นและค่าสนับสนุนมากกว่าหรือเท่ากับค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด แต่เนื่องจากต้องนำตัวแบบที่ได้ไปใช้ในการหาแนวโน้มการประกอบอาชีพต่อไป ดังนั้นจึงต้องมีการจำแนกกฎความสัมพันธ์ให้กฎความสัมพันธ์ที่เกิดขึ้นต้องไม่ซ้ำกัน นั่นคือแต่ละเงื่อนไขสามารถมีผลลัพธ์ได้เพียงหนึ่งเดียว โดยพิจารณาหลักเกณฑ์จากค่าความเชื่อมั่นสูงสุด และค่าสนับสนุนสูงสุดของแต่ละเงื่อนไข ต่อจากนั้นจึงทำการแปลความหมายโดยนำข้อมูลรายบุคคลนักศึกษาที่ผ่านการแปลงข้อมูลแล้วกับตัวแบบที่ผ่านการจำแนกกฎความสัมพันธ์แล้วมา

ประมวลผล โดยการตรวจสอบสาขาวิชาและเพศของนักศึกษาและนำมาเปรียบเทียบกับตัวแบบก็จะได้ว่านักศึกษาจะประกอบอาชีพอะไรตามตัวแบบที่ได้ เมื่อแปลความหมายเสร็จเรียบร้อยแล้วสามารถเรียกดูรายงานจำนวนผู้ประกอบอาชีพต่างๆ โดยระบุเงื่อนไขต่างๆ เช่น ปี หรืออาชีพ ซึ่งรายงานก็สามารถเลือกแบบรายงานที่ออกรายงานได้จากผลการวิจัยแสดงให้เห็นว่าสามารถสร้างตัวแบบในรูปแบบกฎความสัมพันธ์ได้จริงและทดสอบความถูกต้องของตัวแบบตามสัดส่วนที่แบ่งข้อมูลเป็น 3 ช่วง คือ 60:40 , 70:30 และ 80:20 ช่วงละ 4 ครั้ง และคิดค่าเฉลี่ยร้อยละความถูกต้องได้ผล(Train : Validation) 60:40,70:30 , 80:20 ความถูกต้องของตัวแบบ (ค่าเฉลี่ย) 65.39%, 74.72%, 81.89% ตามลำดับ

บทที่ 3

ระเบียบวิธีวิจัย

3.1 ขั้นตอนการดำเนินการวิจัย

งานวิจัยนี้ได้นำเทคนิคการทำเหมืองข้อมูลมาใช้เพื่อช่วยในการตัดสินใจการอนุมัติสินเชื่อการเช่าซื้อรถยนต์ไร้ค่าของบริษัท เอกนครหลวงอิสซิง จำกัด โดยมีขั้นตอนการดำเนินการวิจัยดังนี้

1. ศึกษาและรวบรวมข้อมูล
2. เตรียมข้อมูล
3. การศึกษา ทดสอบและเลือกเทคนิคที่เหมาะสม
4. วิเคราะห์และสรุปผล
5. พัฒนาระบบสนับสนุนการตัดสินใจอนุมัติสินเชื่อ
6. สรุปผลการวิจัยและข้อเสนอแนะ
7. เรียบเรียงงานค้นคว้าอิสระ

3.2 เครื่องมือ (Tool) ที่ใช้ในการดำเนินการ

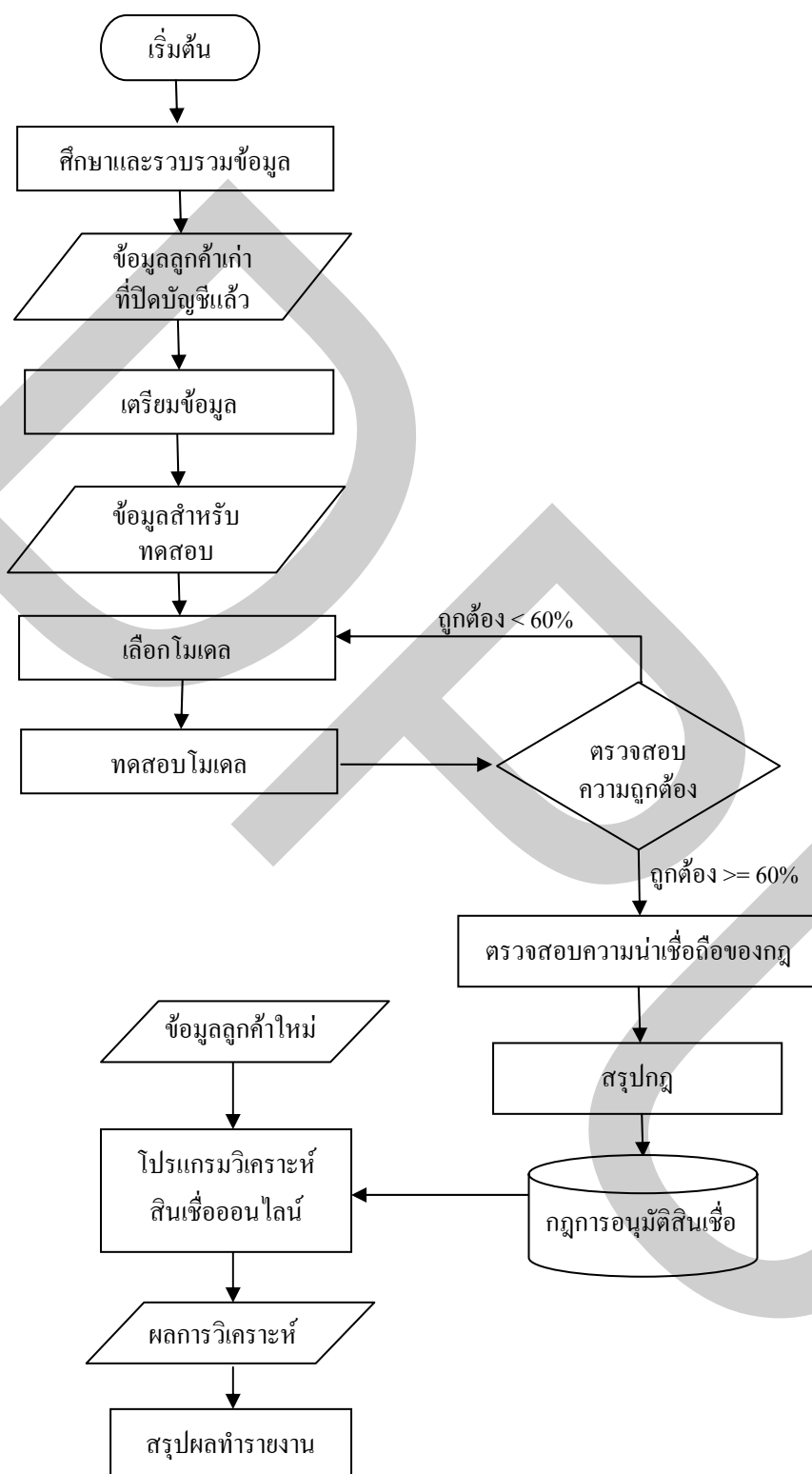
1. โปรแกรมสำเร็จรูป Weka เพื่อใช้ในการทำเหมืองข้อมูลเพื่อจัดกลุ่มลูกค้า
2. โปรแกรมตรวจสอบโครงสร้างของข้อมูลและจัดการข้อมูล จะใช้ DBF Manager
3. ใช้ระบบปฏิบัติการ: Windows XP Professional version 2002 Pack 2
4. โปรแกรม Microsoft Excel ใช้ในการทำความสะอาดข้อมูล ลดขนาดของข้อมูลและเตรียมข้อมูลสำหรับนำเข้าทดสอบในโปรแกรม Weka
5. โปรแกรม Microsoft Word ใช้ในการจัดทำเอกสารรายงานค้นคว้าอิสระ
6. โปรแกรม Macromedia Dreamweaver MX 2004, PHP, Apache และฐานข้อมูล MySQL สำหรับพัฒนาระบบสนับสนุนการตัดสินใจการอนุมัติสินเชื่อ
7. โปรแกรมสำเร็จรูป SQL Server 2008 ใช้ในการค้นหาข้อมูลที่ต้องการ

3.3 ระยะเวลาในการดำเนินการวิจัย

ระยะเวลาในการดำเนินการวิจัย และแผนผังขั้นตอนในการดำเนินงานการวิจัย สามารถสรุปได้ดัง ตารางที่ 3.1

ตารางที่ 3.1 ระยะเวลาในการดำเนินการวิจัย

ลำดับที่	ระยะเวลาดำเนินงาน (เดือน)	1	2	3	4	5	6	7	8	9	10	11	12
1	ศึกษาและรวบรวมข้อมูล	←→											
2	เตรียมข้อมูล		←→										
3	ศึกษา ทดสอบและเลือกเทคนิคที่เหมาะสม				←→								
4	วิเคราะห์และสรุปผล							←→					
5	พัฒนาระบบ								←→				
6	สรุปผลการวิจัยและข้อเสนอแนะ									←→			
7	เรียบเรียงรายงานค้นคว้าอิสระ									←→			



ภาพที่ 3.1 แผนผังขั้นตอนการดำเนินการวิจัย

3.4 สรุป

ขั้นตอนในการดำเนินการวิจัย ผู้วิจัยได้มีการแบ่งขั้นตอนที่จะศึกษาออกเป็น 7 ขั้นตอน ได้แก่ ขั้นตอนการศึกษาและรวบรวมข้อมูล การเตรียมข้อมูล การศึกษา ทดสอบและเลือกเทคนิคที่เหมาะสม การวิเคราะห์และสรุปผล การพัฒนาระบบสนับสนุนการตัดสินใจอัตโนมัติขึ้นชื่อ การสรุปผลการวิจัยและข้อเสนอแนะ และการเรียบเรียงงานค้นคว้าอิสระ

บทที่ 4

ผลดำเนินการวิจัย

หลังจากที่ศึกษาวิธีการต่างๆเพื่อนำมาประยุกต์ในงานวิจัยแล้วนั้น สามารถอธิบายผลการดำเนินงานวิจัยของผู้วิจัยออกเป็น 5 ส่วน ดังนี้

1. การศึกษาและรวบรวมข้อมูล
2. การเตรียมข้อมูล
3. การศึกษา ทดสอบและเลือกเทคนิคที่เหมาะสม
4. การวิเคราะห์และสรุปผล
5. การพัฒนาระบบสนับสนุนการตัดสินใจการอนุมัติสินเชื่อ

ผลการดำเนินงานวิจัยในแต่ละส่วนสามารถอธิบายโดยละเอียดดังต่อไปนี้

4.1 การศึกษาและรวบรวมข้อมูล

การศึกษาโครงสร้างข้อมูล เนื่องจากการจัดเก็บข้อมูลการเช่าซื้อของบริษัท เอกนครหลวง ลิสซิ่ง จำกัด อยู่ในรูปแบบแฟ้มข้อมูล ในระบบปฏิบัติการ DOS 7 ดังแสดงในภาพที่ 4.1 โดยประกอบด้วยแฟ้มสัญญาการเช่าซื้อ แฟ้มข้อมูลลูกค้า, แฟ้มข้อมูลรายละเอียดของรุ่นรถ และแฟ้มข้อมูลการชำระค่างวด เป็นต้น ตารางที่ 4.1 แสดงโครงสร้างข้อมูลของแฟ้มข้อมูลลูกค้า

โดยข้อมูลที่ใช้ในการศึกษานี้คือข้อมูลของลูกค้าที่ปิดบัญชีไปแล้ว ตั้งแต่ปี พ.ศ. 2545-2550 ซึ่งมีจำนวน 17,964 สัญญาที่ปิดบัญชีไปแล้ว และศึกษาถึงข้อมูลที่ใช้ในการพิจารณาอนุมัติสินเชื่อการเช่าซื้อรถจักรยานยนต์ โดยการข้อมูลที่ใช้ในการพิจารณาจะมีหลากหลายขึ้นอยู่กับสถานการณ์ ความชำนาญและประสบการณ์ของเจ้าหน้าที่วิเคราะห์สินเชื่อ โดยข้อมูลทั้งหมดที่มีการจัดเก็บมีจำนวน 142 แอตทริบิวต์ แต่ข้อมูลที่สำคัญที่ใช้ในการพิจารณาจะประกอบไปด้วยข้อมูล อาชีพ อายุ กลุ่มสินค้า ยี่ห้อสินค้า ปีที่ผลิตสินค้า จำนวนงวด ราคาสินค้า วงเงินที่ขอสินเชื่อ ค่างวด และเงินดาวน์ ดังนี้ อาชีพ อายุ รหัสกลุ่ม ยี่ห้อสินค้า ปีที่ผลิตสินค้า จำนวนเงินที่ขอสินเชื่อ ค่างวดต่อเดือน จำนวนงวด เงินดาวน์ และสถานภาพลูกค้า

ส่วนงานข้อมูล (ฐาน) F2-ใบเสร็จ F3-ภาษี F5-สถานะภาพ F6-สัญญา F7-ตาราง ค/น

BILL

BILL	ACCOUNT	FST.DUE	LAST	DUE	PAID	STAT	PAYMENT	CUSTOMER
0515	0703040013	05/11/20	03/05/20	21.0	15.9	--00	1865.00	เจ้าท่าอ่าว อ
0515	0712040008	05/01/20	03/07/20	19.0	16.7	-OK-	2710.00	สุวิศฯ จันท
0515	0001200002	05/03/20	05/05/20	17.0	12.8	--00	2330.00	ดอกอ้อ ศรีน
0515	0001300004	05/03/20	10/07/20	17.0	14.0	-OK-	2320.00	พุ่ม ทรัพย์
0515	0002200007	05/04/20	03/06/20	16.0	11.8	--00	1970.00	วราภรณ์ น
0515	0003050003	05/04/20	16/07/20	16.0	12.9	-OK-	2105.00	อภัย จันท
0515	0004040006	05/05/20	23/05/20	15.0	11.1	--00	2565.00	คำวิทย์ ปา
0515	0005120004	05/05/20	25/05/20	14.0	9.5	--00	2400.00	อริยารณ น
0515	0005000008	05/05/20	03/05/20	14.0	9.9	--00	2345.00	นภาพร โห
0515	0005310007	05/07/20	01/07/20	13.0	8.8	-OK-	2000.00	สุต นก
0515	0003300005	05/11/20	03/05/20	9.0	6.0	--00	2065.00	โกศล ธารา

ชื่อลูกค้า	: โกศล	CD 0515	: ปาวิ ศิษา
นามสกุล	: ธารา	ค้าง	: 2065.00
ชื่อผู้ค้า	: กิตติ	จริง	: 9.00
นามสกุล	: ธาร	จำนวน	: 6.00
ทะเบียนรถ	: นร 329	วันที่	: 05/11/2008
ทะเบียนรถ	: 07/10/2009	วันที่	: 03/05/2009
ประวัติ	: / /	เลข	: NF100CE-0131899

ภาพที่ 4.1 ข้อมูลในระบบปฏิบัติการ DOS

ตารางที่ 4.1 โครงสร้างข้อมูลของแฟ้มข้อมูลลูกค้า (บางส่วน)

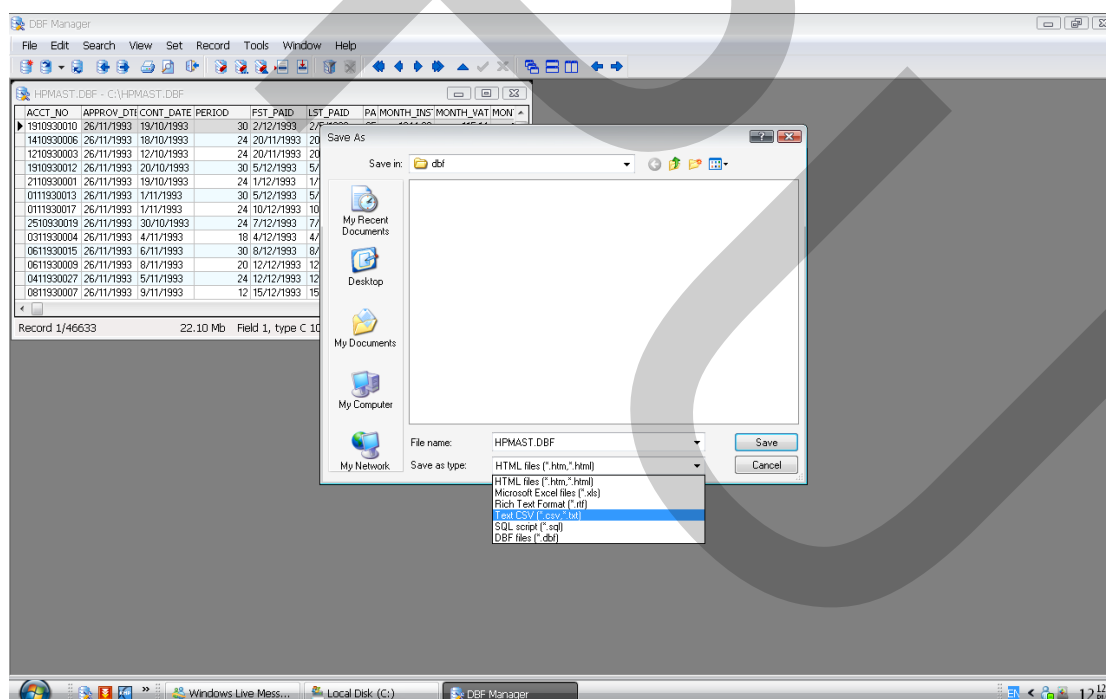
DBF Table structure				
Field	Field Name	Field Type	Field Length	Field Decimal
1	ACCT_NO	C [Character]	10	0
2	TYPE_CODE	C [Character]	1	0
3	CUST_RANK	C [Character]	4	0
4	CUST_NANK	C [Character]	30	0
5	CUST_SUR	C [Character]	30	0
6	PADDRESS1	C [Character]	40	0
7	PADDRESS2	C [Character]	40	0
8	PADDRESS3	C [Character]	40	0
9	PADDRESS4	C [Character]	40	0
10	PZIP	C [Character]	8	0
11	TELEPHONE1	C [Character]	10	0

4.2 การเตรียมข้อมูล (Data Pre-processing)

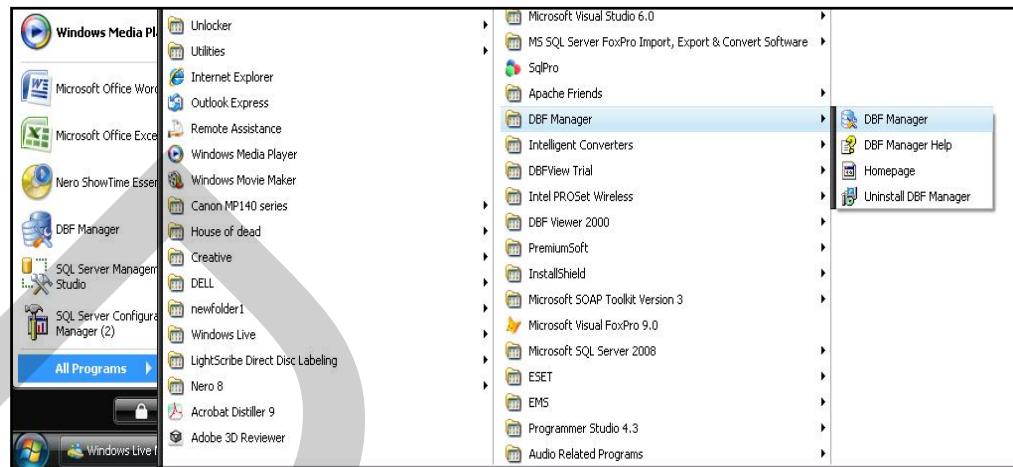
เนื่องจากข้อมูลที่บริษัทได้จัดเก็บไว้ยังไม่อยู่ในรูปแบบที่สามารถนำมาใช้ในการทดสอบและวิเคราะห์ด้วยเครื่องมือการทำเหมืองข้อมูล อีกทั้งยังมีปัญหาทั้งในด้านปริมาณและคุณภาพของข้อมูล จึงต้องมีการเตรียมข้อมูลก่อนนำไปใช้ในการทำเหมืองข้อมูลเพื่อให้ได้ผลการทดลองที่มีประสิทธิภาพสูงสุด โดยในขั้นตอนนี้จะประกอบด้วยขั้นตอนย่อย 3 ขั้นตอนคือ การโอนย้ายข้อมูล (Data Transfer) การลดขนาดของข้อมูล (Data Reduction) และการทำความสะอาดข้อมูล (Data Cleaning)

4.2.1 การโอนย้ายข้อมูล (Data Transfer)

ข้อมูลเดิมที่เก็บจะอยู่ในรูปแบบแฟ้มข้อมูล ในระบบปฏิบัติการ DOS7 จึงจำเป็นต้องมีการโอนย้ายข้อมูลมาอยู่ในรูปแบบที่สามารถนำไปใช้งานได้ ในการวิจัยครั้งนี้ได้โอนย้ายข้อมูลมาอยู่ในรูปแบบไฟล์ Excel ในระบบปฏิบัติการ Windows และบันทึกข้อมูลให้อยู่ในรูปแบบ .CSV เพื่อสะดวกในการนำเข้าไปทดลองในโปรแกรม Weka



ภาพที่ 4.2 การนำออกข้อมูล



ภาพที่ 4.3 การเปิดโปรแกรม DBF Manager เพื่อศึกษาโครงสร้างข้อมูล

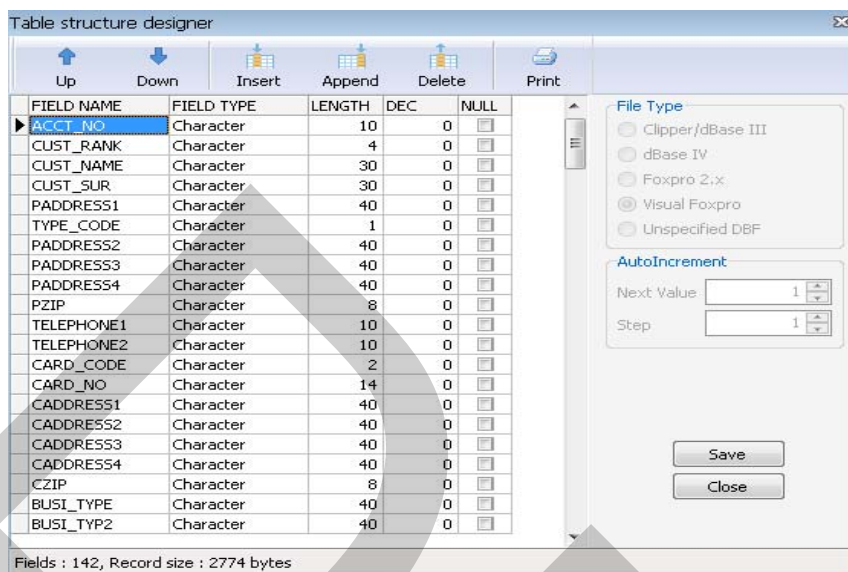
DBF Manager - (hpmasp.dbf - C:\Documents and Settings\Administrator\My Documents\hpmasp.dbf)

File Edit Search View Set Record Tools Window Help

ACCT_NO	CUST_NAME	CUST_SUR	PADDRESS1	TADDRESS2	PADDRESS3	PADDRESS4
211030001	นาย ชิตชัย	จตุรพร	68/185 ต.พิชัยรัตนบุรี	ต.เลาขวัญ ต.บางบอน	ต.บางขุนเทียน	กทม.
211030010	น.ส. สุณี	จันทร์	85/132 ซ.พหลโยธินซอย 1	ต.พหลโยธิน 4 ต.รังสิต	ต.รังสิต	สมุทรสาคร
211030012	นาย สุรินทร์	วิเศษ	89 ม.13 ซ.สุขุม	ต.สุขุมวิท 2	ต.สุขุมวิท 2	จ.ขอนแก่น
251030019	นาย สุทธิชัย	วิเศษ	231 หมู่ 2	ต.พหลโยธิน 4 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
011130013	นาย ชัยชนะ	วิเศษ	25 ม.8 ต.วิเศษนครนายก	ต.พหลโยธิน 4 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
011130017	นาย วิเศษ	วิเศษ	177/11 ม.5 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
031130014	นาย กิติ	พชรพร	8/123 ม.14 ต.พหลโยธิน 103	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
061130015	น.ส. อรุณพร	วิเศษ	886/25 ซ.ก. 132	ต.พหลโยธิน 4 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
061130019	น.ส. กมล	วิเศษ	15/1 หมู่ 10 ต.พหลโยธิน	ต.พหลโยธิน 4	จ.พหลโยธิน	จ.ขอนแก่น
041130027	นาย ประเสริฐ	คำพิพร	32 ม.11 ต.พหลโยธิน ต.บางบอน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
081130020	นาย อธิชัย	วิเศษ	116/221 หมู่ 4	ต.พหลโยธิน 4 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
081130027	นาย สมชาย	คำพิพร	30/68 หมู่ 19 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
081130018	นาย อธิชัย	วิเศษ	43 ต.พหลโยธิน	ต.พหลโยธิน (75) ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
081130030	นาย ประเสริฐ	คำพิพร	13/15 ต.พหลโยธิน ต.พหลโยธิน	ต.พหลโยธิน 27 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
121130006	น.ส. ประเสริฐ	วิเศษ	ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
171130007	นาย ชิตชัย	วิเศษ	12/1 ม.8	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
171130013	นาย คำพิพร	วิเศษ	60/145 ต.พหลโยธิน	ต.พหลโยธิน 2	จ.พหลโยธิน	จ.ขอนแก่น
171130021	นาย สุทธิชัย	วิเศษ	61/45 หมู่ 7 ต.พหลโยธิน 75	ต.พหลโยธิน 4 ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
201130006	นาย สมชาย	คำพิพร	176 ต.พหลโยธิน 30	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
231130008	นาย อธิชัย	วิเศษ	330/14-15 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
231130007	นาย ชิตชัย	วิเศษ	204/127 ม.4	ต.พหลโยธิน 103	จ.พหลโยธิน	จ.ขอนแก่น
251130013	นาย ประเสริฐ	คำพิพร	69/5 ต.พหลโยธิน 18	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
261130010	น.ส. อธิชัย	วิเศษ	151/128 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301130003	นาย อธิชัย	วิเศษ	113/105 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301200011	นาย วิเศษ	วิเศษ	1266/2 ม.2 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301200015	นาย ประเสริฐ	คำพิพร	32/1 ม.1	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301200022	นาย นภาพร	วิเศษ	310/21 ซ. 2 ม. 5 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
271130019	นาย สมชัย	วิเศษ	290 ต.พหลโยธิน 82	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301207009	นาย สุทธิชัย	วิเศษ	290/3 ม.5 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301204015	นาย วิเศษ	วิเศษ	240/304 ม.14 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301200004	นาย ประเสริฐ	คำพิพร	11/12/12 (ซอย) ต.พหลโยธิน 2	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
301208009	นาย ชิตชัย	วิเศษ	311 ต.พหลโยธิน 8	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
3012090025	น.ส. ราณี	วิเศษ	844/16 ต.พหลโยธิน 107	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
3012090032	นาย สุทธิชัย	วิเศษ	113 ม.6 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
3012110012	นาย ชิตชัย	วิเศษ	304/2 ซ. 3 ม. 5 ต.พหลโยธิน	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
3012160011	นาย กมล	วิเศษ	4/6 ม.19 ต.พหลโยธิน 1	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น
3012040034	นาย ประเสริฐ	คำพิพร	52 ต.พหลโยธิน 230/ก. 1	ต.พหลโยธิน	จ.พหลโยธิน	จ.ขอนแก่น

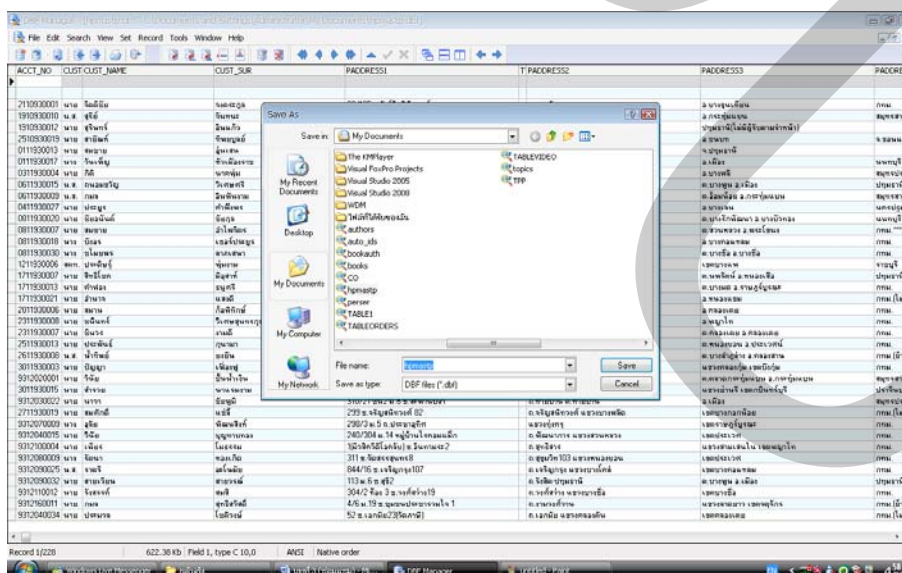
Record 1/228 622.38 kb Field 1, type C 10,0 ANSI Native order

ภาพที่ 4.4 หน้าจอแสดงข้อมูลลูกค้าในโปรแกรม DBF Manager



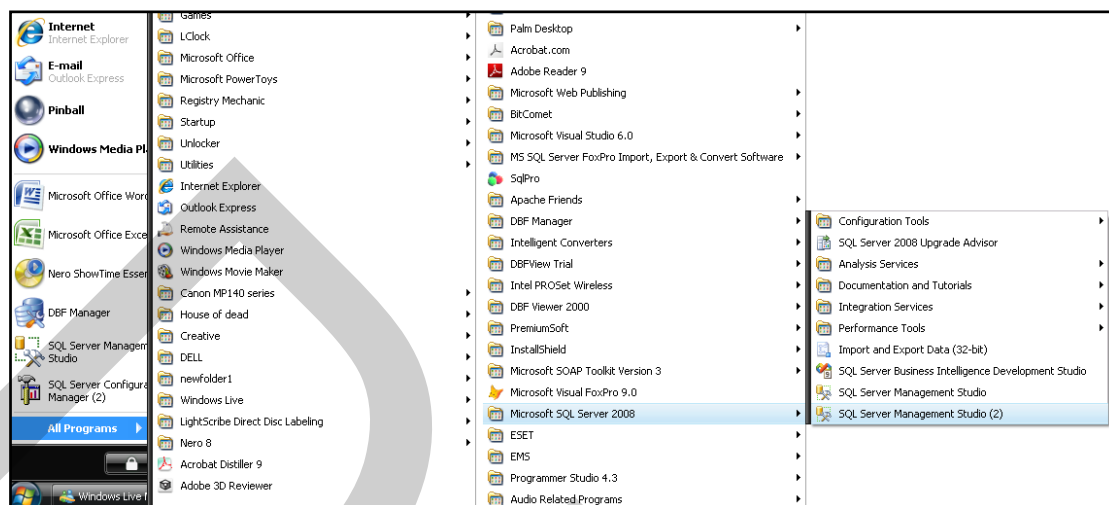
ภาพที่ 4.5 การตรวจสอบโครงสร้างของข้อมูล (Data Structure)

การตรวจสอบโครงสร้างของข้อมูลเพื่อให้ทราบว่าข้อมูลที่มีอยู่มีแอทริบิวต์ใดบ้างและชนิดของแอทริบิวต์ใดที่จะสามารถนำมาใช้ในการทำวิจัย

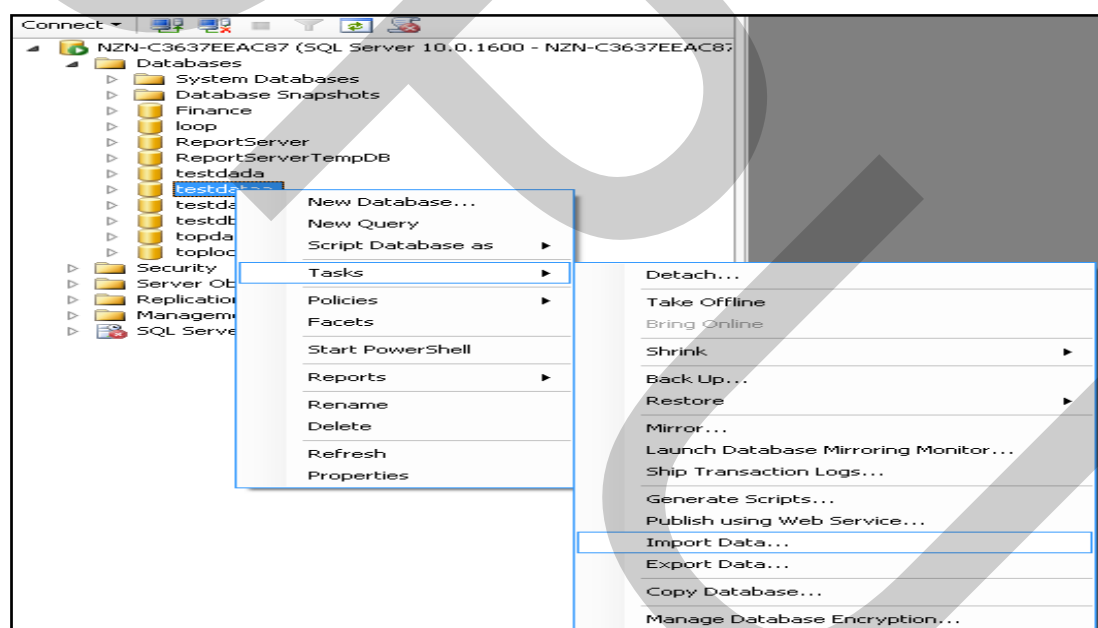


ภาพที่ 4.6 การบันทึกข้อมูล

เมื่อทราบโครงสร้างของข้อมูลจะทำการทำการบันทึกข้อมูลเพื่อที่จะนำไปใช้ในการวิจัยต่อไป

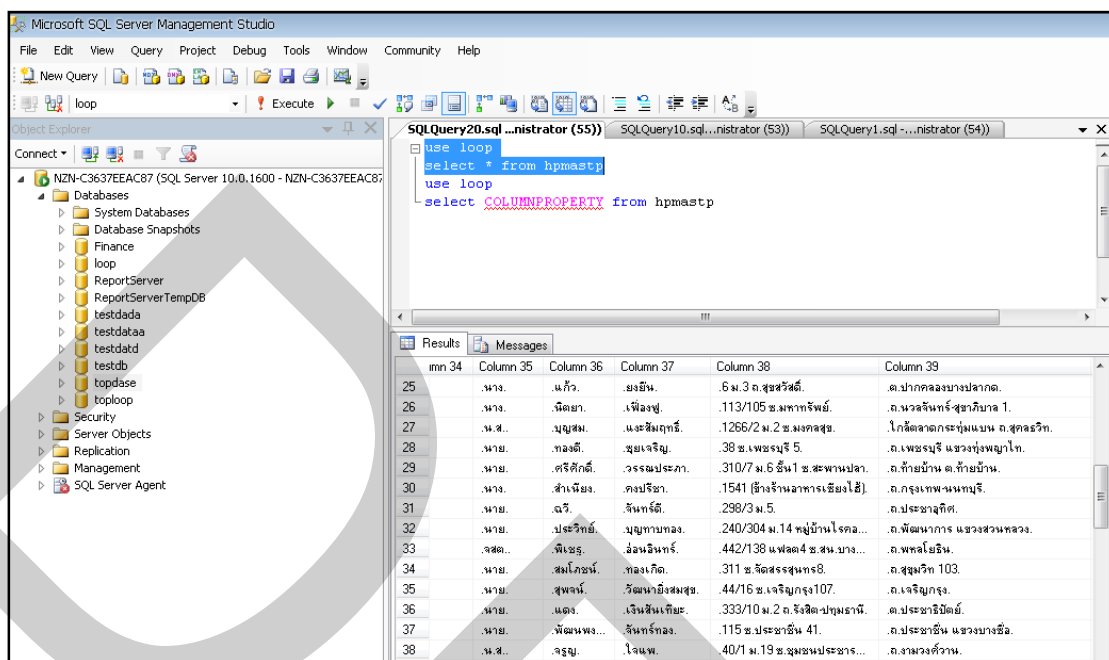


ภาพที่ 4.7 การเปิดโปรแกรม SQL Server 2008



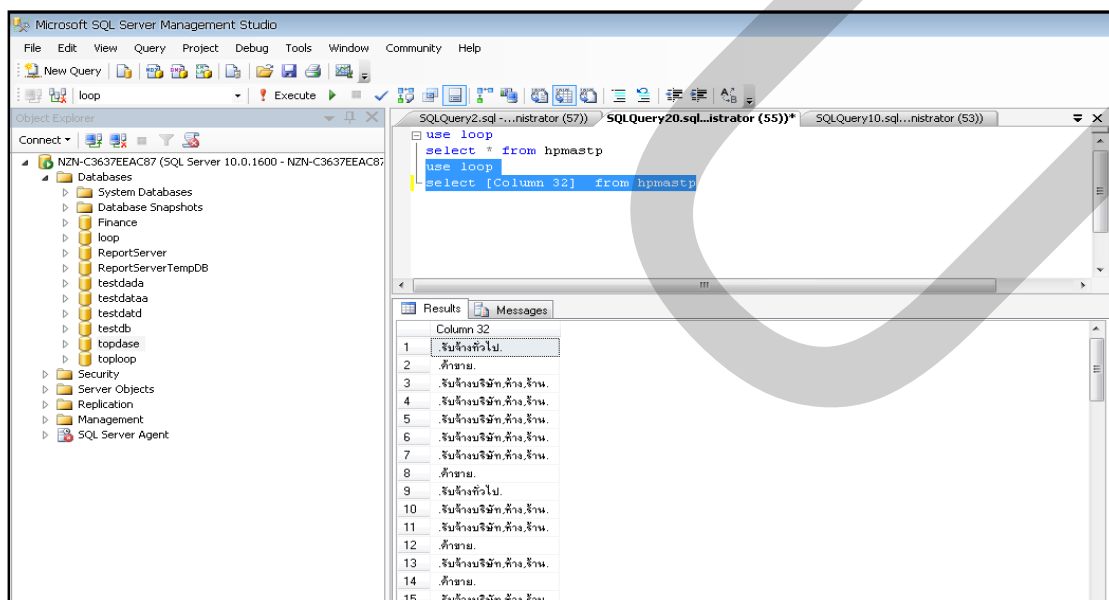
ภาพที่ 4.8 การนำเข้าไฟล์ DBF เข้าโปรแกรม SQL Server 2008

การนำเข้าไฟล์ DBF ให้มาอยู่ใน SQL Server 2008 เพื่อที่จะสร้างข้อมูลเดิมให้เป็นฐานข้อมูลใหม่ เนื่องจากข้อมูลเดิมที่มีอยู่มีการจัดเก็บเป็นไฟล์ไม่ได้อยู่ในลักษณะของฐานข้อมูล ทำให้ไม่สามารถที่จะสืบค้นข้อมูลที่ต้องการนำมาทำการวิจัยได้



ภาพที่ 4.9 การสืบค้นข้อมูลในฐานข้อมูลใหม่ที่สร้างขึ้น

แสดงการสืบค้นหาแอทริบิวต์ที่ต้องการก่อนที่จะนำออกข้อมูลเป็นไฟล์Excel เพื่อนำไปใช้งานต่อไป



ภาพที่ 4.10 แสดงการสืบค้นหาแอทริบิวต์ที่ต้องการ

	A	B	C	D	E	F	G	H	I	J	K	L
	ACCT_NO	OCCUP_NAME	CUST_AGE	GROUP_CODE	MODEL_CODE	PROD_BRAND	YEAR	PERIOD	FINAN_AMT	MONTH_TOT	BRAND_NAME	ACCT_STA
1												
2	1210930003	รับจ้างทั่วไป	22	1	3	2	93	24	22150	1605	YAMAHA	CP
3	1410930006	ค้าขาย	28	1	9	2	93	24	30280	2195	YAMAHA	NF
4	111930017	ค้าขาย	36	1	10	1	93	24	30280	2195	HONDA	NF
5	311930004	รับจ้างทั่วไป	35	1	1	1	93	18	23925	2090	HONDA	NH
6	411930027	ค้าขาย	42	1	2	4	93	24	19907	1445	SUZUKI	HI
7	811930007	ค้าขาย	25	1	8	3	93	12	24579	2880	KAWASAKI	NF
8	1211930006	ข้าราชการพลเรือน	29	1	12	2	93	24	38037	2755	YAMAHA	NF
9	2011930006	รับจ้างทั่วไป	21	1	11	1	93	12	31495	3685	HONDA	NF
10	2311930008	รับจ้างทั่วไป	36	1	11	1	93	24	31495	2285	HONDA	NF
11	2311930007	ค้าขาย	19	1	12	2	93	30	38037	2415	YAMAHA	NF
12	9312020001	รับจ้างทั่วไป	18	1	8	1	93	30	29439	1870	HONDA	CP
13	3011930015	ชั้นกรรับจ้าง	41	1	12	2	93	26	38037	2625	YAMAHA	NF
14	9312070009	ค้าขาย	40	1	8	3	93	20	24579	2000	KAWASAKI	CP
15	9312040015	ค้าขาย	26	1	13	2	93	24	43271	3135	YAMAHA	NF
16	9312090032	รับจ้างทั่วไป	24	1	9	1	93	24	31028	2250	HONDA	NF
17	9312160011	ค้าขาย	45	1	8	3	93	24	24579	1780	KAWASAKI	NH
18	9312110016	ค้าขาย	20	1	3	1	93	24	26542	1925	HONDA	NF
19	9312210007	ค้าขาย	46	1	3	2	93	24	22617	1640	YAMAHA	NF
20	9312220004	เกษมตรกร	40	1	5	3	93	24	42617	3090	KAWASAKI	NH
21	9312210027	ซิมมิลเลอร์ไรซ์รับจ้าง	33	2	5	3	92	24	32000	2455	KAWASAKI	NF
22	9312290008	ซิมมิลเลอร์ไรซ์รับจ้าง	33	1	15	3	93	30	45514	2890	KAWASAKI	NF
23	9312220021	รับจ้างทั่วไป	21	1	15	3	93	30	45514	2890	KAWASAKI	NF
24	9312280019	ซิมมิลเลอร์ไรซ์รับจ้าง	35	1	11	3	93	20	36355	2960	KAWASAKI	NF
25	9312300017	พนักงานรัฐวิสาหกิจ	35	1	2	2	93	24	30280	2195	YAMAHA	NH
26	9312270038	ค้าขาย	20	1	16	1	93	24	25514	1850	HONDA	NF
27	9401050014	ค้าขาย	34	1	9	2	93	24	30280	2195	YAMAHA	NF
28	9401060016	ค้าขาย	23	1	9	2	93	12	23832	2750	YAMAHA	CP
29	9401080008	ซิมมิลเลอร์ไรซ์รับจ้าง	21	1	9	4	93	24	31776	2305	SUZUKI	NF
30	9401060033	ข้าราชการทหาร	20	1	12	2	93	30	38505	2445	YAMAHA	NF
31	9401130011	ลูกจ้างหน่วยงานของ	24	1	2	4	93	23	20561	1530	SUZUKI	NF
32	9401140003	ค้าขาย	32	1	15	1	93	24	24299	1760	HONDA	NF
33	9401170004	พนักงานรัฐวิสาหกิจ	43	1	15	1	93	24	24299	1760	HONDA	CP
34	9401170004	พนักงานรัฐวิสาหกิจ	43	1	15	1	93	24	24299	1760	HONDA	CP

ภาพที่ 4.11 ข้อมูลที่นำออกมาอยู่ในไฟล์ Excel

เนื่องจากข้อมูลที่น่าออกมาซึ่งได้ทำการทำความสะอาดต้องนำมาผ่านกระบวนการเตรียมข้อมูลก่อนที่จะนำเข้าข้อมูล

4.2.2 การลดขนาดของข้อมูล (Data Reduction)

เนื่องจากข้อมูลมีจำนวนมากจึงจำเป็นต้องมีการลดขนาดของข้อมูลลง เพื่อให้สะดวกในการวิเคราะห์จึงได้มีการสืบค้นโดยเลือกเฉพาะข้อมูลบัญชีสัญญาการเช่าซื้อรถจักรยานยนต์ ไม่นับรวมบัญชีรถยนต์และบัญชีอื่นๆ เลือกเฉพาะสัญญาที่อยู่ระหว่างปี 2545-2550 ซึ่งข้อมูลที่ใช้ในการพิจารณาอนุมัติสินเชื่อเช่าซื้อรถจักรยานยนต์จะมีหลากหลายขึ้นอยู่กับสถานการณ์ ความชำนาญและประสบการณ์ของเจ้าหน้าที่วิเคราะห์สินเชื่อ แต่ข้อมูลหลักที่สำคัญจะประกอบไปด้วยการพิจารณา อาชีพ, อายุ, กลุ่มสินค้า, ยี่ห้อสินค้า, ปีที่ผลิตสินค้า, จำนวนงวด, ราคาสินค้า, วงเงินที่ขอสินเชื่อ, ค่างวด และเงินดาวน์ จึงได้มีการเลือกข้อมูลเฉพาะแถวข้อมูลที่มีข้อมูลสมบูรณ์ครบทุกแอททริบิวต์ที่ต้องการ ซึ่งได้ข้อมูลจำนวน 17,964 แถว และมีการลดความหลากหลายของค่าของข้อมูลในบางแอททริบิวต์ที่ลง เช่น สถานะภาพลูกค้าแต่เดิมจะมีความหลากหลายมาก เช่น ลูกค้าที่ปิดบัญชีปกติ, ลูกค้าเร่งรัดปิดปกติ,ลูกค้าที่ออกมาขายรอซื้อรถ, ลูกค้าที่ยึดได้รถตั้งขาย, ลูกค้าที่ยึดไม่ได้รอการฟ้องร้องโดยค่าของข้อมูลในแอททริบิวต์นี้จะถูกลดค่าลงเหลือเพียง GOOD กับ BAD ซึ่งหมายถึง ลูกค้าดี และ ลูกค้าไม่ดี ตามลำดับ

Field No	Field Name	Description	Field Type	Field Length
1	OCCUP_NAME	อาชีพ	C [Character]	40
2	CUST_AGE	อายุ	N [Numeric]	2
3	GROUP_NAME	กลุ่มสินค้า	C [Character]	40
4	BRAND_NAME	ยี่ห้อสินค้า	C [Character]	40
5	YEAR	ปีที่ผลิตสินค้า	C [Character]	4
6	PERIOD	จำนวนงวด	N [Numeric]	2
7	TOT_PRICE	ราคาสินค้า	N [Numeric]	12
8	FINAN_TOT	ยอดกู้	N [Numeric]	12
9	MONTH_TOT	ค่างวด	N [Numeric]	10
10	DOWN_TOT	เงินดาวน์	N [Numeric]	10
11	ACCT_STAT	สถานะภาพ	C [Character]	2

ภาพที่ 4.12 ข้อมูลที่ได้จากการลดขนาดของข้อมูล

ข้อมูลที่ได้มาต้องทำการลดขนาดของข้อมูลลงด้วยการลบรายการสินค้าสำหรับรถยนต์ออก ให้เหลือเพียงรายการสินค้าสำหรับรถจักรยานยนต์ และลบรายการที่ยังไม่ปิดบัญชีออก เพื่อให้ได้ข้อมูลที่ต้องการนำไปใช้จริง

ตารางที่ 4.2 โครงสร้างของข้อมูลที่ใช้ในการวิเคราะห์การพิจารณาอนุมัติสินเชื่อ

Field No	Field Name	Description	Field Type	Field Length
1	OCCUP_NAME	อาชีพ	C [Character]	40
2	CUST_AGE	อายุ	N [Numeric]	2
3	GROUP_NAME	กลุ่มสินค้า	C [Character]	40
4	BRAND_NAME	ยี่ห้อสินค้า	C [Character]	40
5	YEAR	ปีที่ผลิตสินค้า	C [Character]	4
6	PERIOD	จำนวนงวด	N [Numeric]	2
7	TOT_PRICE	ราคาสินค้า	N [Numeric]	12
8	FINAN_TOT	ยอดกู้	N [Numeric]	12
9	MONTH_TOT	ค่างวด	N [Numeric]	10
10	DOWN_TOT	เงินดาวน์	N [Numeric]	10
11	ACCT_STAT	สถานะภาพ	C [Character]	2

4.2.3 การทำความสะอาดข้อมูล (Data Cleaning)

ในขั้นตอนนี้เป็นตรวจสอบ และแก้ไขค่าของข้อมูลให้มีความหมายที่ถูกต้องตรงกัน เช่น ค่าของข้อมูลกลุ่มสินค้า มีการบันทึกเป็น สินค้าเก่า สินค้าใหม่ รถจักรยานยนต์เก่า รถจักรยานยนต์ใหม่ ดังนั้นจึงต้องมีการแก้ไขข้อมูล จาก รถจักรยานยนต์เก่า ให้เป็น สินค้าเก่า และ รถจักรยานยนต์ใหม่ ให้เป็น สินค้าใหม่ เป็นต้น ค่าของข้อมูลต่างๆมีรายละเอียดดังนี้

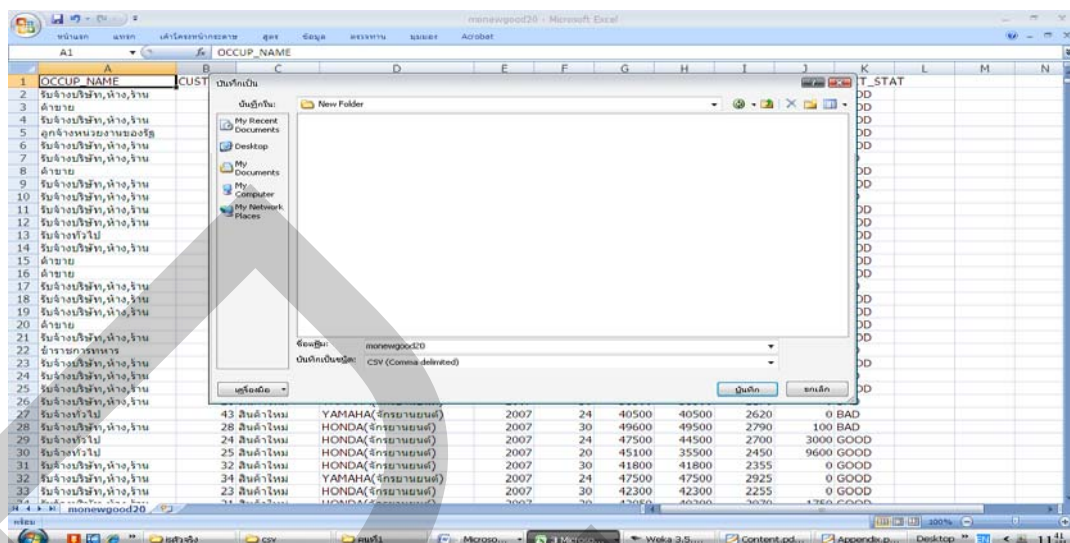
ตารางที่ 4.3 การแปลงค่าของข้อมูลสถานะภาพ

ลำดับ	สถานะภาพเดิม	สถานะภาพใหม่
1	CP = ลูกค้าปิดบัญชีปกติ	GOOD
2	NP = ปิดบัญชีปกติหยุด vat	
3	HI = เร่งรัดยึดได้	
4	FF = ยึดได้รอขาย	BAD
5	ED = เร่งรัดรอยึด	
6	NH = ตัดขาดทุนสินค้ายึด	
7	NF = ยึดไม่ได้ตัดหนี้สูญ	

เนื่องจากสถานะภาพของบัญชีที่บริษัทมีการจัดเก็บ แบ่งออกเป็นหลายประเภทจึงต้องมีการแปลงข้อมูลสถานะให้เหลือเพียงกลุ่มลูกค้าดี (GOOD) และลูกค้าไม่ดี (BAD)

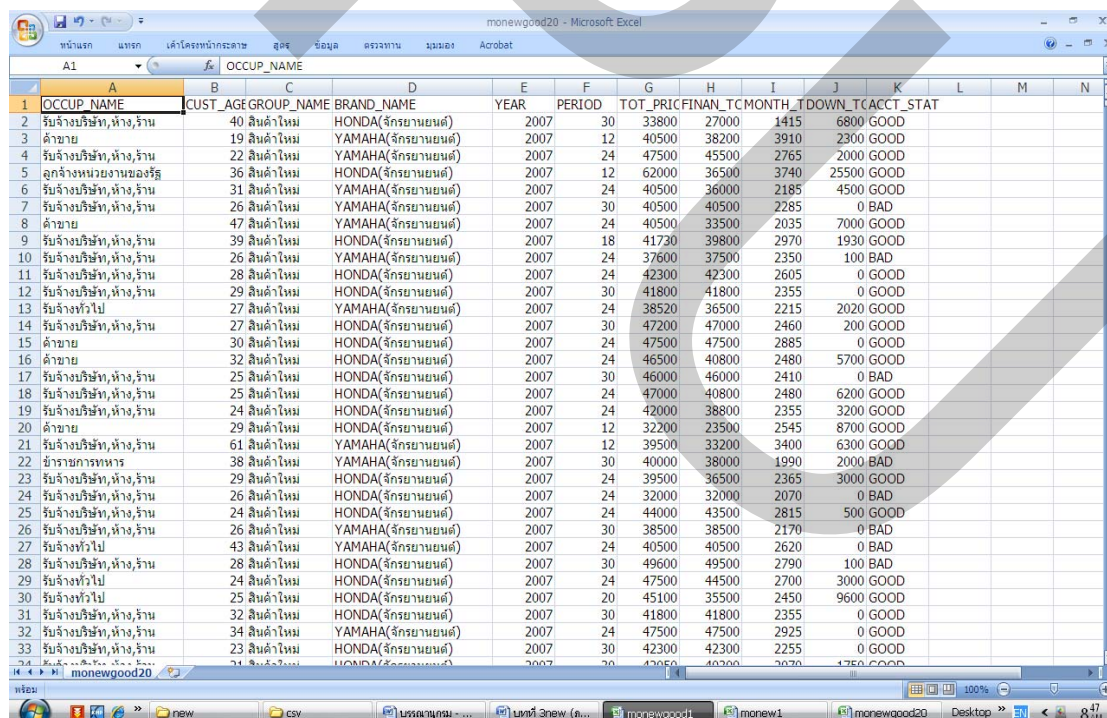
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	OCCUP_NAME	CUST_AGE	GROUP_NAME	BRAND_NAME	YEAR	PERIOD	TOT_PRICE	FINANCIAL	MONTH	TDOWN	CACCT	STAT		
1	ร้านจิวเวลรี่	40	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	33800	27000	1415	680	GOOD			
2	ร้านจิวเวลรี่	19	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	12	40500	38200	3910	2300	GOOD			
3	ร้านจิวเวลรี่	22	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	47500	45500	2765	2000	GOOD			
4	ร้านจิวเวลรี่	36	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	12	62000	36500	3740	2550	GOOD			
5	ร้านจิวเวลรี่	31	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	40500	36000	2185	4500	GOOD			
6	ร้านจิวเวลรี่	26	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	30	40500	40500	2285	0	BAD			
7	ร้านจิวเวลรี่	47	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	40500	33500	2035	7000	GOOD			
8	ร้านจิวเวลรี่	39	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	18	41730	39800	2970	1900	GOOD			
9	ร้านจิวเวลรี่	26	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	37600	37500	2350	100	BAD			
10	ร้านจิวเวลรี่	28	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	42300	42300	2605	0	GOOD			
11	ร้านจิวเวลรี่	29	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	41800	41800	2355	0	GOOD			
12	ร้านจิวเวลรี่	27	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	38520	36500	2215	2000	GOOD			
13	ร้านจิวเวลรี่	27	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	47200	47000	2460	200	GOOD			
14	ร้านจิวเวลรี่	30	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	47500	47500	2885	0	GOOD			
15	ร้านจิวเวลรี่	32	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	46500	40800	2480	570	GOOD			
16	ร้านจิวเวลรี่	25	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	46000	46000	2410	0	BAD			
17	ร้านจิวเวลรี่	25	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	47000	40800	2480	6200	GOOD			
18	ร้านจิวเวลรี่	24	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	42000	38800	2355	3200	GOOD			
19	ร้านจิวเวลรี่	29	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	12	32200	23500	2545	870	GOOD			
20	ร้านจิวเวลรี่	61	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	12	39500	33200	3400	6300	GOOD			
21	ร้านจิวเวลรี่	38	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	30	40000	38000	1990	2000	BAD			
22	ร้านจิวเวลรี่	29	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	39500	36500	2365	3000	GOOD			
23	ร้านจิวเวลรี่	26	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	32000	32000	2070	0	BAD			
24	ร้านจิวเวลรี่	24	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	44000	43500	2815	500	GOOD			
25	ร้านจิวเวลรี่	26	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	30	38500	38500	2170	0	BAD			
26	ร้านจิวเวลรี่	43	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	40500	40500	2620	0	BAD			
27	ร้านจิวเวลรี่	28	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	49600	49500	2790	100	BAD			
28	ร้านจิวเวลรี่	24	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	24	47500	44500	2700	3000	GOOD			
29	ร้านจิวเวลรี่	25	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	20	45100	35500	2450	960	GOOD			
30	ร้านจิวเวลรี่	32	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	41800	41800	2355	0	GOOD			
31	ร้านจิวเวลรี่	34	สินค้าใหม่	YAMAHA(จักรยานยนต์)	2007	24	47500	47500	2925	0	GOOD			
32	ร้านจิวเวลรี่	23	สินค้าใหม่	HONDA(จักรยานยนต์)	2007	30	42300	42300	2255	0	GOOD			

ภาพที่ 4.13 ผลลัพธ์การแปลงค่าของข้อมูลสถานะภาพ



ภาพที่ 4.14 การบันทึกข้อมูลในรูปแบบนามสกุล .csv

ในการนำเข้าข้อมูลของโปรแกรม Weka3.5.7 ต้องทำการแปลงไฟล์ข้อมูลให้เป็นนามสกุล csv ก่อน จึงจะสามารถนำเข้าข้อมูลในโปรแกรม Weka ได้



ภาพที่ 4.15 แสดงไฟล์ข้อมูลในรูปแบบ .csv

ตารางที่ 4.4 ผลการวิเคราะห์ค่าของข้อมูลประเภทอาชีพ

ลำดับ	ประเภทอาชีพ	จำนวน	ร้อยละ(%)
1	รับจ้างบริษัท,ห้าง,ร้าน	10,874	60.53
2	ค้าขาย	2,868	15.96
3	ลูกจ้างหน่วยงานของรัฐ	365	2.03
4	รับจ้างทั่วไป	1,238	6.89
5	ข้าราชการทหาร	189	1.05
6	จี้มอเตอร์ไซด์รับจ้าง	1025	5.71
7	ขับรถรับจ้าง	182	1.01
8	เกษตรกร	639	3.56
9	มีทรัพย์สินให้เช่า	77	0.43
10	ผู้รับเหมาก่อสร้าง	158	0.88
11	ข้าราชการตำรวจ	74	0.41
12	รับจ้างก่อสร้าง	64	0.36
13	พนักงานรัฐวิสาหกิจ	158	0.88
14	ข้าราชการพลเรือน	24	0.13
15	อื่นๆ	29	0.16
	รวม	17,954	100.00

ตารางที่ 4.5 ผลการวิเคราะห์ค่าของข้อมูลอายุ

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	1
ค่าสูงสุด	70
ค่ากลาง	31.346

ตารางที่ 4.6 ผลการวิเคราะห์ค่าของข้อมูลกลุ่มสินค้า

ลำดับ	กลุ่มสินค้า	จำนวน	ร้อยละ(%)
1	สินค้าใหม่	12,751	70.98
2	สินค้าเก่า	5,213	29.02
	รวม	17,964	100.00

ตารางที่ 4.7 ผลการวิเคราะห์ค่าของข้อมูลยี่ห้อสินค้า

ลำดับ	ยี่ห้อ	จำนวน	ร้อยละ(%)
1	Honda	12,362	68.81
2	Yamaha	4,962	27.62
3	JRD	255	1.42
4	Kawasaki	217	1.29
5	Suzuki	168	0.94
	รวม	17,964	100.00

ตารางที่ 4.8 ผลการวิเคราะห์ค่าของข้อมูลปีที่ผลิตสินค้า

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	1994
ค่าสูงสุด	2009
ค่ากลาง	2004

ตารางที่ 4.9 ผลการวิเคราะห์ค่าของข้อมูลจำนวนงวดที่ผ่อนชำระ

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	2
ค่าสูงสุด	36
ค่ากลาง	21

ตารางที่ 4.10 ผลการวิเคราะห์ค่าของข้อมูลราคาสินค้า

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	5,000
ค่าสูงสุด	85,000
ค่ากลาง	33,090

ตารางที่ 4.11 ผลการวิเคราะห์ค่าของข้อมูลยอดกฐ์

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	4,000
ค่าสูงสุด	65,000
ค่ากลาง	30,509

ตารางที่ 4.12 ผลการวิเคราะห์ค่าของข้อมูลค่างวด

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	870
ค่าสูงสุด	12,1300
ค่ากลาง	2,168

ตารางที่ 4.13 ผลการวิเคราะห์ค่าของข้อมูลเงินค่าน้ำ

สถิติ	ค่าของข้อมูล
ค่าต่ำสุด	0
ค่าสูงสุด	41,032
ค่ากลาง	2,580

ตารางที่ 4.14 ผลการวิเคราะห์ค่าของข้อมูลสถานภาพลูกค้า

สถานภาพ	จำนวน	ร้อยละ(%)
GOOD	11,801	65.96
BAD	6,163	34.31

4.3 การศึกษาทดสอบ และเลือกเทคนิคที่เหมาะสม

เทคนิคการจำแนกกลุ่มข้อมูลมีหลากหลายเทคนิค และแต่ละเทคนิคก็ให้ผลลัพธ์ที่มีรูปแบบแตกต่างกัน ในการศึกษาครั้งนี้ต้องการนำโมเดลที่ได้ไปพัฒนาระบบสนับสนุนการตัดสินใจการพิจารณาการอนุมัติสินเชื่อ ดังนั้นเทคนิคการจำแนกกลุ่มที่นำมาใช้ในการศึกษาครั้งนี้

จึงเป็นเทคนิคที่ให้ผลลัพธ์เป็นกฎทั้งหมด โดยรายละเอียดการทำงานของแต่ละเทคนิคได้อธิบายไว้ในบทที่ 2 เทคนิคที่ศึกษาได้แก่

1) โมเดลการจำแนกกลุ่มแบบ Decision Tree: C4.5 หรือเรียกว่าโมเดล Classifiers trees J48 ในโปรแกรม Weka

2) โมเดลการจำแนกกลุ่มแบบ Decision Rules: Part หรือเรียกว่าโมเดล Classifiers PART ในโปรแกรม Weka

3) โมเดลการจำแนกกลุ่มแบบ Decision Table หรือเรียกว่าโมเดล Classifiers Decision Table ในโปรแกรม Weka

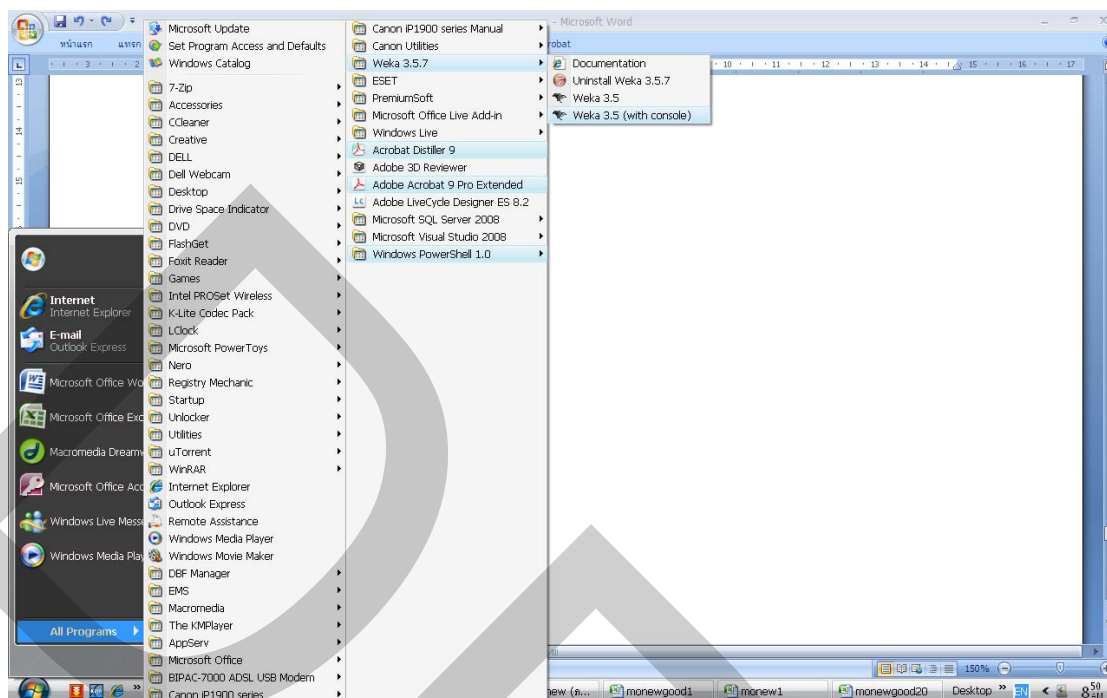
การศึกษาค้างนี้ทำการทดสอบโมเดลโดยใช้โปรแกรมประยุกต์ Weka เวอร์ชัน 3.5.7 ข้อมูลที่ใช้ในการทดสอบมีจำนวน 17,964 แถว ในการทดสอบได้ใช้การทดสอบแบบ 10-fold cross validation ซึ่งได้แก่ การแบ่งกลุ่มข้อมูลออกเป็น 10 กลุ่ม ในแต่ละรอบจะนำกลุ่มข้อมูลจำนวน 9 กลุ่ม เป็นกลุ่มศึกษา (Training Set) และกลุ่มข้อมูลที่เหลือเป็นกลุ่มทดสอบ (Testing Set) โดยจะทำการซ้ำเป็นจำนวน 10 รอบ เพื่อเปลี่ยนกลุ่มทดสอบให้ครบทุกกลุ่ม และนำผลลัพธ์ที่ได้มาหาค่าเฉลี่ย ซึ่งจะทำการทดลองหลายๆ โมเดลโดยมีการเปลี่ยนแปลงข้อมูลนำเข้า ดังนี้

— ทดลองปรับค่าของข้อมูลที่เป็นตัวเลขแบบต่อเนื่อง ให้เป็นกลุ่มข้อมูลที่ไม่ต่อเนื่องแบบต่างๆ เช่น ข้อมูลยอด้ว ค่างวด จำนวนงวด มีการปรับค่าเป็น 10 ช่วง ได้แก่การแบ่งช่วงแบบให้ระยะห่างที่เท่ากัน และแบ่งช่วงแบบความถี่เท่ากัน ผลที่ได้พบว่าจำนวนกฎที่ออกมามากเกินไป โดยแต่ละกฎที่ได้มีจำนวนตัวอย่างสนับสนุนน้อยทำให้ไม่น่าเชื่อถือ

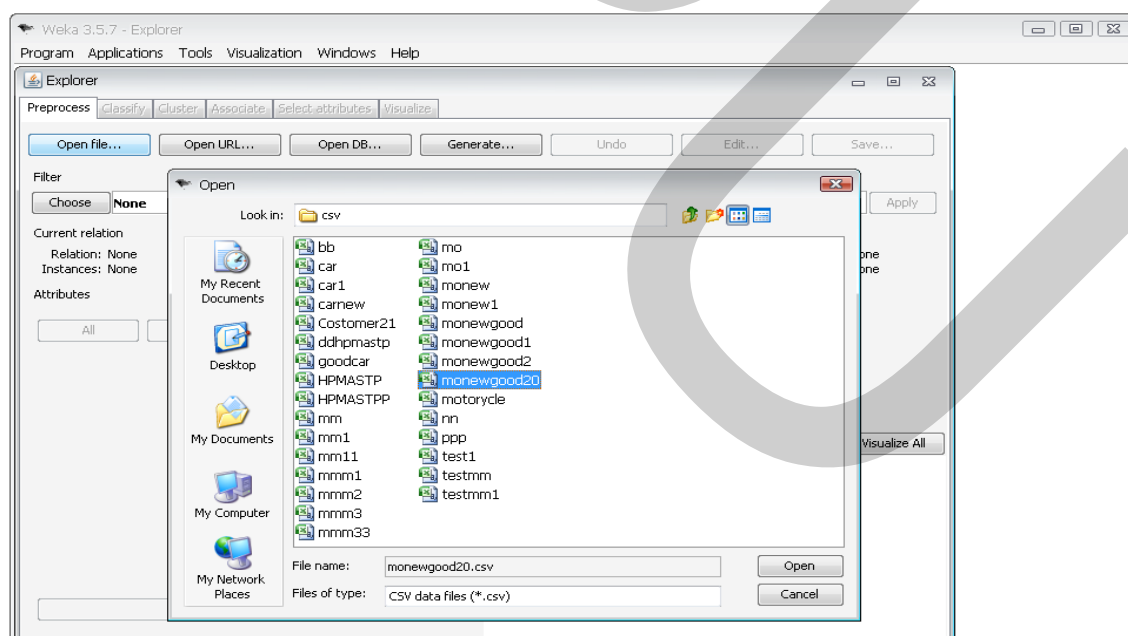
— ทดลองตัดข้อมูลบางแอทริบิวต์ออก เช่น ปีรด ผลที่ได้พบว่า จำนวนกฎที่ออกมาน้อยเกินไป และมีความถูกต้องต่ำ ทำให้ไม่น่าเชื่อถือ

— ทดลองตัดแถวบางแถวออก เช่น อายุที่น้อยๆ อาชีพบางอาชีพ ผลที่ได้พบว่าจำนวนกฎที่ออกมาน้อยเกินไป และมีความถูกต้องต่ำ ทำให้ไม่น่าเชื่อถือ

เกณฑ์ที่ใช้ในการเลือกโมเดลในการศึกษานี้ พิจารณาจากเปอร์เซ็นต์ความถูกต้องที่มากที่สุด และได้จำนวนกฎที่น้อยที่สุด

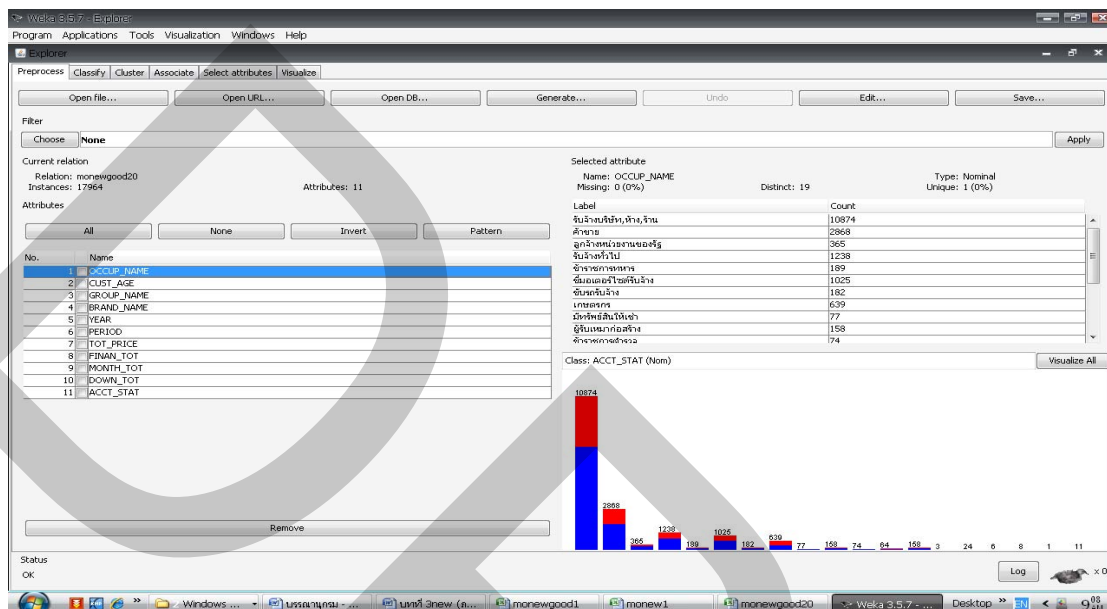


ภาพที่ 4.16 การเปิดโปรแกรม Weka3.5.7



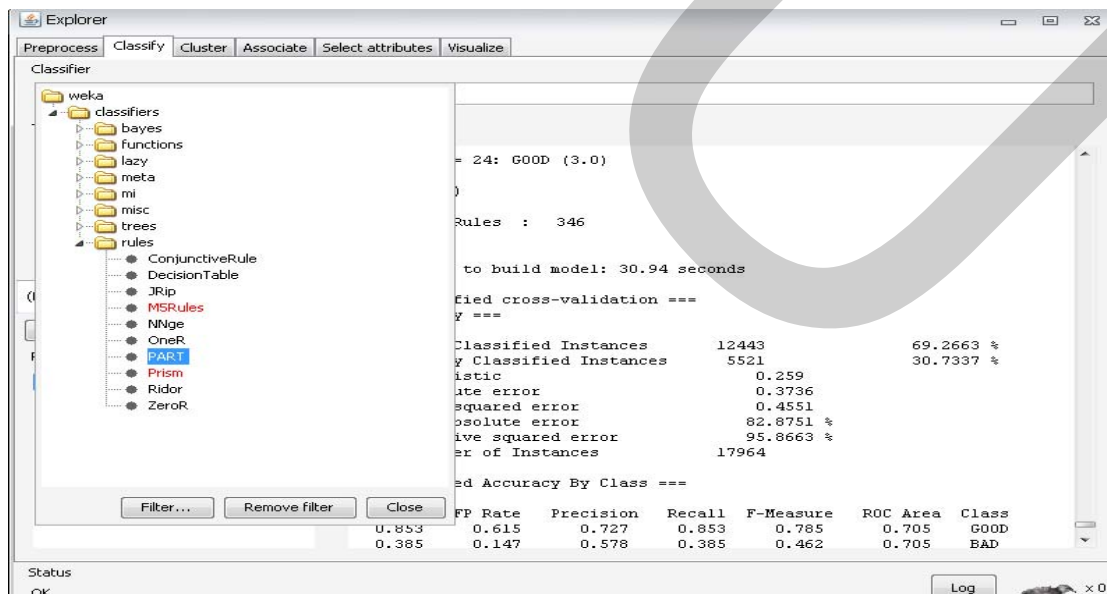
ภาพที่ 4.17 การนำเข้าข้อมูลในโปรแกรม Weka

เปิดโปรแกรม เลือก Application >> Explorer >> Open file เลือกไฟล์ข้อมูลที่ต้องการนำเข้าข้อมูล



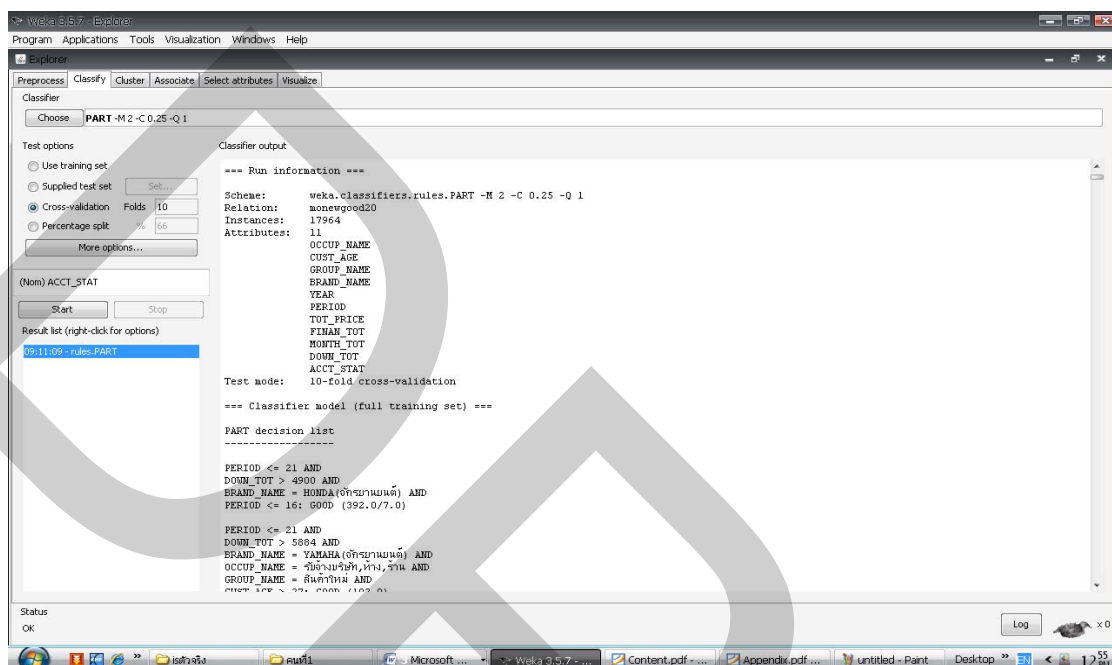
ภาพที่ 4.18 หน้าจอการวิเคราะห์ข้อมูล

หน้าจอนี้จะทำการวิเคราะห์หาจำนวนของข้อมูลที่ใช้ในการวิจัยเพื่อให้ทราบจำนวนเปอร์เซ็นต์ต่อข้อมูลทั้งหมด และค่าเฉลี่ยของข้อมูลในแต่ละแอททริบิวต์

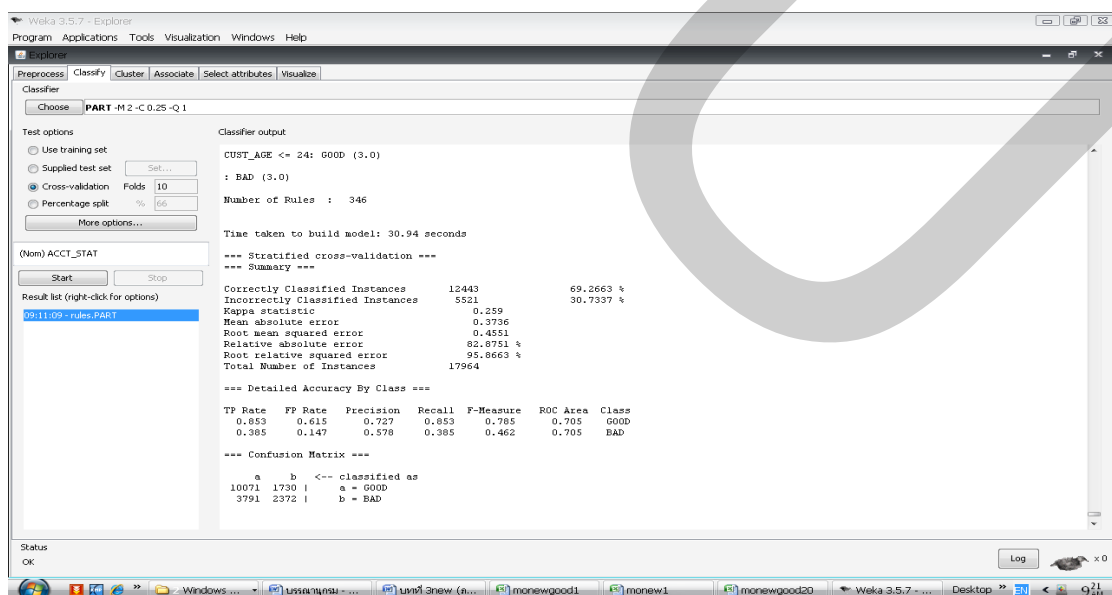


ภาพที่ 4.19 การเลือกเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูล

การจัดกลุ่มข้อมูลทำได้โดยเลือก Classify >> Choose >> Rule >> เลือกเทคนิคการจำแนก
ข้อมูลที่ต้องการ เช่น Part กดปุ่ม Start



ภาพที่ 4.20 หน้าจอแสดงผลการจำแนกกลุ่มข้อมูลด้วยเทคนิค Decision Rule: Part



ภาพที่ 4.21 หน้าจอแสดงผลประสิทธิภาพของโมเดล

4.4 การวิเคราะห์และสรุปผล

4.4.1 ผลการจำแนกกลุ่มแบบ Decision Tree C4.5

จากการทดลองพบว่าเทคนิค Decision Tree C4.5 ให้ผลลัพธ์การจำแนกกลุ่มลูกค้าดีและไม่ดี มีความถูกต้องเป็น 72.9 % และได้กฎจำนวน 1,729 กฎ ภาพที่ 4.1 แสดงผลลัพธ์บางส่วนที่ได้จากการจำแนกกลุ่มแบบ Decision Tree C4.5 ซึ่งจะแสดงผลลัพธ์เป็นกฎในลักษณะของต้นไม้การตัดสินใจ ที่มีกิ่งต้นไม้จำนวนมาก

```

Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Instances:   17964
Attributes:  11      OCCUP_NAME  CUST_AGE  GROUP_NAME  BRAND_NAME
YEAR  PERIOD  TOT_PRICE  FINAN_TOT  MONTH_TOT  DOWN_TOT
ACCT_STAT
PERIOD <= 21: GOOD (5653.0/964.0)
PERIOD > 21
|   DOWN_TOT <= 5100
|   |   PERIOD <= 26
|   |   |   GROUP_NAME = สินค้าใหม่
|   |   |   DOWN_TOT <= 1985
|   |   |   |   FINAN_TOT <= 31800
|   |   |   |   |   YEAR = 2007.0: GOOD (26.0/6.0)
|   |   |   |   |   YEAR = 2004.0: GOOD (384.0/129.0)
|   |   |   |   |   YEAR = 2006.0: GOOD (71.0/18.0)
|   |   |   |   |   YEAR = 2005.0
|   |   |   |   |   |   MONTH_TOT <= 1890
|   |   |   |   |   |   |   MONTH_TOT <= 1815
|   |   |   |   |   |   |   |   FINAN_TOT <= 26000: BAD (7.0/3.0)
|   |   |   |   |   |   |   |   FINAN_TOT > 26000
|   |   |   |   |   |   |   |   |   TOT_PRICE <= 28700
|   |   |   |   |   |   |   |   |   |   CUST_AGE <= 33: GOOD (14.0/3.0)
|   |   |   |   |   |   |   |   |   |   CUST_AGE > 33: BAD (5.0)
|   |   |   |   |   |   |   |   |   |   TOT_PRICE > 28700: BAD (17.0/6.0)
|   |   |   |   |   |   |   |   |   MONTH_TOT > 1815: GOOD (28.0/4.0)
|   |   |   |   |   |   |   |   MONTH_TOT > 1890
|   |   |   |   |   |   |   |   |   TOT_PRICE <= 31244: BAD (6.0/1.0)
|   |   |   |   |   |   |   |   |   TOT_PRICE > 31244: GOOD (6.0/1.0)
|   |   |   |   |   |   |   |   YEAR = 2003.0
|   |   |   |   |   |   |   |   |   CUST_AGE <= 35: GOOD (35.0/13.0)
|   |   |   |   |   |   |   |   |   CUST_AGE > 35: BAD (13.0/3.0)
|   |   |   |   |   |   |   |   YEAR = 2000.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 2002.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 2001.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 2009.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 6.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 1999.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 1.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 0.0: GOOD (0.0)
|   |   |   |   |   |   |   |   YEAR = 1998.0: GOOD (0.0)

```

ภาพที่ 4.22 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Tree C4.5

ผลลัพธ์ที่ได้จะแสดงเงื่อนไขของกฎ โดยค่าของข้อมูลของแอทริบิวต์ที่มีค่าต่อเนื่องจะใช้เครื่องหมาย “<” “<=” และ “>” “>=” ในการพิจารณา ส่วนข้อมูลที่เป็นค่าไม่ต่อเนื่องจะใช้เครื่องหมาย “=” และใช้เงื่อนไข “AND” ในการเชื่อมโยงระหว่างแอทริบิวต์ โดยจะเชื่อมโยงไป

จนถึงแอทริบิวต์ที่จะแบ่งกลุ่มได้ และจะมีข้อมูลความถูกต้องของกฎ ซึ่งได้แก่ จำนวนสัญญาณที่สอดคล้องกับกฎและจำนวนสัญญาณที่ไม่สอดคล้องกับกฎ เช่น

PERIOD <= 21: GOOD (5653.0/964.0)

หมายถึง ถ้าจำนวนจำนวนงวดที่ผ่อนชำระน้อยกว่า 21 เดือนจะเป็นลูกค้าดี และมีข้อมูลที่สอดคล้องกับกฎนี้หรือให้ผลลัพธ์ตรงกันจำนวน 5653 ตัวอย่าง และข้อมูลให้ผลลัพธ์ไม่ตรงกัน 964 ตัวอย่าง

4.4.2 ผลการจำแนกกลุ่มแบบ Decision Tree: Partial Rules

เทคนิค Decision Tree: Partial Rules ให้ผลลัพธ์การจำแนกกลุ่มลูกค้าดีและไม่ดี มีความถูกต้องเป็น 72.7 % และได้กฎจำนวน 346 กฎ ภาพที่ 4.2 แสดงผลลัพธ์บางส่วนที่ได้จากการจำแนกกลุ่มแบบ Decision Tree: Partial Rules ซึ่งแสดงผลลัพธ์เป็นกฎที่เข้าใจง่ายให้จำนวนกฎน้อย

```
Scheme:      weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Instances:    17964
Attributes:   11      OCCUP_NAME  CUST_AGE  GROUP_NAME  BRAND_NAME  YEAR  PERIOD
                TOT_PRICE  FINAN_TOT  MONTH_TOT  DOWN_TOT    ACCT_STAT

PART decision list
PERIOD <= 21 AND DOWN_TOT > 4900 AND BRAND_NAME = HONDA(จักรยานยนต์) ANDPERIOD <= 16: GOOD
(392.0/7.0)
PERIOD <= 21 ANDDOWN_TOT > 5884 ANDBRAND_NAME = YAMAHA(จักรยานยนต์) ANDGROUP_NAME = สิ้นค้าใหม่
ANDOCCUP_NAME = รับจ้างบริษัท, ห้าง, ร้าน AND
YEAR = 2004.0: GOOD (45.0/2.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDCUST_AGE > 43 ANDYEAR = 2003.0: GOOD (53.0/1.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDDOWN_TOT > 5884 ANDBRAND_NAME = YAMAHA(จักรยานยนต์)
ANDOCCUP_NAME = รับจ้างบริษัท, ห้าง, ร้าน ANDTOT_PRICE > 36400: GOOD (116.0/1.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDCUST_AGE > 34 ANDPERIOD <= 16 AND
BRAND_NAME = YAMAHA(จักรยานยนต์) ANDMONTH_TOT > 2115: GOOD (119.0/3.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDOCCUP_NAME = พนักงานรัฐวิสาหกิจ: GOOD (48.0/4.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDOCCUP_NAME = ลูกจ้างหน่วยงานของรัฐ ANDGROUP_NAME = สิ้นค้าเก่า(02)
ANDYEAR = 2002.0: GOOD (13.0/1.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDPERIOD <= 7 ANDYEAR = 2002.0: GOOD (159.0/20.0)
PERIOD <= 21 ANDMONTH_TOT > 1395 ANDYEAR = 2001.0 ANDOCCUP_NAME = รับจ้างบริษัท, ห้าง, ร้าน
ANDCUST_AGE > 32: GOOD (125.0/7.0)
PERIOD <= 26 ANDPERIOD <= 19 ANDMONTH_TOT > 1395 ANDOCCUP_NAME = ลูกจ้างหน่วยงานของรัฐ ANDYEAR =
2004.0: GOOD (13.0/1.0)
```

ภาพที่ 4.23 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Tree: Partial Rules

ตัวอย่าง เช่น PERIOD <= 26 AND DOWN_TOT > 5100 AND GROUP_NAME = สิ้นค้าใหม่ AND BRAND_NAME = YAMAHA (จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัท ห้าง, ร้าน AND YEAR = 2006 AND MONTH_TOT > 2315: GOOD (30.0)

หมายถึง จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด และเงินดาวน์มากกว่า 5,100 บาท และเป็นรถใหม่ ยี่ห้อยามาฮ่า และลูกค้ามีอาชีพรับจ้างบริษัทห้างร้าน และเป็นรถปี 2006 และค้างงวด

มากกว่า 2,315 จะเป็นลูกค้าดี และมีข้อมูลที่สอดคล้องกับกฎนี้หรือให้ผลลัพธ์ตรงกัน จำนวน 30 ตัวอย่าง

4.4.3 ผลการจำแนกกลุ่มแบบ Decision Table

เทคนิค Decision Table ให้ผลลัพธ์การจำแนกกลุ่มลูกค้าดีและไม่ดี มีความถูกต้องเป็น 71.9% และได้กฎจำนวน 184 กฎ ภาพที่ 4.3 แสดงผลลัพธ์บางส่วนที่ได้จากการจำแนกกลุ่มแบบ Decision Table

```

Scheme:      weka.classifiers.rules.DecisionTable
Instances:    17964
Attributes:   11 OCCUP_NAME  CUST_AGE  GROUP_NAME  BRAND_NAME
              YEAR  PERIOD  TOT_PRICE  FINAN_TOT  MONTH_TOT  DOWN_TOT
              ACCT_STAT
Number of Rules : 184
Non matches covered by Majority class.
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 65
  Merit of best subset found: 69.456
Evaluation (for feature selection): CV (leave one out)
Feature set: 2,3,6,7,11
Time taken to build model: 3.45 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      12498      69.5725 %
Incorrectly Classified Instances    5466      30.4275 %
Kappa statistic                    0.2483
Mean absolute error                 0.3934
Root mean squared error            0.4433
Relative absolute error            87.2707 %
Root relative squared error        93.3757 %
Total Number of Instances          17964

=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
  0.88    0.657   0.719     0.88    0.792     0.716    GOOD
  0.343   0.12     0.599     0.343   0.436     0.716    BAD

=== Confusion Matrix ===
      a      b  <-- classified as
10384 1417 |    a = GOOD
 4049 2114 |    b = BAD
  
```

ภาพที่ 4.24 ผลลัพธ์บางส่วนจากการจำแนกกลุ่มแบบ Decision Table

4.4.4 การวิเคราะห์และการคัดเลือกกฎ

จากผลการทดลองพบว่า การจำแนกกลุ่มแบบ ให้ผลลัพธ์ที่มีความถูกต้องที่ใกล้เคียงกัน คือ 72.9 เปอร์เซ็นต์ 72.7 เปอร์เซ็นต์ และ 71.9 เปอร์เซ็นต์ แต่ให้กฎแตกต่างกันคือ 1729, 346 และ

184 กฎตามลำดับ ซึ่ง Decision Tree: Partial Rules สามารถแบ่งกลุ่มได้ตามเงื่อนไขได้ชัดเจนกว่า และสามารถแสดงกฎให้เห็นสามารถนำไปวิเคราะห์กฎต่อไปได้ โดยแบ่งเป็น 4 กลุ่มหลักๆ ดังนี้

1) กฎที่จำแนกกลุ่มลูกค้าดี (Good) ได้ถูกต้อง 100 เปอร์เซ็นต์ มีจำนวน 85 กฎ ซึ่งกฎที่อยู่ในกลุ่มนี้เป็นกฎที่มีความน่าเชื่อถือมากที่สุด เพราะสามารถจำแนกกลุ่มตัวอย่างได้ถูกต้องทั้งหมด

ตัวอย่างกฎในกลุ่มนี้ เช่น PERIOD <= 19 AND MONTH_TOT > 1925 AND BRAND_NAME = HONDA (จักรยานยนต์) AND FINAN_TOT <= 32400 AND OCCUP_NAME = ค้าขาย: GOOD (58.0)

หมายถึง จำนวนงวดน้อยกว่าหรือเท่ากับ 19 งวด เงินค่างวดมากกว่า 1,925 บาท ยี่ห้อฮอนด้า ยอดกู้ น้อยกว่าหรือเท่ากับ 32,400 อาชีพค้าขาย เป็นกลุ่มลูกค้าดี โดยมีข้อมูลที่สอดคล้องกับกฎนี้ จำนวน 58 ตัวอย่าง

2) กฎที่จำแนกกลุ่มลูกค้าไม่ดี (Bad) ได้ถูกต้อง 100 เปอร์เซ็นต์ มีจำนวน 41 กฎ ซึ่งกฎที่อยู่ในกลุ่มนี้เป็นอีกกลุ่มหนึ่งที่มีความน่าเชื่อถือมากที่สุด เพราะสามารถจำแนกกลุ่มตัวอย่างได้ถูกต้องทั้งหมด

ตัวอย่างกฎในกลุ่มนี้ เช่น PERIOD > 26 AND YEAR = 2006 AND OCCUP_NAME = ค้าขาย AND FINAN_TOT > 32700 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = HONDA (จักรยานยนต์) AND DOWN_TOT <= 200 AND CUST_AGE <= 45: BAD (25.0)

หมายถึง จำนวนงวดมากกว่า 26 งวด เป็นรถปี 2006 อาชีพค้าขาย ยอดกู้มากกว่า 32,700 บาท เป็นรถใหม่ ยี่ห้อฮอนด้า เงินค่างวดน้อยกว่าหรือเท่ากับ 200 บาท อายุน้อยกว่าหรือเท่ากับ 45 ปี เป็นกลุ่มลูกค้าไม่ดี โดยมีข้อมูลที่สอดคล้องกับกฎนี้ จำนวน 25 ตัวอย่าง

3) กฎที่จำแนกกลุ่มลูกค้าดี (Good) ได้ถูกต้องไม่ถึง 100 เปอร์เซ็นต์ มีจำนวน 131 กฎ ตัวอย่างกฎในกลุ่มนี้ เช่น PERIOD <= 19 AND YEAR = 2005 AND GROUP_NAME = สินค้าใหม่ AND CUST_AGE > 24 AND OCCUP_NAME = รับจ้างบริษัท ห้าง ร้าน AND BRAND_NAME = HONDA (จักรยานยนต์) AND PERIOD > 12 AND CUST_AGE <= 38: GOOD (18.0/1.0)

หมายถึง จำนวนงวดมากกว่า 12 งวดและน้อยกว่าหรือเท่ากับ 19 งวด เป็นรถปี 2005 รถใหม่ ลูกค้ามีอายุมากกว่า 24 ปี อาชีพรับจ้างบริษัท ห้าง ร้าน รถยี่ห้อฮอนด้า ลูกค้าอายุน้อยกว่าหรือเท่ากับ 38 ปี เป็นกลุ่มลูกค้าดี โดยมีข้อมูลที่สอดคล้องกับกฎนี้ จำนวน 18 ตัวอย่าง และมีข้อมูลที่ไม่สอดคล้องกับกฎนี้จำนวน 1 ตัวอย่าง

4) กฎที่จำแนกกลุ่มลูกค้าไม่ดี (Bad) ได้ถูกต้องไม่ถึง 100 เปอร์เซ็นต์ มีจำนวน 89 กฎ ตัวอย่างกฎในกลุ่มนี้ เช่น

YEAR = 2003 AND OCCUP_NAME = รับจ้างบริษัท ห้าง ร้าน AND BRAND_NAME = SUZUKI (จักรยานยนต์) AND GROUP_NAME = สินค้าใหม่ AND FINAN_TOT <= 30400: BAD (20.0/8.0)

หมายถึง รถปี 2003 อาชีพรับจ้างบริษัทห้างร้าน ยี่ห้อซูซูกิ เป็นรถใหม่ ยอดกั๊นน้อยกว่าหรือเท่ากับ 30,400 บาท เป็นกลุ่มลูกค้าเสีย โดยมีข้อมูลที่สอดคล้องกับกฎนี้ จำนวน 20 ตัวอย่าง และมีข้อมูลที่ไม่สอดคล้องกับกฎนี้จำนวน 8 ตัวอย่าง

แต่เนื่องจากกฎที่ได้มีจำนวนมาก และกฎบางกฎมีเงื่อนไขไม่เหมาะสม บางกฎมีเงื่อนไขที่ซ้ำซ้อน ดังนั้น 'ผู้เชี่ยวชาญจะทำหน้าที่ตรวจสอบ แก้ไขและลบกฎบางกฎทิ้งไป ดังตัวอย่างต่อไปนี้

1. กฎที่ไม่เหมาะสมและถูกลบทิ้ง เช่น

OCCUP_NAME = ข้าราชการพลเรือน : GOOD (4.0)

หมายถึง ถ้าลูกค้ามีอาชีพเป็น ข้าราชการพลเรือน จะเป็นลูกค้าดี กฎไม่เหมาะสมเนื่องจาก มีเงื่อนไขในการพิจารณาจากอาชีพเพียงเงื่อนไขเดียว และมีกลุ่มตัวอย่างสนับสนุนจำนวนน้อยมาก

2. กฎที่มีเงื่อนไขที่ซ้ำซ้อนและถูกปรับให้เข้าใจง่ายขึ้น เช่น

PERIOD <= 26 AND YEAR = 2005 AND GROUP_NAME = สิ้นค้าใหม่ AND CUST_AGE > 24 AND PERIOD <= 19 AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND BRAND_NAME = HONDA (จักรยานยนต์) AND PERIOD > 12 AND CUST_AGE <= 38: GOOD (18.0/1.0)

เงื่อนไขที่ซ้ำซ้อนคือ **PERIOD <= 26 PERIOD <= 19** ปรับใหม่เป็น **PERIOD <= 19** เพียงค่าเดียวได้กฎใหม่คือ

PERIOD <= 19 AND YEAR = 2005 AND GROUP_NAME = สิ้นค้าใหม่ AND CUST_AGE > 24 AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND BRAND_NAME = HONDA (จักรยานยนต์) AND PERIOD > 12 AND CUST_AGE <= 38: GOOD (18.0/1.0)

หมายถึง จำนวนงวดมากกว่า 12 งวดและน้อยกว่าหรือเท่า 19 งวด เป็นรถปี 2005 รถใหม่ ลูกค้ามีอายุมากกว่า 24 ปี อาชีพรับจ้างบริษัทห้างร้าน รถยี่ห้อฮอนด้า ลูกค้าอายุน้อยกว่าหรือเท่ากับ 38 ปี เป็นกลุ่มลูกค้าดี โดยมีข้อมูลที่สอดคล้องกับกฎนี้ จำนวน 18 ตัวอย่าง และมีข้อมูลที่ไม่สอดคล้องกับกฎนี้จำนวน 1 ตัวอย่าง

กฎที่อยู่ในกลุ่มลูกค้าดี (Good)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวดเงินดาวน์ มากกว่า 5,100 บาท รถใหม่ ยี่ห้อ ยามาฮ่า อาชีพรับจ้างบริษัทห้างร้าน รถปี 2006 ค่างวดมากกว่า 2,315 จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 5100 AND GROUP_NAME = สินค้าใหม่ AND BR AND_NAME = YAMAHA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND YEAR = 2006 AND MONTH_TOT > 2315: GOOD (30.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวน้นมากกว่า 4,935 บาท เป็นรถใหม่ ยี่ห้อฮอนด้า อาชีพรับจ้างบริษัทห้างร้าน รถปี 2005 อายุมากกว่า 22 ปี จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 4935 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = HONDA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND YEAR = 2005 AND CUST_AGE > 22 : GOOD (40.0/3.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวน้นมากกว่า 4,095 บาท เป็นรถใหม่ ยี่ห้อฮอนด้า อาชีพรับจ้างบริษัทห้างร้าน อายุมากกว่า 30 ปีแต่น้อยกว่าหรือเท่ากับ 42 ปี รถปี 2003 จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 4095 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = HONDA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND CUST_AGE > 30 AND YEAR = 2003 AND CUST_AGE <= 42 : GOOD (49.0/1.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวน้นมากกว่า 5160 บาท เป็นรถใหม่ ยี่ห้อยามาฮ้า อาชีพรับจ้างบริษัทห้างร้าน จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 5160 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = YAMAHA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัท, ห้าง,ร้าน AND YEAR = 2004: GOOD (93.0/16.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวน้นมากกว่า 5,160 บาท เป็นรถใหม่ ยี่ห้อยามาฮ้า อาชีพรับจ้างทั่วไป ค่างวดน้อยกว่าหรือเท่ากับ 2,250 จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 5160 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = YAMAHA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างทั่วไป AND MONTH_TOT <= 2250: GOOD (30.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวน้นมากกว่า 4,935บาท เป็นรถใหม่ ยี่ห้อฮอนด้า อาชีพรับจ้างทั่วไป ค่างวดน้อยกว่าหรือเท่ากับ 2,250 จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND OWN_TOT > 4935 AND GROUP_NAME = สิ้นค้าใหม่ AND BRAND_NAME = HONDA(จักรยานยนต์) AND DOWN_TOT > 15810: GOOD (92.0/1.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด เงินคาวนั้มากกว่า 9,032บาท เป็นรถใหม่ ยี่ห้อ ยามาฮ่าอาชีพค้าขาย จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 4935 AND GROUP_NAME = สิ้นค้าใหม่ AND BRAND_NAME = YAMAHA(จักรยานยนต์) AND OCCUP_NAME = ค้าขาย AND DOWN_TOT > 9032: GOOD (33.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวด อายุมากกว่า 31 ปี เป็นรถใหม่ เงินคาวนั้มากกว่า 2,596 บาท รถปี 2006 อาชีพ รับจ้างบริษัท อายุมากกว่า 32 ปีจะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND CUST_AGE > 31 AND GROUP_NAME = สิ้นค้าใหม่ AND DOWN_TOT > 2596 AND YEAR = 2006.0 AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND CUST_AGE > 32 : GOOD (25.0/1.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 19งวดที่มีค่างวดมากกว่า 1,925 บาท ยี่ห้อ ฮอนด้า ยอดกู้ น้อยกว่าหรือเท่ากับ 32,400 บาท อาชีพค้าขาย ค่างวดมากกว่า 2,655 บาท จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 19 AND MONTH_TOT > 1925 AND BRAND_NAME = HONDA (จักรยานยนต์) AND FINAN_TOT <= 32400 AND OCCUP_NAME = ค้าขาย AND MONTH_TOT > 2655: GOOD (58.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวดที่มีเงินคาวนั้มากกว่า 4,095 บาทเป็นรถใหม่ ยี่ห้อฮอนด้า ค่างวดอาชีพลูกจ้างหน่วยงานของรัฐ จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND DOWN_TOT > 4095 AND GROUP_NAME = สิ้นค้าใหม่ AND BRAND_NAME = HONDA (จักรยานยนต์) AND OCCUP_NAME = ลูกจ้างหน่วยงานของรัฐ: GOOD (16.0)

จำนวนงวดน้อยกว่าหรือเท่ากับ 26 งวดอาชีพรับจ้างบริษัทห้างร้าน รถปี 2004 ยี่ห้อ ฮอนด้าที่มีเงินคาวนั้ น้อยกว่าหรือเท่ากับ 3,227บาท เป็นรถเก่า จะเป็นลูกค้าดี

กฎที่อ้างอิง: PERIOD <= 26 AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND YEAR = 2004 AND BRAND_NAME = HONDA(จักรยานยนต์) AND DOWN_TOT <= 3227 AND GROUP_NAME = สิ้นค้าเก่า(02): GOOD (326.0/67.0)

เงินดาวน์ น้อยกว่า 4,095 บาท รถใหม่ ยี่ห้อฮอนด้าอาชีพรับจ้างบริษัท ห้าง ร้าน จำนวนงวดน้อยกว่าหรือเท่ากับ 19 งวดรถปี 2004 จะเป็นลูกค้าดี

กฎที่อ้างอิง: DOWN_TOT > 4095 AND GROUP_NAME = สินค้าใหม่ AND BRAND_NAME = HONDA(จักรยานยนต์) AND OCCUP_NAME = รับจ้างบริษัทห้างร้าน AND PERIOD <= 19 AND YEAR = 2004.0: GOOD (29.0/1.0)

กฎที่อยู่ในกลุ่มลูกค้าไม่ดี (Bad)

จำนวนงวดมากกว่า 26 งวดรถปีน้อยกว่าหรือเท่ากับ 2006 แต่มากกว่าปี 2003 ยี่ห้อยามาฮ่า อาชีพรับจ้างทั่วไปเป็นรถใหม่ ค่างวดน้อยกว่าหรือเท่ากับ 2,670 เงินดาวน์ น้อยกว่า หรือเท่ากับ 3,784บาท จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: PERIOD > 26 AND YEAR <= 2006 AND OCCUP_NAME = รับจ้างทั่วไป AND YEAR > 2003 AND BRAND_NAME = YAMAHA (จักรยานยนต์) AND GROUP_NAME = สินค้าใหม่ AND MONTH_TOT <= 2670 AND DOWN_TOT <= 3784: BAD (25.0/5.0)

จำนวนงวดมากกว่า 26 งวดรถปีน้อยกว่าหรือเท่ากับ 2006 เงินดาวน์ น้อยกว่า หรือเท่ากับ 430 บาท อาชีพ ค้าขาย ยี่ห้อฮอนด้า ราคามากกว่า 33,900บาท จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: PERIOD > 26 AND YEAR <= 2006 AND DOWN_TOT <= 430 AND OCCUP_NAME = ค้าขาย AND BRAND_NAME = HONDA(จักรยานยนต์) AND TOT_PRICE > 33900: BAD (11.0)

จำนวนงวดมากกว่า 24 งวดเงินดาวน์ น้อยกว่า หรือเท่ากับ 1,065 บาทแต่มากกว่า 346 บาท รถปีน้อยกว่าหรือเท่ากับ 2006 อาชีพ ค้าขายยี่ห้อยามาฮ่า จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: PERIOD > 24 AND DOWN_TOT <= 1065 AND YEAR <= 2006 AND OCCUP_NAME = ค้าขาย AND BRAND_NAME = YAMAHA(จักรยานยนต์) AND DOWN_TOT > 346: BAD (33.0/5.0)

จำนวนงวดมากกว่า 26 รถปีน้อยกว่าหรือเท่ากับ 2006 แต่มากกว่า 2004 งวดเงินดาวน์ น้อยกว่า หรือเท่ากับ 43 บาทอาชีพ รับจ้างบริษัทห้างร้าน ยี่ห้อฮอนด้าค่างวดมากกว่า 1,780 บาทแต่น้อยกว่าหรือเท่ากับ 2,235 บาท อายุมากกว่า 23 จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: PERIOD > 26 AND YEAR <= 2006 AND DOWN_TOT <= 43 AND OCCUP_NAME = รับจ้างบริษัท ห้าง ร้าน AND YEAR > 2004 AND BRAND_NAME =

HONDA(จักรยานยนต์) AND YEAR <= 2005 AND MONTH_TOT > 1780 AND FINAN_TOT > 33900 AND MONTH_TOT <= 2235 AND CUST_AGE > 23: BAD (58.0/11.0)

จำนวนงวดมากกว่า 26 ยี่ห้อยามาฮ้า อาชีพ ข้าราชการทหาร อายุไม่น้อยกว่าหรือเท่ากับ 39 ยอดกู้ไม่น้อยกว่าหรือเท่ากับ 39,800 บาท จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: PERIOD > 26 AND BRAND_NAME = YAMAHA (จักรยานยนต์) AND OCCUP_NAME = ข้าราชการทหาร AND CUST_AGE <= 39 AND FINAN_TOT <= 39800: BAD (12.0)

อาชีพ เกษตรกร ยี่ห้อยอนด้า รถเก่า เงินคาวน้นน้อยกว่าหรือเท่ากับ 1,950 บาท จำนวนงวดมากกว่า 23 งวด ราคา น้อยกว่าหรือเท่ากับ 23,500 บาท จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: OCCUP_NAME = เกษตรกร AND BRAND_NAME = HONDA (จักรยานยนต์) AND GROUP_NAME = สิ้นค้าเก่า (02) AND DOWN_TOT <= 1950 AND PERIOD > 23 AND TOT_PRICE <= 23500: BAD (12.0)

อาชีพ ค้าขาย ยี่ห้อยอนด้า รถเก่า เงินคาวน้นน้อยกว่าหรือเท่ากับ 3,550 บาท เป็นรถเก่า ราคา น้อยกว่า25,680 บาท จำนวนงวดน้อยกว่า หรือเท่ากับ 27 งวด จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: OCCUP_NAME = ค้าขาย AND BRAND_NAME = HONDA (จักรยานยนต์) AND DOWN_TOT <= 3550 AND GROUP_NAME = สิ้นค้าเก่า(02) AND TOT_PRICE > 25680 AND PERIOD <= 27: BAD (16.0/4.0)

อาชีพ รับจ้างบริษัท ห้าง ร้าน จำนวนงวดมากกว่า 27 งวด ยี่ห้อยามาฮ้า ราคา น้อยกว่าหรือเท่ากับ38,214 บาท รถปีน้อยกว่า 2005 ยอดกู้มากกว่า 34,500 บาท จะเป็นลูกค้าไม่ดี

กฎที่อ้างอิง: OCCUP_NAME = รับจ้างบริษัท ห้าง ร้าน AND PERIOD > 27 AND BRAND_NAME = YAMAHA (จักรยานยนต์) AND TOT_PRICE <= 38214 AND YEAR < 2005 AND FINAN_TOT > 34500: BAD (12.0)

สรุปกฎจากผลการทดสอบการจำแนกกลุ่มข้อมูลแบบ Decision Rules: Part เป็นการจับกลุ่มที่เหมาะสมที่สุดในการทดสอบเนื่องจากมีกฎจำนวนมากแต่กฎบางกฎมีเงื่อนไขไม่เหมาะสม บางกฎมีเงื่อนไขที่ซ้ำซ้อน ดังนั้นผู้เชี่ยวชาญจะทำหน้าที่ตรวจสอบ และลบกฎบางกฎทิ้งไป โดยตัดกฎที่ไม่เหมาะสมออก และปรับกฎที่มีเงื่อนไขที่ซ้ำซ้อนให้เข้าใจง่ายขึ้น จากทั้งหมด 346 กฎ เหลือเพียง 188 กฎ

ตารางที่ 4.4 การหาค่ากฎที่น่าเชื่อถือ

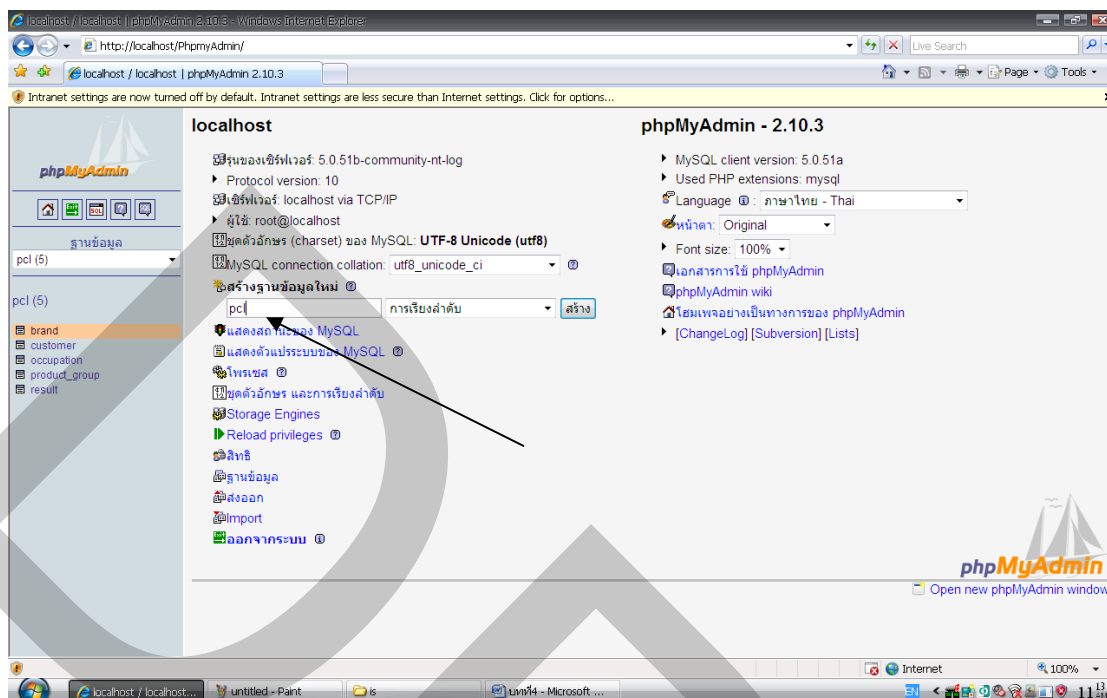
กฎ	อาชีพ	อายุ	กลุ่มสินค้า	ยี่ห้อสินค้า	ปีที่ผลิต	งวด	ราคา	ยอดกู้	ค่างวด	ค่าวัน	สถานะภาพ	True	False	%T	%F
1				HONDA		<= 21				> 4900	GOOD	392	7	3.32	0.11
2	รับจ้างบริษัท	> 27	สินค้าใหม่	YAMAHA		<= 21				> 5884	GOOD	102	0	0.86	0.00
3		> 43	สินค้าเก่า			<= 16				> 9000	GOOD	283	17	2.40	0.28
4	พนักงานรัฐวิสาหกิจ					<= 21			> 1395		GOOD	44	3	0.37	0.05
5	รับจ้างทั่วไป		สินค้าใหม่	YAMAHA	2005	<= 21	> 37800		> 1395		GOOD	22	0	0.19	0.00
6	รับจ้างทั่วไป	> 25	สินค้าใหม่	HONDA	2001	<= 21			> 1395		GOOD	27	0	0.23	0.00
7	ข้าราชการทหาร	> 34	สินค้าเก่า			<= 21			> 1395		GOOD	17	0	0.14	0.00
8	ลูกจ้างหน่วยงานของรัฐ	> 25		HONDA		<= 21			> 1395		GOOD	26	0	0.22	0.00
9	ข้าราชการตำรวจ	<= 38		HONDA		<= 21			> 1395		GOOD	6	0	0.05	0.00
10	รับจ้างบริษัท, ห้าง, ร้าน				<= 2001	<= 10			> 1925		GOOD	130	12	1.10	0.19
11	ขี่มอเตอร์ไซด์รับจ้าง			HONDA	> 1999	<= 19			> 1395	> 3000	GOOD	28	1	0.24	0.02
12	รับจ้างทั่วไป				> 1999	<= 19			> 1395		GOOD	66	4	0.56	0.06
13	รับจ้างบริษัท			HONDA	> 2003	<= 19				> 3000	GOOD	13.6	10	0.12	0.16
14	ค้าขาย	> 39	สินค้าใหม่		> 1998	<= 16					GOOD	51	0	0.43	0.00
15	รับจ้างบริษัท	> 27		HONDA	> 1998	<= 16		<= 38300	<= 15400		GOOD	81	2	0.69	0.03
16	รับจ้างบริษัท	<= 47	สินค้าใหม่	YAMAHA	> 1999	<= 16			> 1395		GOOD	108	8	0.91	0.13
17	รับจ้างบริษัท	<= 47			> 1999	<= 21			> 1395		GOOD	28	0	0.24	0.00
18	รับจ้างทั่วไป		สินค้าใหม่	YAMAHA	> 2003	<= 26			> 2250	> 5160	GOOD	35	7	0.30	0.11
19	เกษตรกร		สินค้าใหม่	YAMAHA		<= 26			> 5160		GOOD	31	1	0.26	0.02
20	ขี่มอเตอร์ไซด์รับจ้าง				> 1998	<= 16					GOOD	52	4	0.44	0.06
21			สินค้าใหม่	YAMAHA		<= 26			>		GOOD	108	4	0.91	0.06

เมื่อได้ผลของการทำเหมืองข้อมูลมาแล้วทำการเลือกกฎที่มีความน่าเชื่อถือมากที่สุดโดยหาจากกฎที่มีค่าความเป็นจริงมากที่สุดเพื่อใช้เป็นเงื่อนไขในการเขียนโปรแกรมวิเคราะห์สินเชื่อ

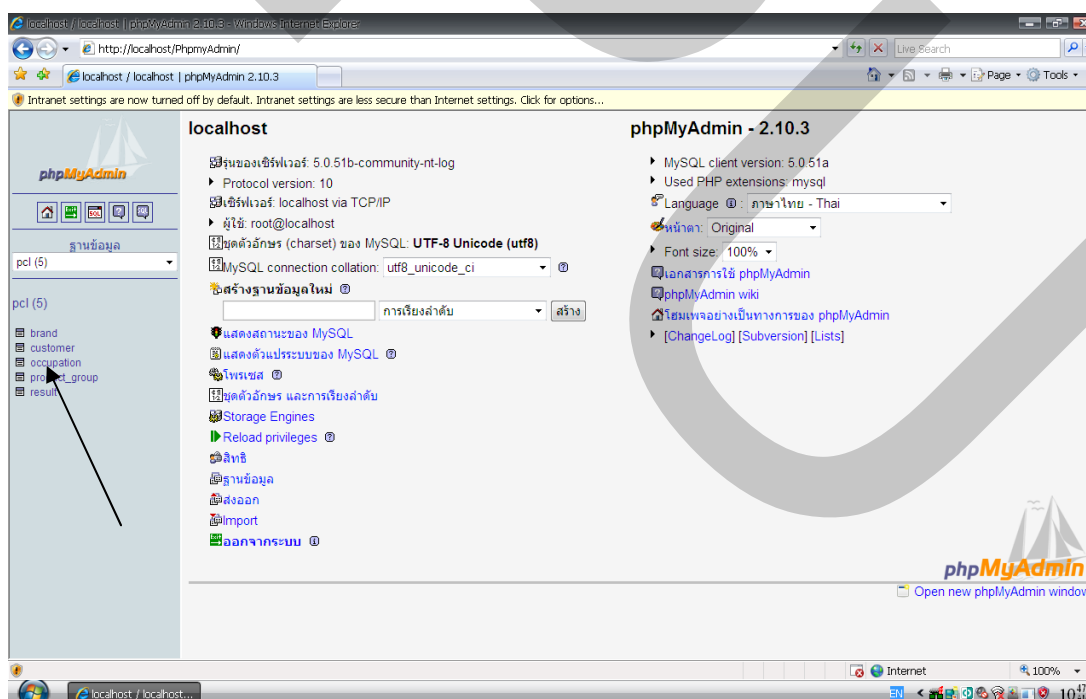
4.5 การพัฒนาระบบสนับสนุนการตัดสินใจการอนุมัติสินเชื่อ

ในขั้นตอนนี้เป็นการนำกฎที่ได้จากการทำเหมืองข้อมูลมาเป็นเงื่อนไขในการพัฒนาโปรแกรมระบบสินเชื่อออนไลน์ให้เป็นเครื่องมือช่วยในการตัดสินใจอนุมัติสินเชื่อเข้าชื่อรถจักรยานยนต์ โปรแกรมที่ใช้ในการพัฒนาเป็นโปรแกรม PHP และใช้ฐานข้อมูล My SQL

4.5.1 การออกแบบฐานข้อมูล เปิด URL <http://localhost/PhpmyAdmin/> เพื่อเข้าไปในโปรแกรมฐานข้อมูล My SQL ใส่ user name และ Password เมื่อเข้าฐานข้อมูลแล้วทำการสร้างฐานข้อมูลชื่อ PCL ต่อไปทำการสร้าง ตาราง 5 ตาราง ชื่อ brand เพื่อเก็บข้อมูลยี่ห้อสินค้า customer เพื่อเก็บข้อมูลรายละเอียดลูกค้าที่ต้องการบันทึก occupation เพื่อเก็บข้อมูลอาชีพ product group เพื่อเก็บข้อมูลกลุ่มสินค้า result เพื่อเก็บข้อมูลสถานะของลูกค้า



ภาพที่ 4.25 การสร้างฐานข้อมูล



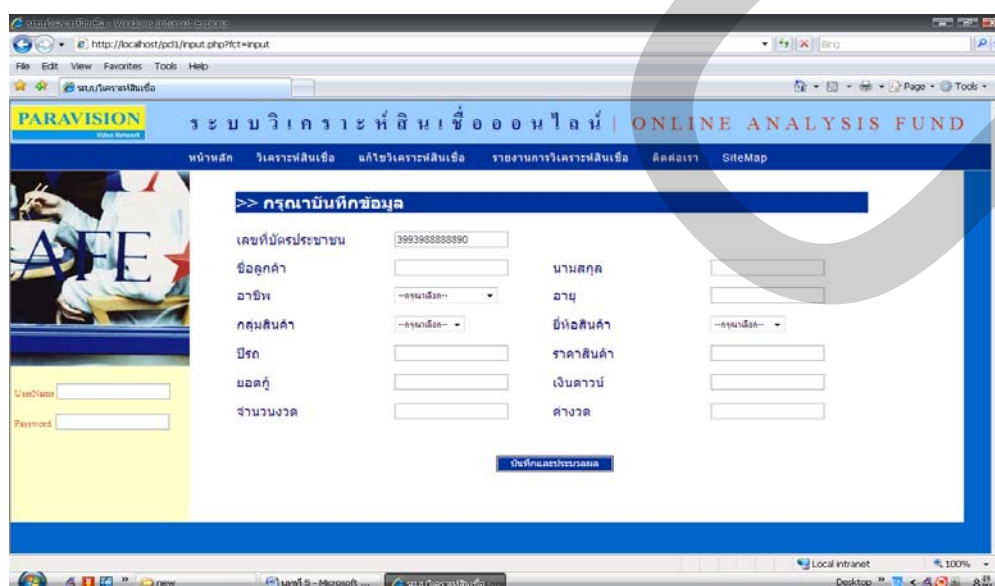
ภาพที่ 4.26 การสร้างตารางในฐานข้อมูล

4.5.2 การออกแบบหน้าจอ Input

ใช้โปรแกรม Dreamweaver CS4 ในการออกแบบจะประกอบด้วยหน้าจอหลักเพื่อแสดงเมนูทั้งหมดของโปรแกรม หน้าจอวิเคราะห์สินเชื่อเพื่อบันทึกข้อมูลลูกค้าที่ต้องการวิเคราะห์ การอนุมัติสินเชื่อ หน้าจอแก้ไขรายงานการวิเคราะห์สินเชื่อเพื่อนำข้อมูลเดิมที่เคยบันทึกไปแล้วกลับมาแก้ไข หน้าจอรายงานวิเคราะห์สินเชื่อเพื่อแสดงรายงานการลูกค้าที่ขอสินเชื่อ หน้าจอติดต่อเราเพื่อที่จะติดต่อ Admin ในกรณีโปรแกรมมีปัญหา

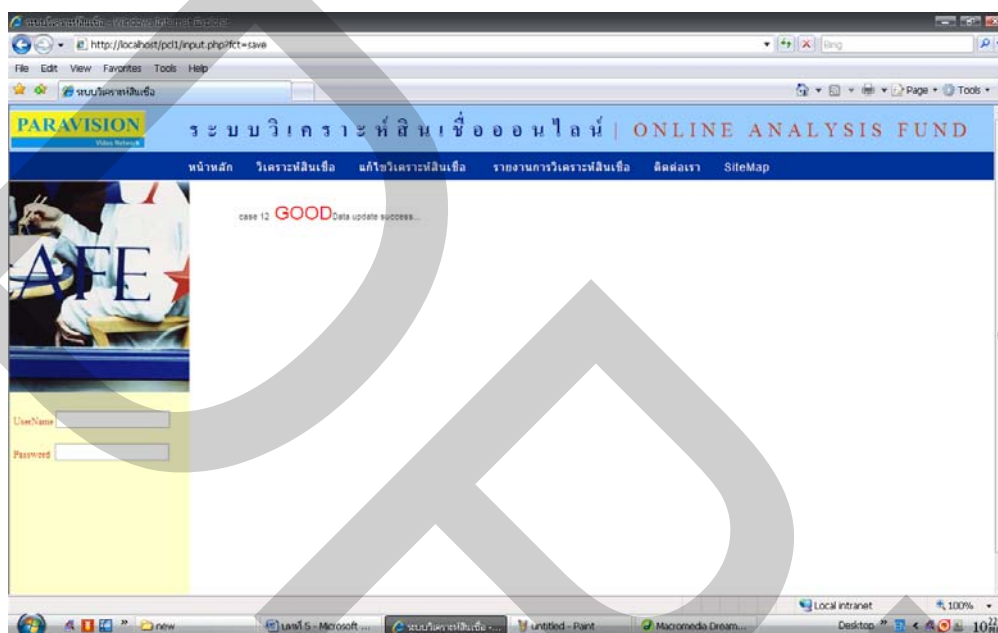


ภาพที่ 4.27 หน้าจอหลักของระบบวิเคราะห์สินเชื่อออนไลน์

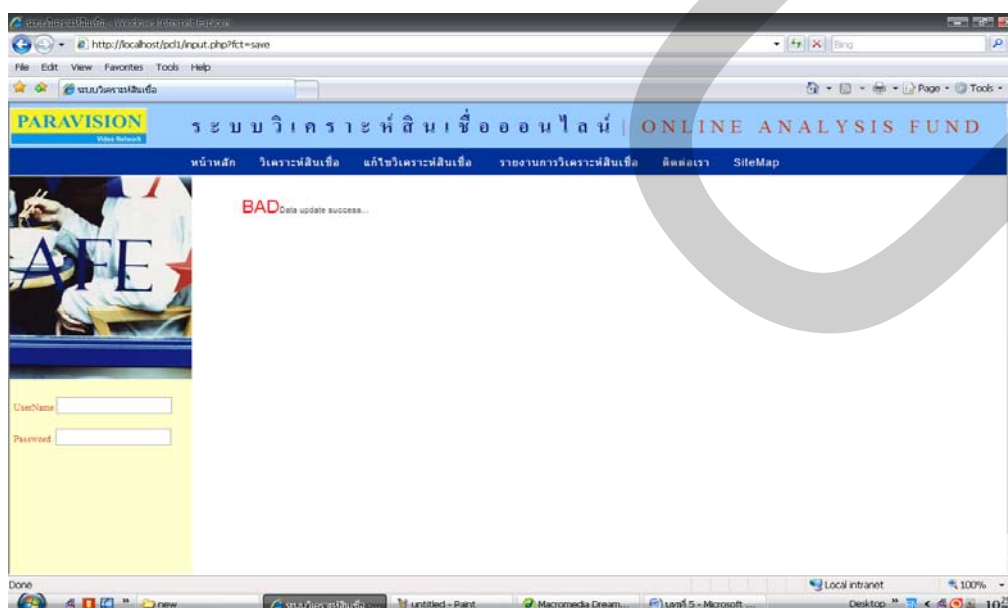


ภาพที่ 4.28 หน้าจอการรับข้อมูลลูกค้าที่ต้องการขอสินเชื่อ

เมื่อผู้ใช้งานทำการป้อนข้อมูลรายละเอียดของลูกค้าที่ต้องการขอสินเชื่อตามรายการต่างๆที่ ออกแบบไว้แล้ว ทำการคลิกที่ปุ่มบันทึกและประมวลผล โปรแกรมจะทำการประมวลผลและ จำแนกลูกค้านี้ว่าเป็น ลูกค้าที่ดี (GOOD) หรือไม่ดี (BAD)



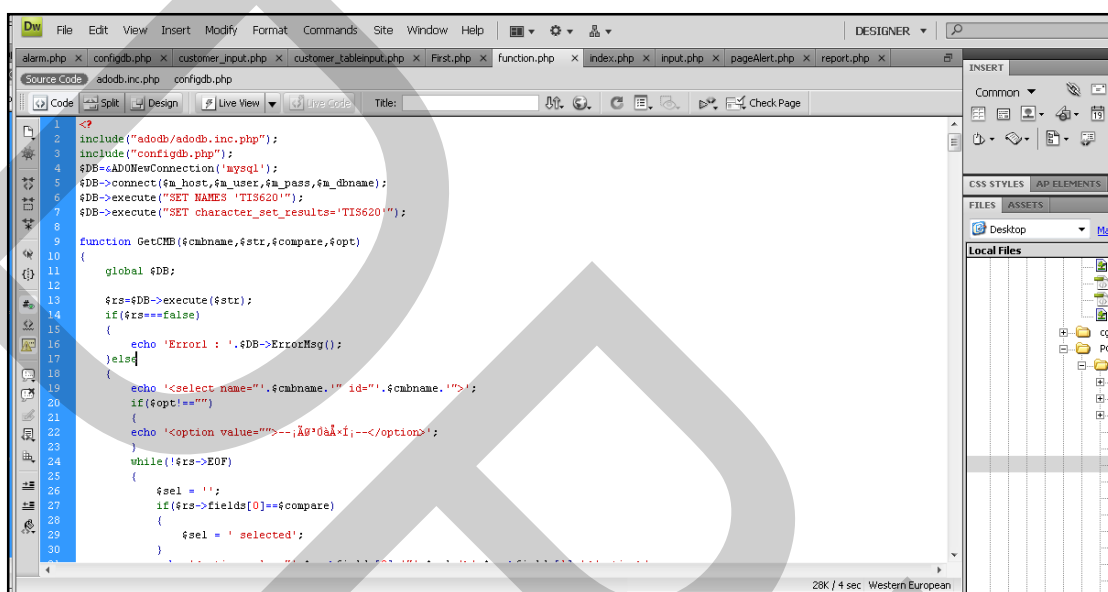
ภาพที่ 4.29 หน้าจอแสดงผลการวิเคราะห์ข้อมูลลูกค้าที่ดี



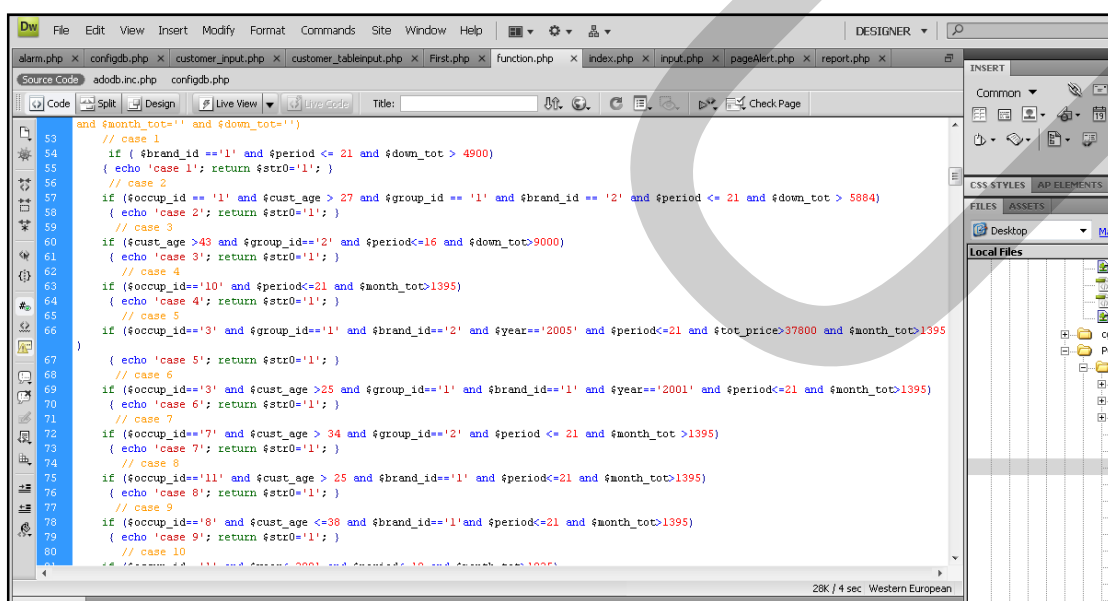
ภาพที่ 4.30 หน้าจอแสดงผลการวิเคราะห์ข้อมูลลูกค้าที่ไม่ดี

4.5.3 การเขียนโปรแกรม

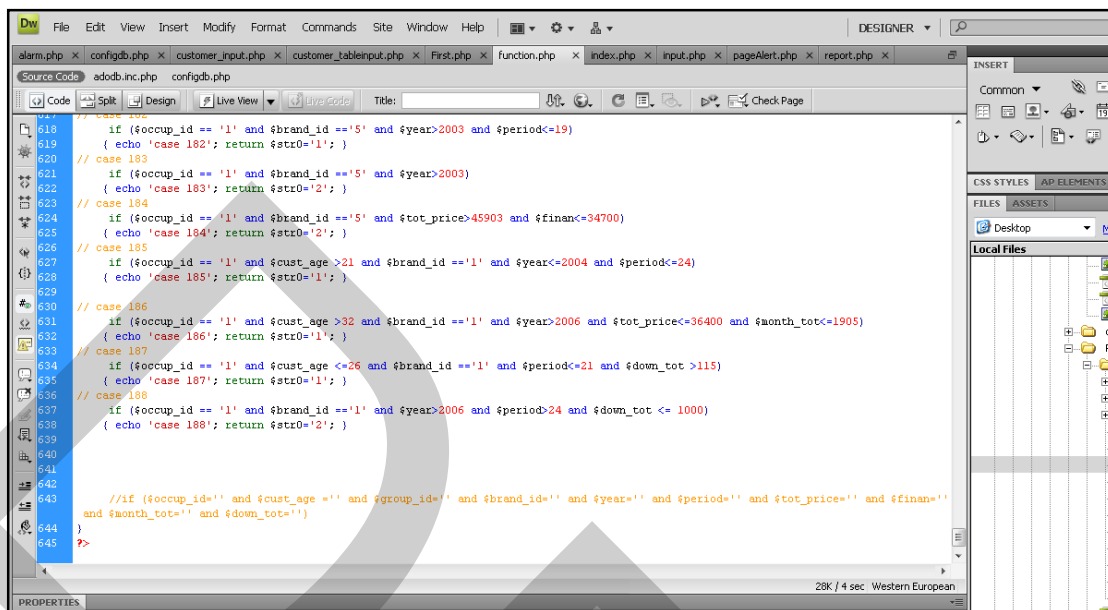
ใช้กฎที่ได้จากการทำเหมืองข้อมูลมาเป็นเงื่อนไขในการเขียนโปรแกรม จากผลการสรุปกฎที่ได้ทั้งหมด 188 กฎโดยจะทำการแปลงกฎที่ได้มาให้เป็น โค้ด ของตัวโปรแกรมใน 1 กฎที่ได้จะเท่ากับ 1 กรณี ใน โค้ด ของโปรแกรมจะได้ทั้งหมด 188 กรณี



ภาพที่ 4.31 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนแรก



ภาพที่ 4.32 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนที่เป็นเงื่อนไข



ภาพที่ 4.33 หน้าจอแสดงการนำกฎมาเขียนโปรแกรมส่วนสุดท้าย

4.5.3 การออกแบบรายงาน

เป็นการนำผลที่ได้จากการประมวลผลของโปรแกรมมาออกรายงานเพื่อให้ทราบข้อมูลต่างๆของลูกค้าที่มาขอสินเชื่อกับทางบริษัท จะมีทั้งข้อมูลของลูกค้าดีและข้อมูลลูกค้าที่ไม่ดี

ระบบวิเคราะห์สินเชื่อออนไลน์ | ONLINE ANALYSIS FUND

หน้าหลัก

วิเคราะห์สินเชื่อ

แก้ไขวิเคราะห์สินเชื่อ

รายงานการวิเคราะห์สินเชื่อ

ติดต่อเรา

SiteMap

รายงานผลการวิเคราะห์ข้อมูลลูกค้า

เลขที่บัตรประชาชน	ชื่อลูกค้า	อาชีพ	อายุ	กลุ่มสินค้า	มีบัตรสินค้า	มีสินเชื่อของสินค้า	จำนวนงวด	มูลค่าสินค้า	ยอดหนี้	ค้างงวด	เงินคงรวม	ผลการวิเคราะห์
0		13	42	1	1	2007	30	43500	43500	2320	0	
0		1	26	1	1	2007	12	34000	34000	3620	0	
0		1	37	1	1	2007	12	50200	49000	5215	1200	
2147483647	ลออจามประเสริฐ	2	35	2	1	2009	24	42500	40000	2570	2500	
0		1	34	1	2	2009	24	34000	30000	2350	4000	
2147483647	นิพัทธ์พรอ์เอสุจินต์	1	33	1	1	2009	24	55000	50000	3200	5000	
2147483647	ศิริกรนามศรี	1	24	2	1	2007	24	25000	22100	1500	2900	
0		0	0	0	0		0	0	0	0	0	
2147483647	กรรณิการามณเฑียร	3	52	1	2	2007	24	43000	32800	1990	10200	
0		0	0	0	0		0	0	0	0	0	
2147483647	สสสม	1	0	0	0		21	0	0	0	4900	
2147483647	สสสม	1	0	0	0		21	0	0	0	4900	
2147483647	สสสม	1	0	0	0		21	0	0	0	4900	
2147483647	ธนาพลสุขดี	3	52	1	2	2007	24	43000	32800	1990	10200	
2147483647	สมิทธธนะ	13	24	1	1	2007	24	43500	43500	2680	0	
2147483647	อนงค์สาธิต	2	23	1	1	2007	30	43500	43500	2320	0	
2147483647	แสงวาทดี	14	53	1	1	2007	30	44000	44000	2345	0	
0		1	24	1	2	2007	24	42000	42000	2590	0	
0		1	28	1	2	2007	24	42000	42000	2720	0	
0		1	23	1	2	2007	12	48000	32100	3345	15900	
0		1	0	1	1	2007	24	32550	32000	1910	550	
0		1	24	1	1	2007	24	42000	41500	2480	500	
0		1	23	1	1	2007	30	43500	43500	2320	0	
0		2	54	1	1	2007	30	43500	43500	2320	0	
0		2	25	1	2	2007	24	39700	33000	2005	6700	
0		7	30	1	2	2007	24	39700	33000	2005	6700	

ภาพที่ 4.34 หน้าจอแสดงรายงานข้อมูลลูกค้า

บทที่ 5

สรุปผลการศึกษา และข้อเสนอแนะการวิจัย

5.1 สรุปผลการวิจัย

จากการดำเนินการวิจัยผู้วิจัยได้ดำเนินการวิจัยจนสำเร็จตามวัตถุประสงค์การวิจัยคือได้ศึกษาโครงสร้างข้อมูลที่จะนำมาวิจัย การโอนย้ายข้อมูล (Data transfer), การทำความสะอาดข้อมูล (Data cleaning), การลดขนาดของข้อมูล (Data Reduction) การเลือกโมเดลที่เหมาะสมในการแบ่งกลุ่มของข้อมูลเพื่อเป็นแนวทางในการพิจารณาอนุมิตินเชื่อโดยเสนอขั้นตอนวิธีการทำเหมืองข้อมูล (Data Mining) เพื่อแบ่งกลุ่มลูกค้า (Classify) และนำตัวแบบที่ได้ไปประยุกต์ใช้เป็นแนวทางในการพิจารณาอนุมิตินเชื่อการเข้าซื้อรถจักรยานยนต์ ซึ่งการทำเหมืองข้อมูลครั้งนี้มีขั้นตอนหลักๆ 5 ขั้นตอน ดังนี้

ส่วนที่ 1 การศึกษาและรวบรวมข้อมูล การศึกษาโครงสร้างข้อมูล เนื่องจากการจัดเก็บข้อมูลการเข้าซื้อของบริษัท เอกนครหลวงลิสซิ่ง จำกัด อยู่ในรูปแบบแฟ้มข้อมูล ในระบบปฏิบัติการ DOS 7 โดยประกอบด้วยแฟ้มสัญญาการเข้าซื้อ, แฟ้มข้อมูลลูกค้า, แฟ้มข้อมูลรายละเอียดของรุ่นรถ และแฟ้มข้อมูลการชำระค่างวด เป็นต้น จึงได้มีการศึกษาโครงสร้างข้อมูลและเลือกเฉพาะข้อมูลของลูกค้าที่ปิดบัญชีไปแล้ว ตั้งแต่ปี พ.ศ. 2545-2550 ซึ่งมีจำนวน 17,964 สัญญา และศึกษาถึงข้อมูลที่ใช้ในการพิจารณาอนุมิตินเชื่อการเข้าซื้อรถจักรยานยนต์ โดยจากข้อมูลที่มีการจัดเก็บทั้งหมดจำนวน 142 ไฟล์ คัดเลือกเฉพาะข้อมูลที่สำคัญที่ใช้ในการพิจารณาจะประกอบไปด้วยข้อมูล ดังนี้ อาชีพ อายุ รหัสกลุ่ม ยี่ห้อสินค้า ปีที่ผลิตสินค้า, จำนวนเงินที่ขอสินเชื่อ ค่างวดต่อเดือน จำนวนงวด เงินคาวน และสถานภาพลูกค้า

ส่วนที่ 2 การเตรียมข้อมูล เนื่องจากข้อมูลที่มีอยู่เดิมมีการจัดเก็บในรูปแบบของ ไฟล์ DBF ในระบบปฏิบัติการ DOS 7 ไม่สามารถที่จะนำมาใช้กับโปรแกรม Weka3.5.7 ได้จึงต้องทำการโอนย้ายข้อมูลให้มาอยู่ในรูปแบบ ไฟล์ Excel นามสกุล .csv หลังจากที่ได้ทำการโอนย้ายข้อมูลมาแล้วข้อมูลที่ไดยังไม่เหมาะสมต้องมีการทำความสะอาดข้อมูล (Data cleaning) และทำการลดขนาดของข้อมูล (Data Reduction) เพื่อให้ได้ข้อมูลที่สมบูรณ์จริงๆ เพียง 17,964 สัญญา ที่ใช้ในการทดสอบการจำแนกกลุ่ม

ส่วนที่ 3 การทดสอบและเลือกเทคนิคที่เหมาะสม ในการศึกษาครั้งนี้ได้ทดสอบเทคนิคในการจำแนกกลุ่ม 3 เทคนิค ที่ให้ผลลัพธ์เป็นกฎ คือ โมเดลการจำแนกกลุ่มแบบ Decision Tree:

C4.5 หรือเรียกว่าโมเดล Classifiers trees J48 โมเดลการจำแนกกลุ่มแบบ Decision Rules: Part หรือเรียกว่าโมเดล Classifiers PART และโมเดลการจำแนกกลุ่มแบบ Decision Table หรือเรียกว่าโมเดล Classifiers Decision Table ในโปรแกรม Weka3.5.7 จากการทดสอบการจำแนกกลุ่มทั้ง 3 เทคนิค ผลการทดสอบพบว่าการจัดกลุ่มแบบ Decision Rules: Part ให้ผลลัพธ์การจำแนกกลุ่มที่ดีที่สุด โดยพิจารณาความถูกต้องและจำนวนกฎที่ได้ คือ 72.7 เปอร์เซ็นต์ และได้กฎจำนวน 346 กฎ ซึ่งกฎที่ได้มีความเหมาะสม สอดคล้องกับการพิจารณาอนุมัตินเชื่อมากที่สุด

ส่วนที่ 4 การวิเคราะห์ข้อมูล จากผลของการทดสอบการจำแนกกลุ่มข้อมูลแบบ Decision Rules: Part เนื่องจากมีกฎจำนวนมาก และกฎบางกฎมีเงื่อนไขไม่เหมาะสม บางกฎมีเงื่อนไขที่ซ้ำซ้อน ดังนั้นผู้เชี่ยวชาญจะทำหน้าที่ตรวจสอบ แก้ไขและลบกฎบางกฎทิ้งไป โดยตัดกฎที่ไม่เหมาะสมออก และปรับกฎที่มีเงื่อนไขที่ซ้ำซ้อนให้เข้าใจง่ายขึ้น จากทั้งหมด 346 กฎ เหลือเพียง 188 กฎ

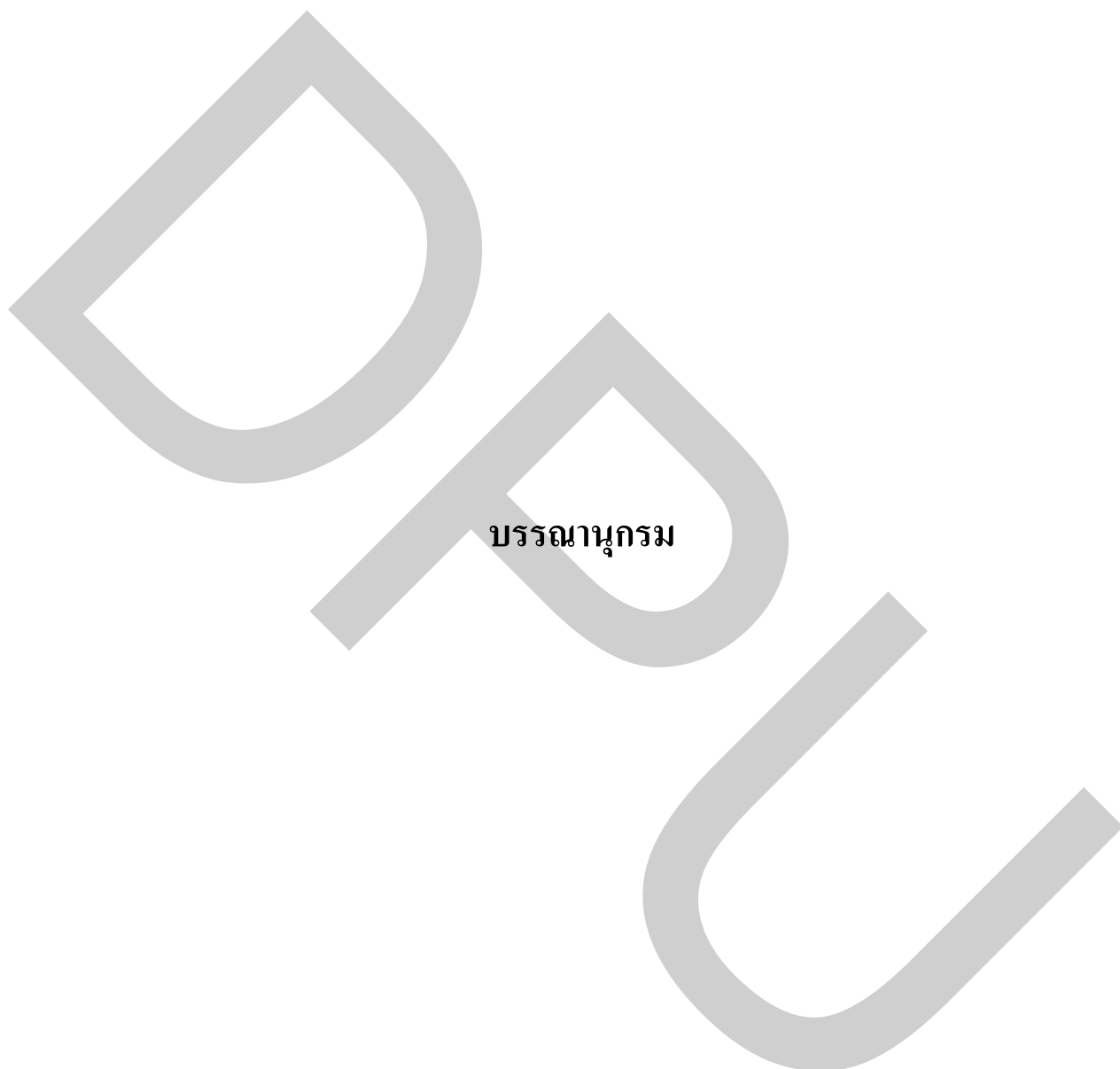
ส่วนที่ 5 การพัฒนาระบบสนับสนุนการตัดสินใจการอนุมัตินเชื่อ เป็นการนำกฎที่มีความน่าเชื่อถือและผ่านการพิจารณา ไปพัฒนาเป็นเป็นเงื่อนไขในการเขียนโปรแกรมเพื่อเป็นเครื่องมือช่วยในการพิจารณาอนุมัตินเชื่อเข้าชื่อรถจักรยานยนต์ โดยโปรแกรมที่ได้เป็นเครื่องมือที่จะสนับสนุนการพิจารณาการอนุมัตินเชื่อซึ่งพิจารณาจากข้อมูลของลูกค้าเท่านั้น แต่ในการพิจารณาสินเชื่อยังต้องมีการพิจารณาข้อมูลในส่วนอื่นๆอีก เช่น สภาวะตลาด การเมือง คู่แข่งทางธุรกิจ

5.2 ปัญหาและแนวทางแก้ไข

1. ในการทำเหมืองข้อมูลครั้งนี้ขั้นตอนการเตรียมข้อมูลยังขาดข้อมูลบางส่วนที่จำเป็นในการพิจารณาอนุมัตินเชื่อเช่น อายุของการทำงานลูกค้า รายได้รวมของลูกค้า หากมีการนำข้อมูลเหล่านี้ในการวิจัยก็จะทำให้ผลการวิจัยที่ได้มีความน่าเชื่อถือมากขึ้น
2. ในการใช้เครื่องมือ (Tool) ต่างๆ เพื่อทำการวิจัยครั้งนี้ ใช้ซอฟต์แวร์แบบโอเพนซอส ซึ่งมีเอกสารประกอบน้อยจึงทำให้ลำบากในการศึกษาค้นคว้าหาวิธีการใช้งาน การวิเคราะห์ และแปลผลลัพธ์
3. ข้อมูลที่บริษัทมีอยู่มีจำนวนมากการโอนย้ายข้อมูลต้องทดลองหลายวิธีเพื่อให้ได้ข้อมูลที่โอนย้ายมามีความสมบูรณ์มากที่สุดจึงทำให้ข้อมูลน่าเชื่อถือมาก
4. ในการทำเหมืองข้อมูลครั้งนี้มีการพิจารณาข้อมูลป้อน ซึ่งหากระยะเวลาผ่านไปนานๆ จะทำให้กฎบางกฎที่เป็นเงื่อนไขของโปรแกรมไม่สามารถใช้งานได้ดีเท่าควรจึงต้องมีการปรับปรุงแก้ไขต่อไป

5.3 ข้อเสนอแนะ

1. รูปแบบเครื่องมือสนับสนุนการพิจารณาอนุมัติสินเชื่อที่ได้จากการทำเหมืองข้อมูลสามารถนำไปเป็นส่วนหนึ่งของการพิจารณาอนุมัติสินเชื่อได้จริง แต่ในการอนุมัติสินเชื่อจะต้องประกอบไปด้วยปัจจัยอื่นอีกหลายๆ ประการ เช่น การเมือง เศรษฐกิจ สภาพตลาดในขณะนั้น
2. ในการทำเหมืองข้อมูล สิ่งที่สำคัญที่สุดในการวิเคราะห์ข้อมูลคือผู้วิจัยจะต้องทราบลักษณะโครงสร้างข้อมูลข้อมูล เพื่อความถูกต้องในการเลือกใช้เทคนิคการทำเหมืองข้อมูลซึ่งจะส่งผลให้ได้ผลลัพธ์ที่มีประสิทธิภาพสูงสุด
3. การทำเหมืองข้อมูลควรใช้ข้อมูลในการทดสอบจำนวนมาก เพื่อให้ได้ผลการวิจัยที่มีความน่าเชื่อถือมากยิ่งขึ้น



ปฐ

บรรณานุกรม

ภาษาไทย

หนังสือ

กิตติพงศ์ กลมกล่อม. (2546). การออกแบบและพัฒนาคลังข้อมูล (Data Warehouse) (พิมพ์ครั้งที่2).
กรุงเทพฯ : เคทีพี คอมพ์ แอนด์ คอนซัลท์ .

ชาญชัย ศุภอรรถกร. (2552). คู่มือการเขียนเว็บอีคอมเมิร์ซด้วย PHP + MySQL. กรุงเทพฯ : ชัค
เชส มีเดีย.

พงษ์พันธ์ ศิวิลัย. (2552). SQL Server 2008. กรุงเทพฯ : ซีเอ็ดดูเคชั่น.

วงศ์ประชา จันทร์สมวงศ์. (2547). อินไซต์ Dreamweaver MX 2004. กรุงเทพฯ : โปรวิชั่น.

วิทยานิพนธ์

ณัฐริน เจริญเกียรติบวร. (2549). การใช้เทคนิคเหมืองข้อมูลเพื่อช่วยในการแนะแนวการศึกษาต่อ
ระดับอุดมศึกษา. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์.
กรุงเทพฯ: มหาวิทยาลัยศิลปากร.

บุญมา เฟ่งชวน. (2548). การใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสนับสนุนการตัดสินใจด้านการ
ผลิต. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์. กรุงเทพฯ:
มหาวิทยาลัยศิลปากร.

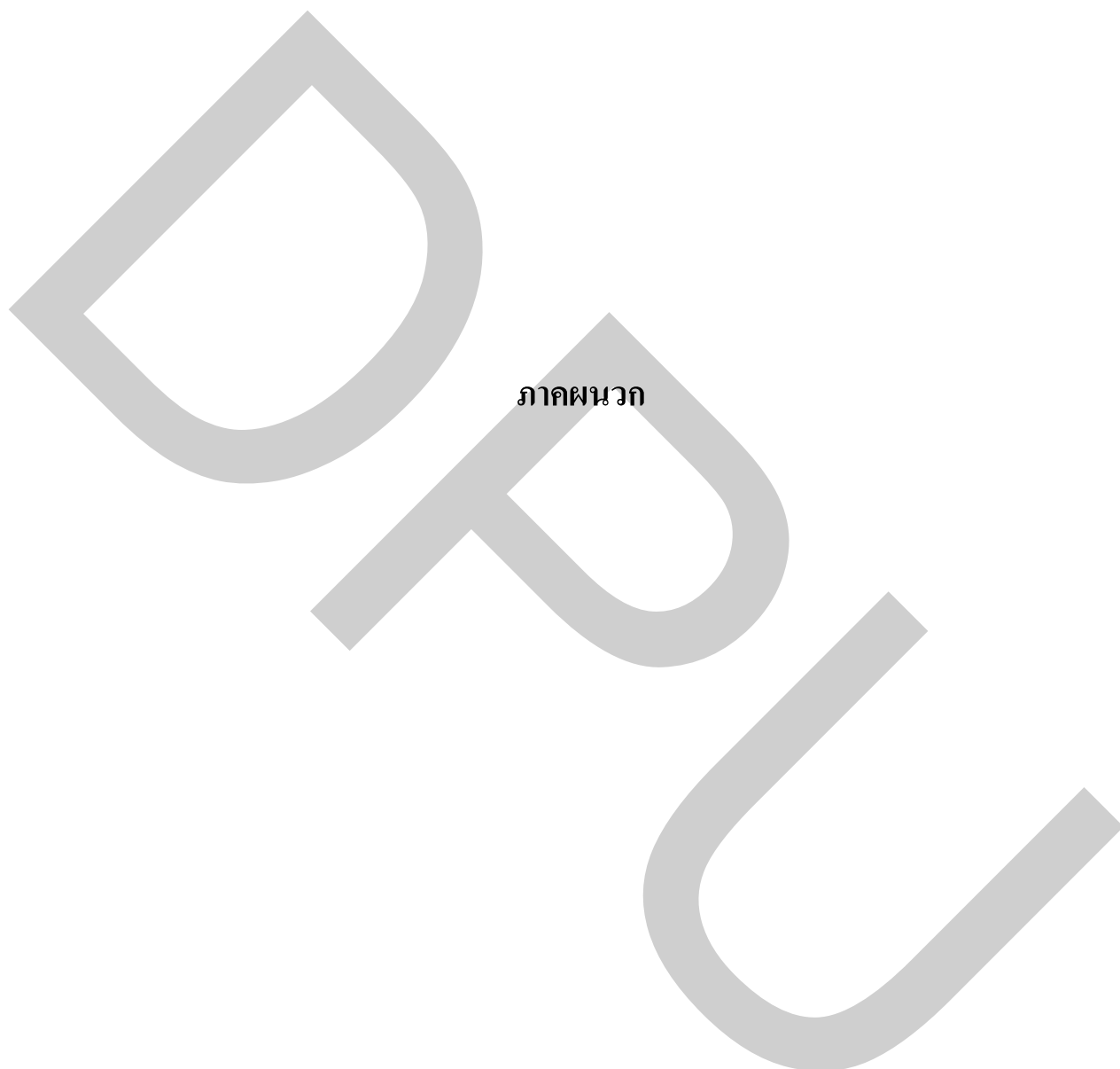
- บวร น้อยแสง.(2549). คลังข้อมูลและเทคนิคเหมืองข้อมูลสำหรับการวิเคราะห์ กรณีศึกษาของ บริษัทเฟ้นซืออาร์ที. วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ สาขาวิชาคณิตศาสตร์และเทคโนโลยีสารสนเทศ. กรุงเทพฯ : มหาวิทยาลัยศิลปากร.
- บุญมา เฟ่งชวน.(2548). การศึกษาตัวแบบและพัฒนาตัวแบบ เพื่อนำตัวแบบที่ได้นำไปประยุกต์ใช้ในการสนับสนุนการตัดสินใจด้านการผลิตบัณฑิต. วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ สาขาวิชาวิทยาการคอมพิวเตอร์. กรุงเทพฯ: มหาวิทยาลัยศิลปากร.
- สุวนีย์ กุลกรนิษฐม. (2549). พัฒนาตัวแบบสำหรับการจัดกลุ่มหลักสูตร ตามกลุ่มสาขาวิชา ISCED . วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ สาขาวิชาวิทยาศาสตร์. กรุงเทพฯ : มหาวิทยาลัยศิลปากร

สารสนเทศจากสื่ออิเล็กทรอนิกส์

- กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวิทย์ รักธรรมานนท์. (2001). การใช้เทคนิคดาต้าไมน์นิ่ง เพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์ : The NECTEC Technical journal. สืบค้นเมื่อ 20 กรกฎาคม 2552,
จาก <http://www.tdc.thislis.or.th>.
- กฤษณะ ไวยมัย และธีระวัฒน์ พงษ์ศิริปริดา. (2001). การใช้เทคนิค Association Rule Discovery เพื่อการจัดสรรกฎหมายในการพิจารณาคดีความ : The NECTEC Technical Journal. สืบค้นเมื่อ 20 กรกฎาคม 2552,
จาก <http://www.tdc.thislis.or.th>.

บุคคล

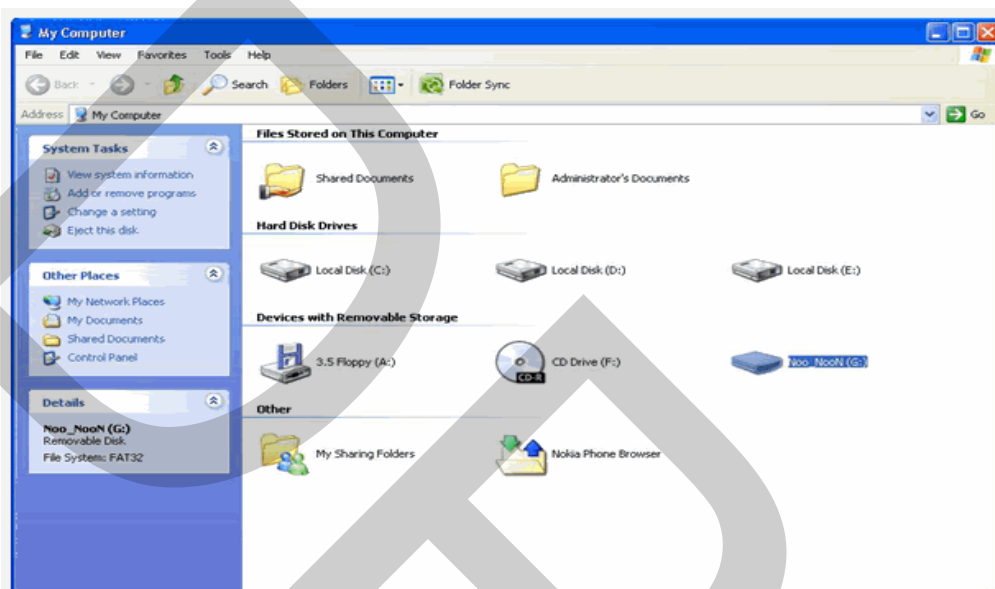
- ¹เกรียงศักดิ์ ชดช้อย. ผู้จัดการฝ่ายการตลาดและสินเชื่อ บริษัทเอกนครหลวงลิซซิ่ง จำกัด ประเทศไทย.สัมภาษณ์ 16 พฤศจิกายน 2552.



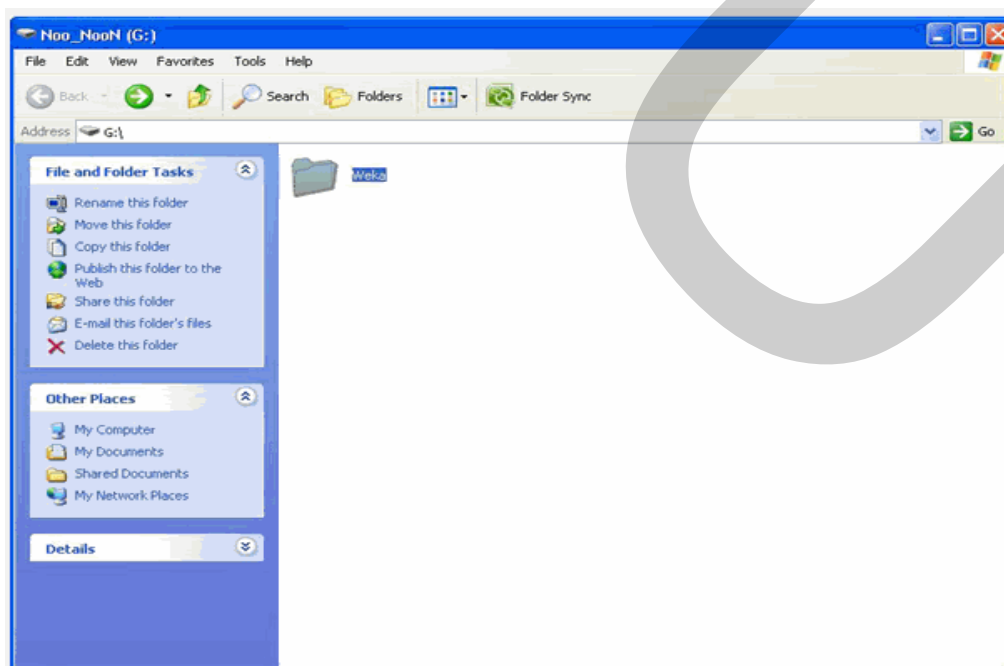
ภาคผนวก

ภาคผนวก

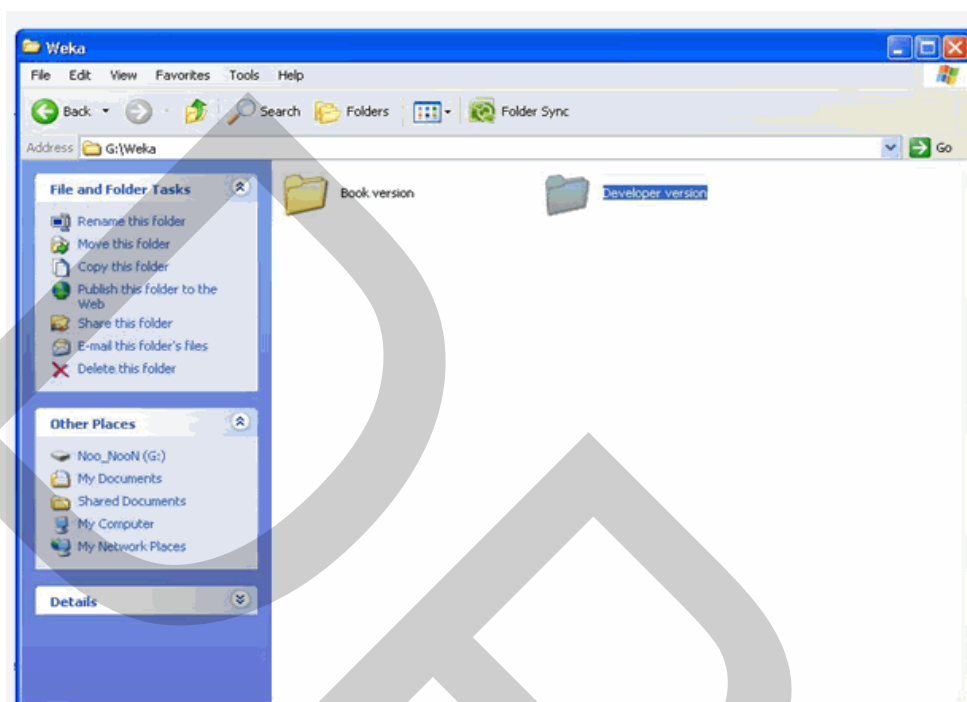
ขั้นตอนการลงโปรแกรม Weka



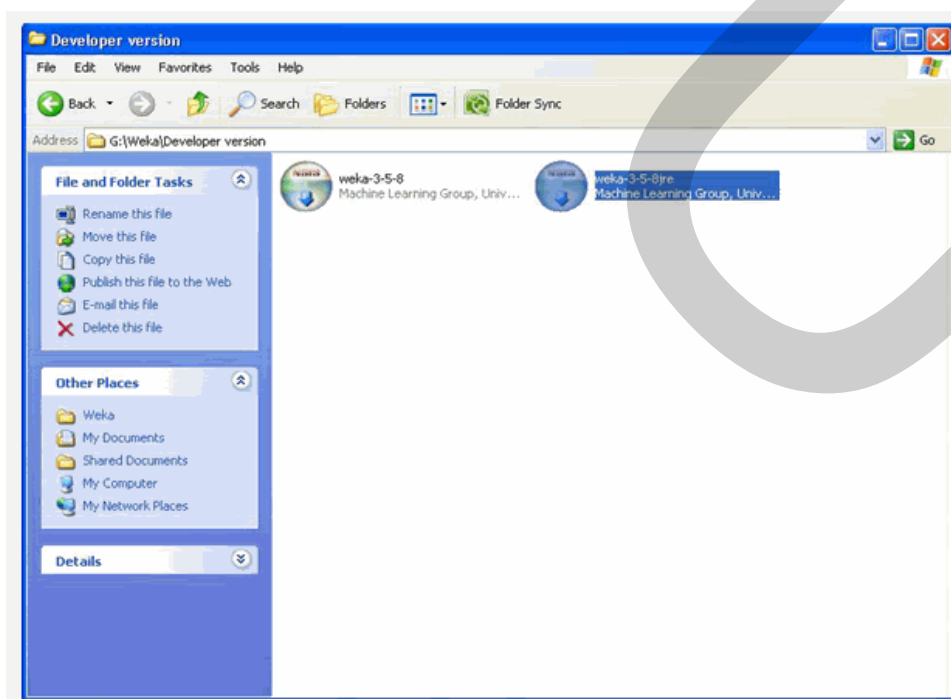
เปิดโปรแกรม Weka 3.5.8 ในกรณีนี้โปรแกรมถูกเก็บอยู่ที่ G: จึงเริ่มต้นการทำงานด้วยการเปิด My Computer จากนั้นคลิกที่ G:\



หลังจากเปิด G:\ แล้ว ให้คลิกที่ Folder ที่มีชื่อว่า Weka

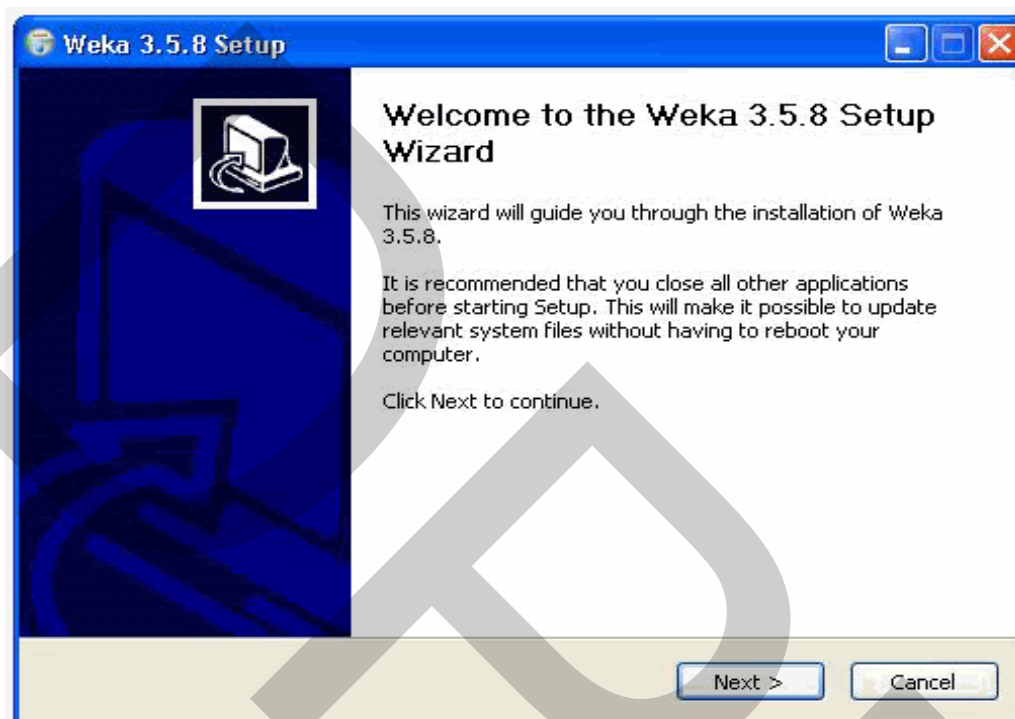


จากภาพจะเห็นได้ว่า ใน Folder ที่มีชื่อว่า Weka มี Folder อยู่ 2 Folder
ให้คลิกที่ Developer version

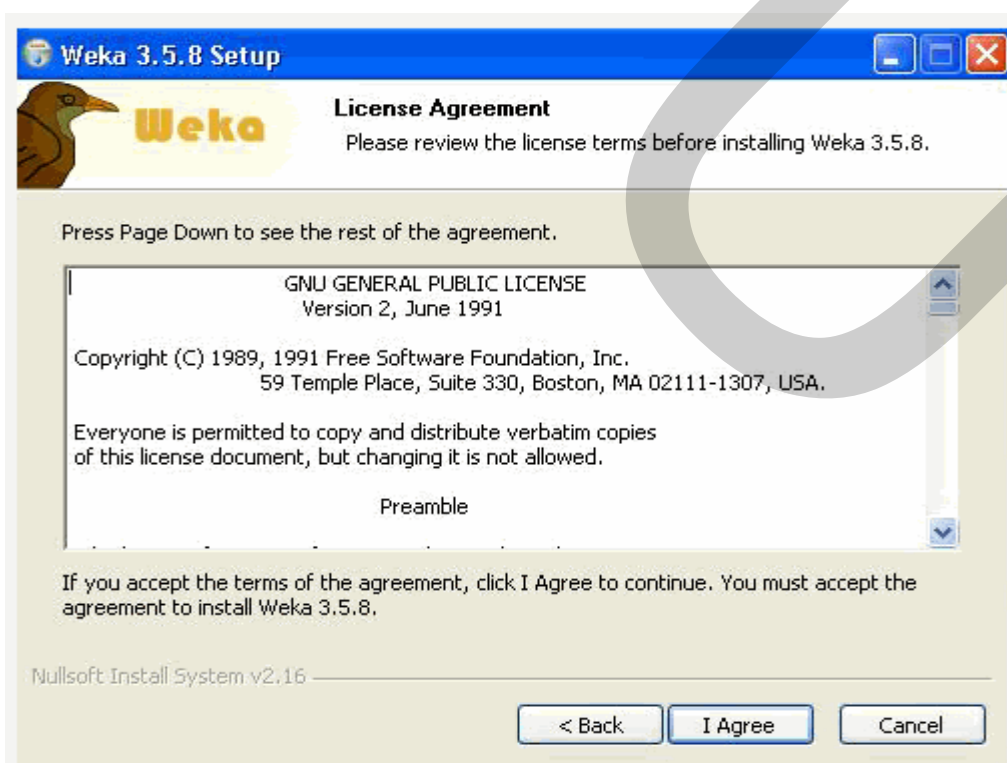


ภายใน Folder จะมีโปรแกรม Weka อยู่ 2 โปรแกรม ให้คลิกที่ weka-3-5-8jre

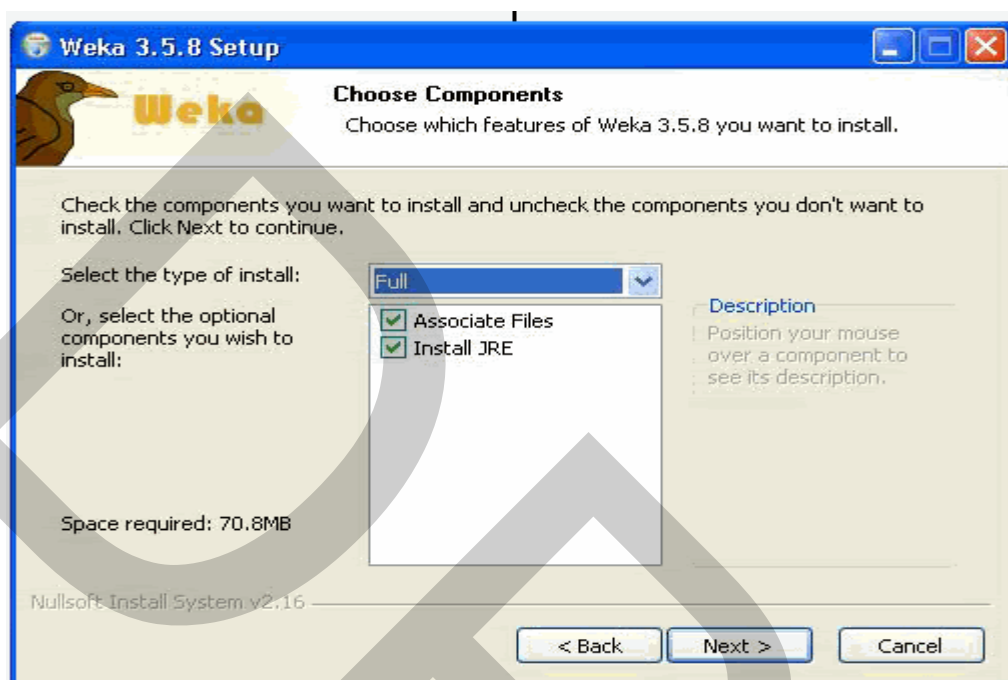
เพื่อทำการรันโปรแกรม



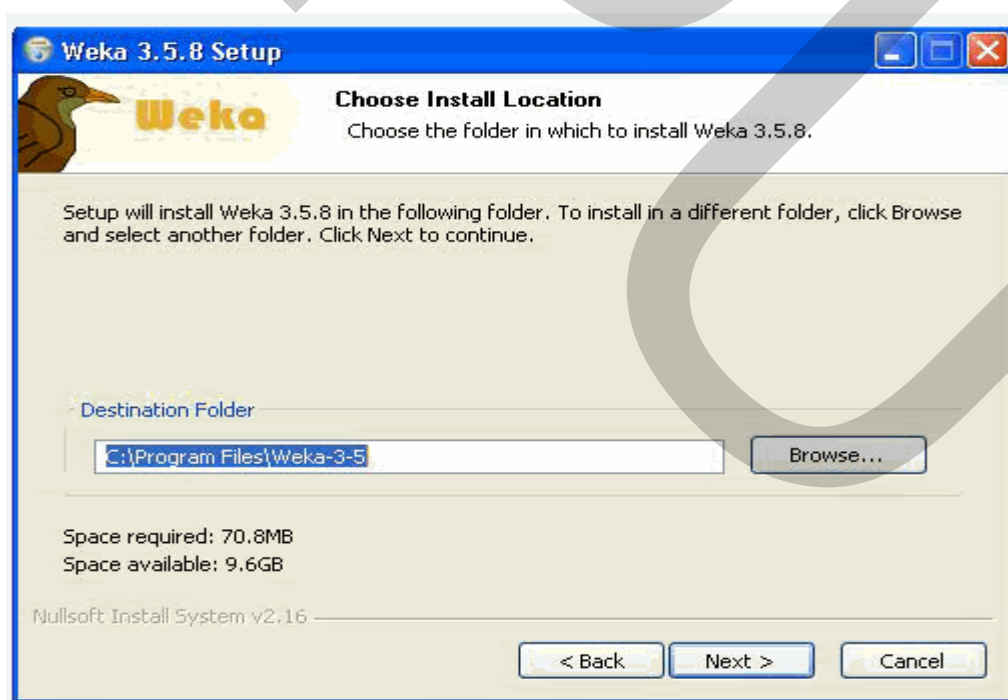
เมื่อรันหน้าจอโปรแกรม Weka 3.5.8 ขึ้นมา ให้คลิกที่ Next เพื่อประมวลผลต่อไป



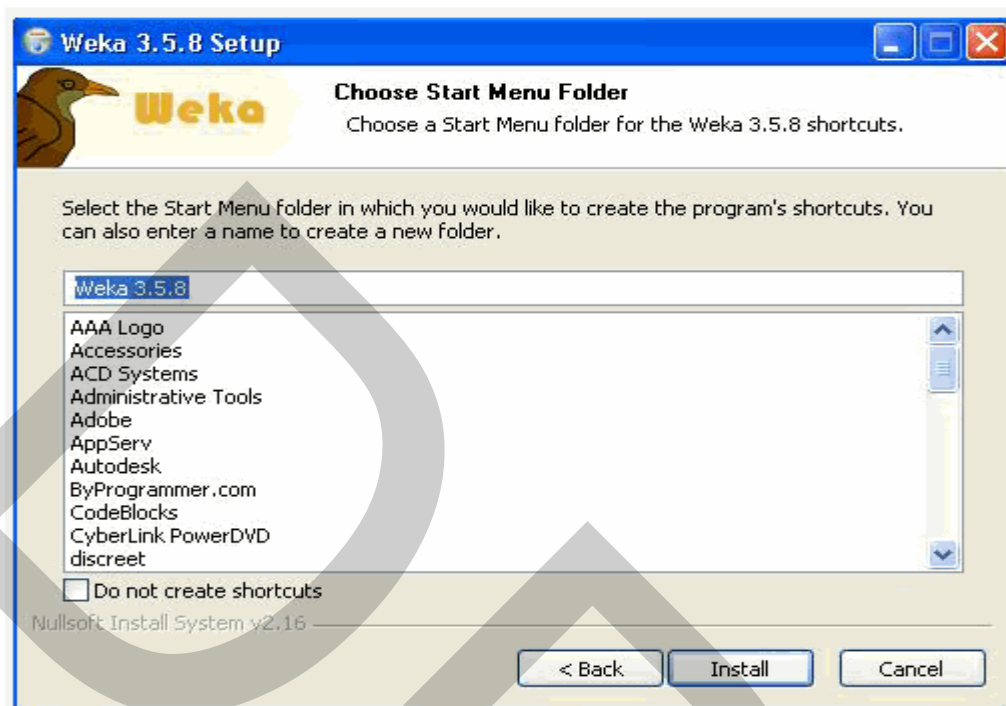
คลิกที่ I Agree เพื่อแสดงการยอมรับ จากนั้นโปรแกรมจะประมวลผลหน้าถัดไป



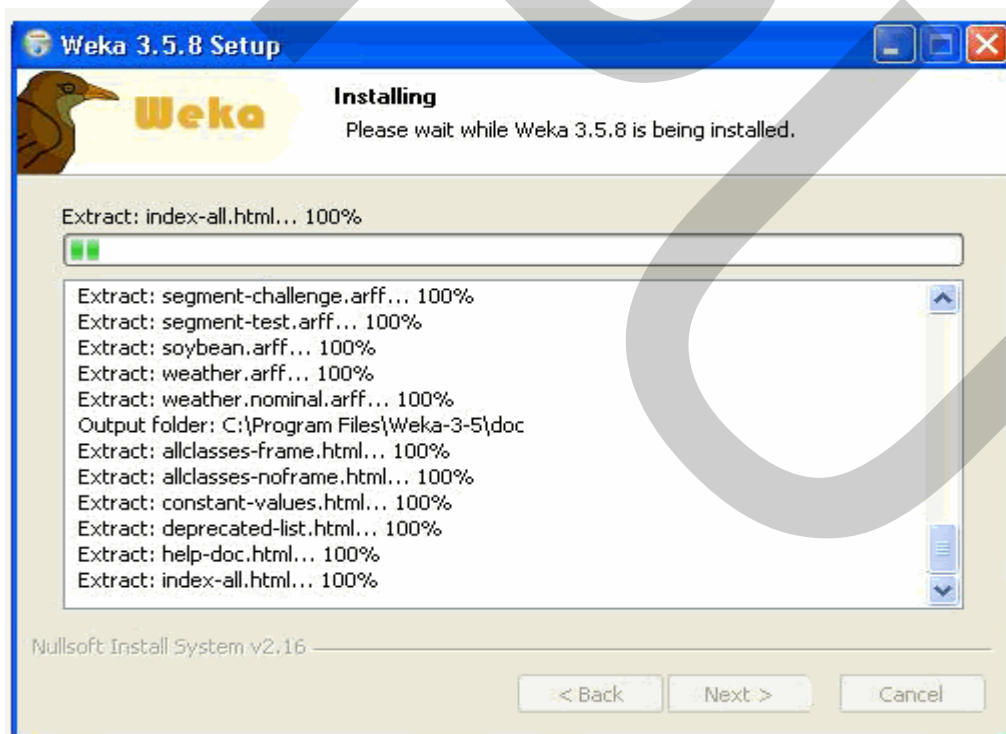
จากภาพนี้ คลิกที่ Next เพื่อทำการ Install



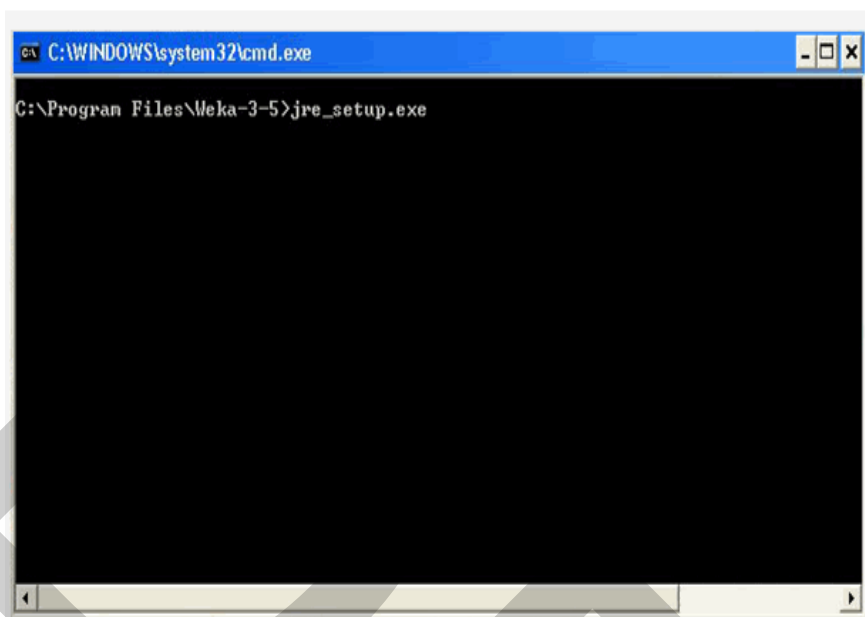
เลือกสถานที่ที่ต้องการเก็บโปรแกรม ในกรณีนี้จะเลือก C:\ จากนั้นคลิกที่ Next เพื่อดำเนินการต่อไป



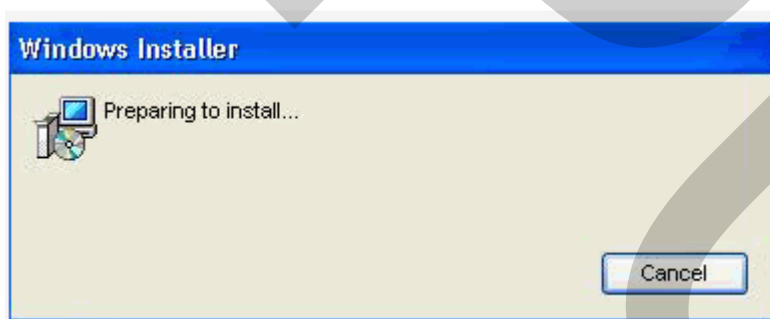
คลิก Install เพื่อเริ่มต้นการติดตั้งโปรแกรม



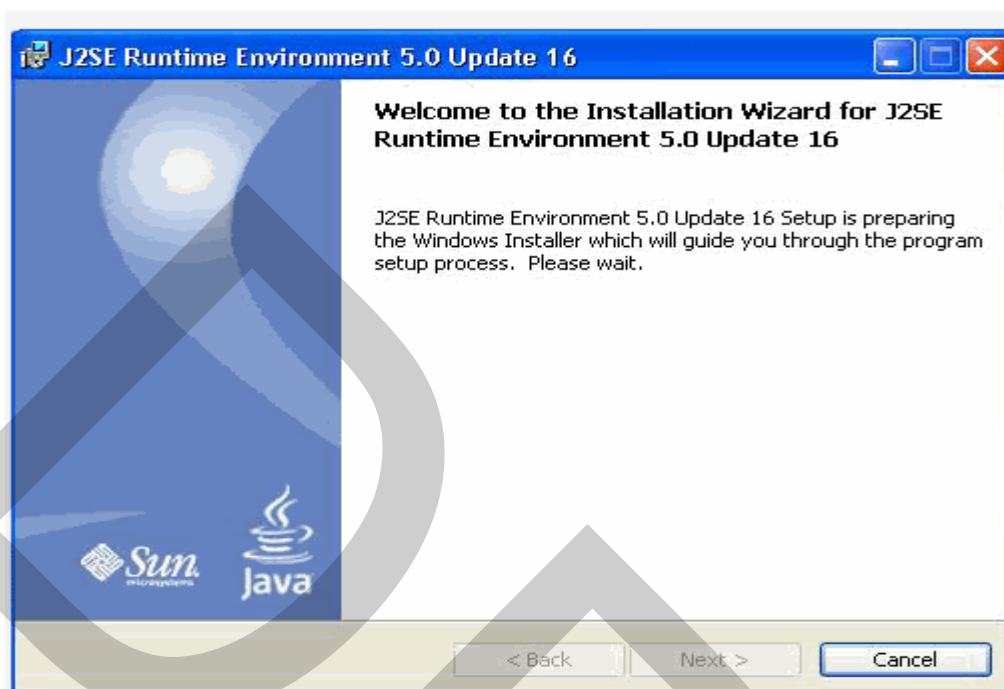
ภาพนี้จะแสดงการประมวลผลการติดตั้งโปรแกรม



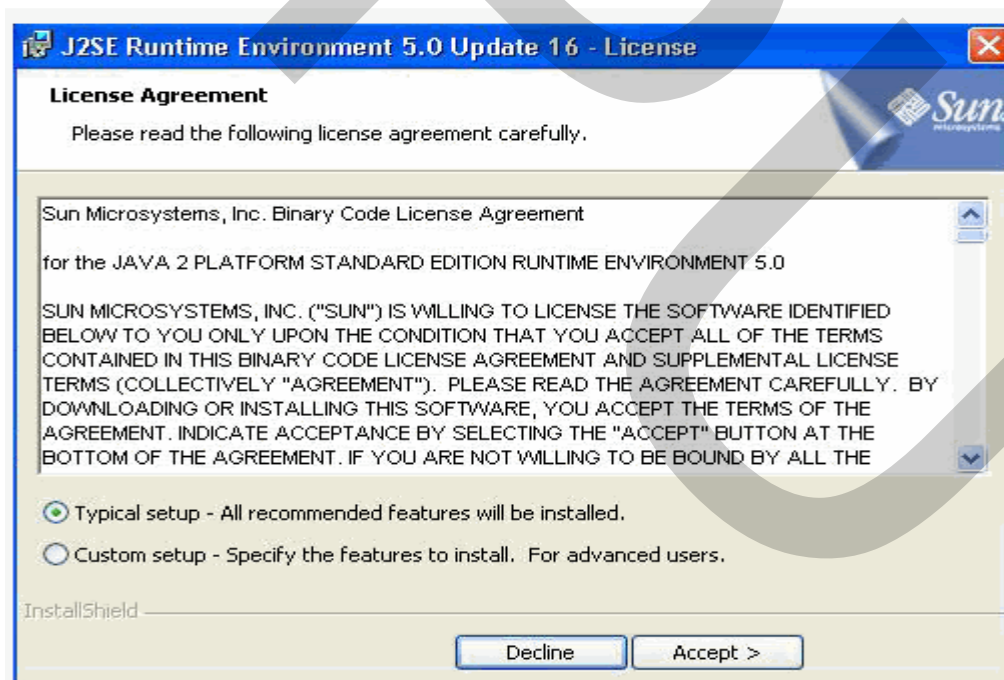
จากภาพที่แล้ว เมื่อประมวลผลเสร็จแล้ว จะแสดงหน้าจอนี้ขึ้นมา เราไม่ต้องคลิกปิด เพราะหน้าจอนี้จะถูกปิดโดยอัตโนมัติหลังจากการติดตั้งเสร็จเรียบร้อยแล้ว



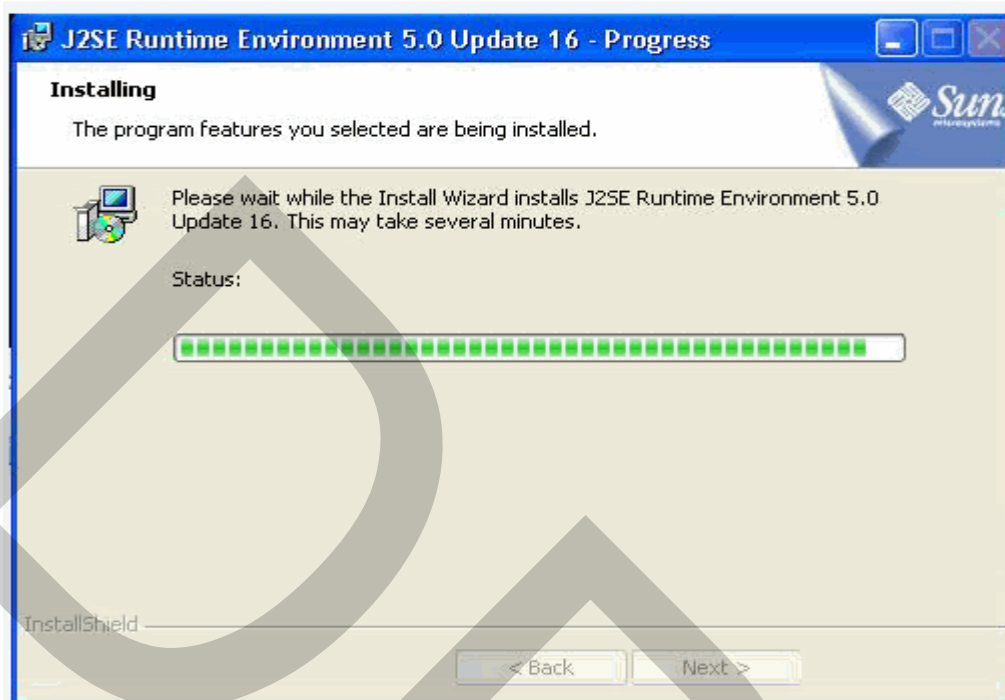
หน้าจอนี้จะแสดงขึ้นมาเพื่อบอกว่าโปรแกรมกำลังจะติดตั้ง
ถ้าหากต้องการยกเลิกการติดตั้งให้คลิกที่ Cancel



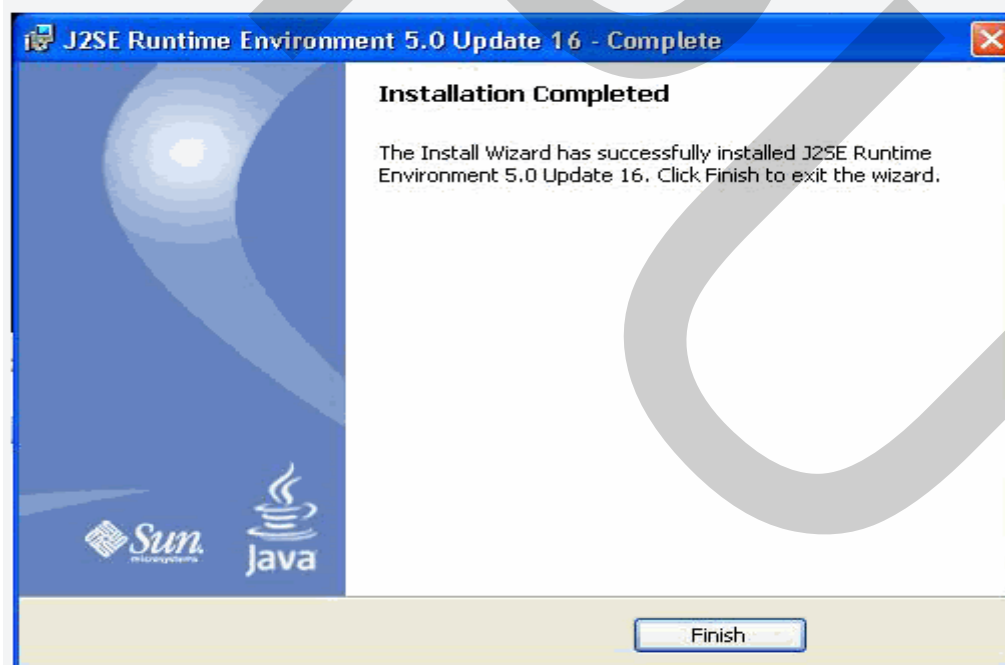
ถ้าต้องการยกเลิกการติดตั้งโปรแกรม คลิกที่ Cancel



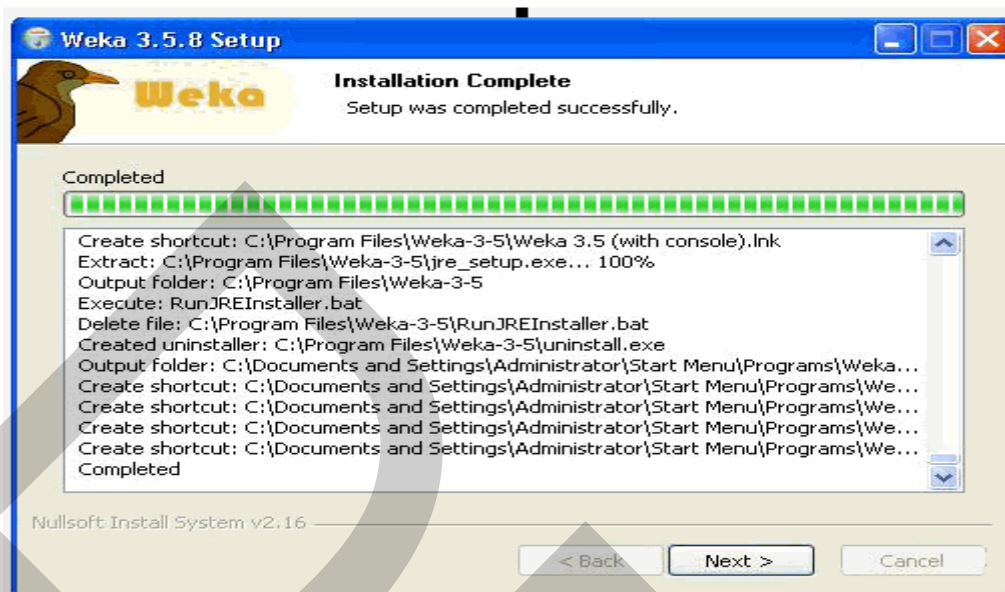
คลิกที่ Accept เพื่อยอมรับและดำเนินการต่อ



โปรแกรมทำการติดตั้งลงสู่ C:\



คลิกที่ Finish เพื่อเป็นการยืนยันการติดตั้งโปรแกรมเสร็จเรียบร้อยแล้ว



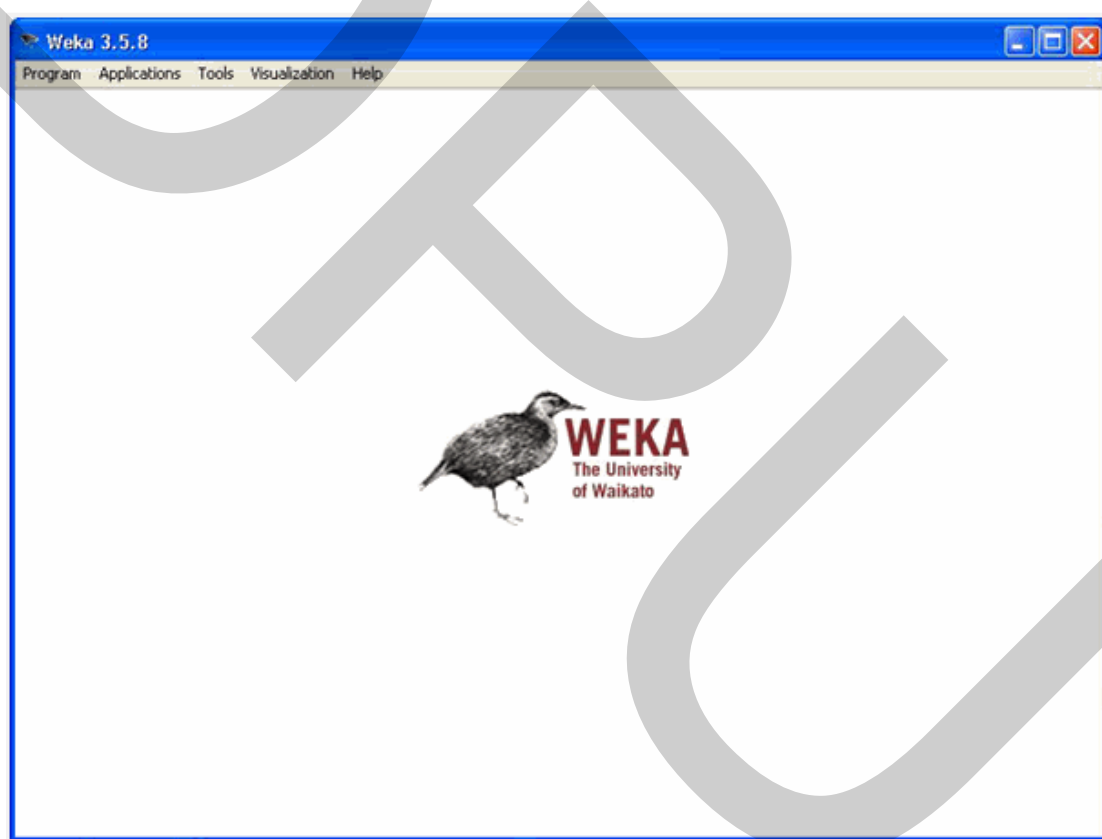
ภาพนี้แสดงให้เห็นว่าการดำเนินการเสร็จแล้ว คลิกที่ Next ต่อไป



แสดงให้ถึงการติดตั้งโปรแกรม Weka ที่เสร็จสิ้นสมบูรณ์แล้ว ตรง Start Weka ที่ถูกเช็คไว้ เพื่อให้รันหน้าจอโปรแกรม Weka 3.5.8 ถ้าไม่ต้องการให้รันโปรแกรมในขณะนี้ ให้คลิกเพื่อให้เครื่องหมายนั้นหายไป สุดท้ายให้คลิกที่ Finish



แสดงการเปิดโปรแกรม Weka 3.5.8



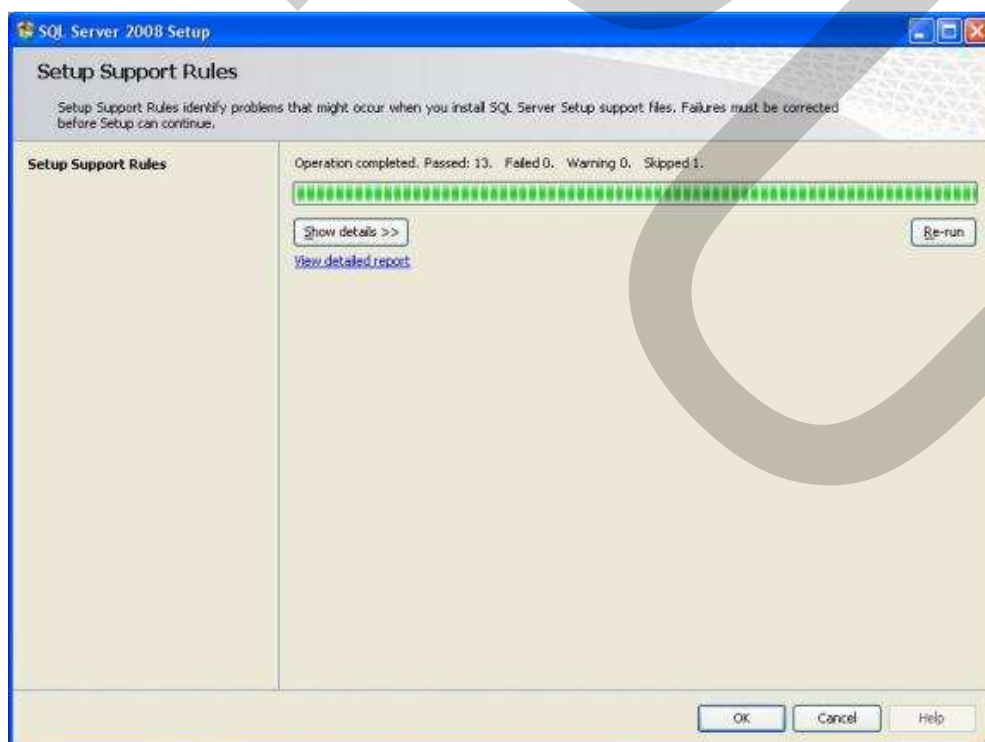
แสดงการหน้าจอการทำงานของโปรแกรม Weka3.5.8จบการลงโปรแกรมWeka

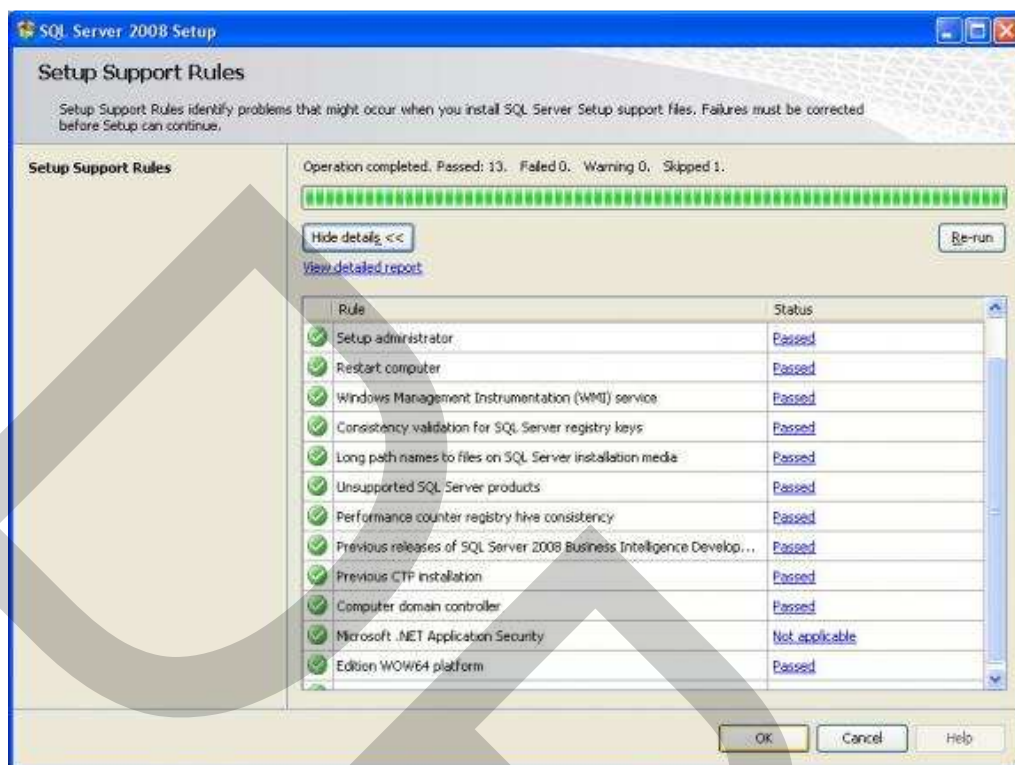
การติดตั้ง SQL Server 2008

รูปภาพ:เลือก System Configuration Checker เพื่อตรวจสอบ



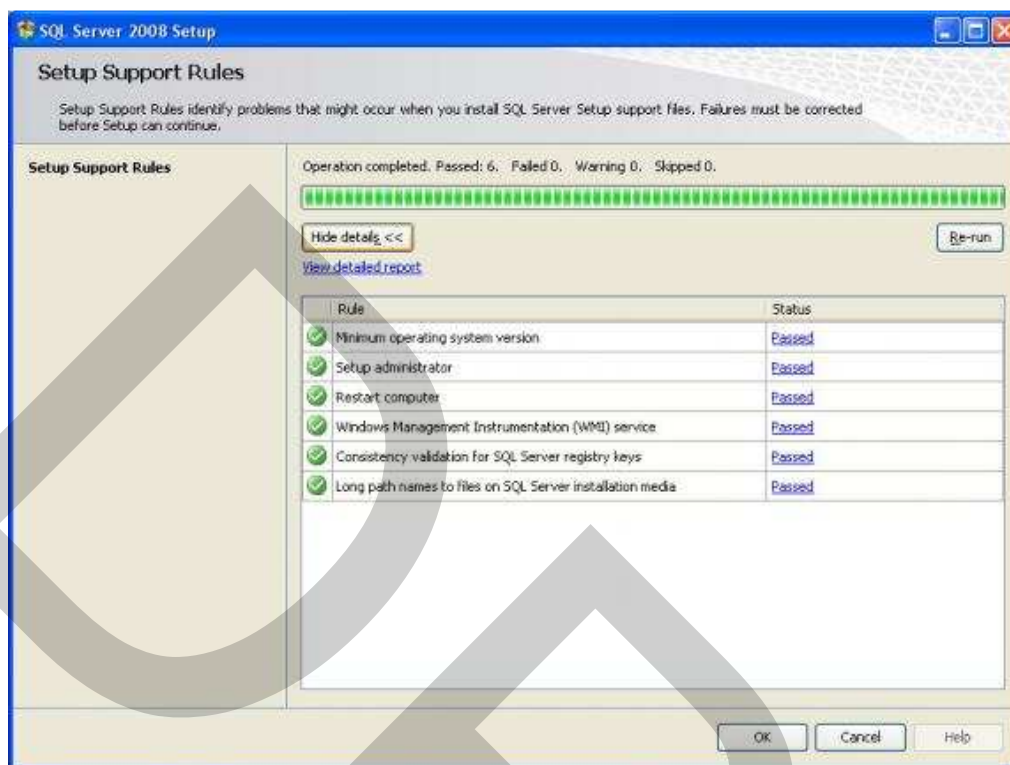
กด Show Details เพื่อดูว่าผ่านหมดหรือไม่



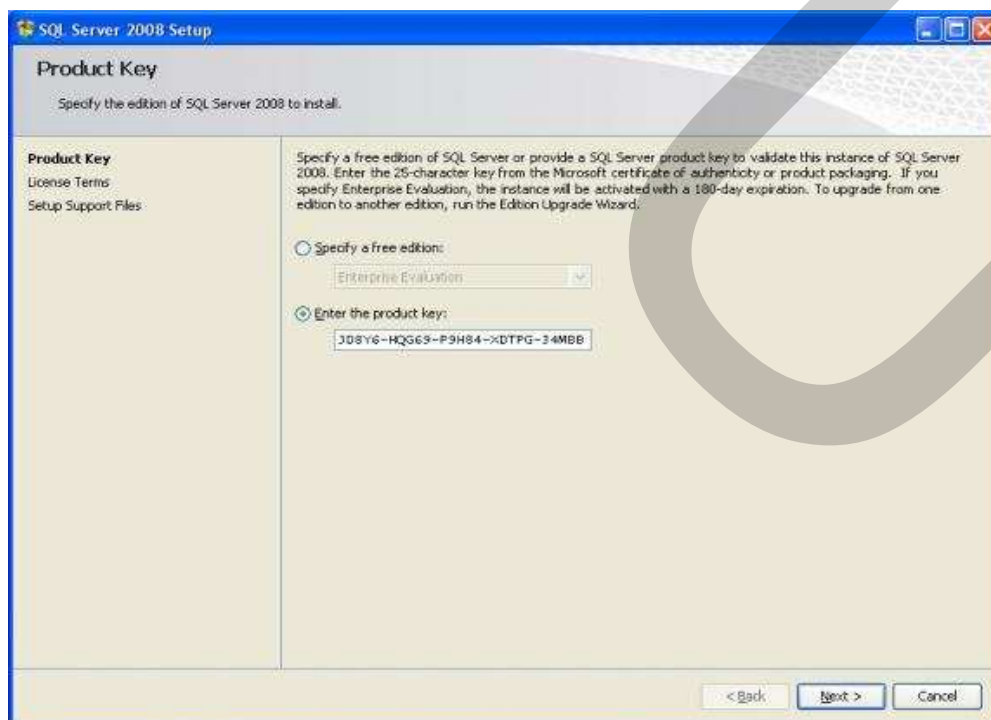


เลือก Installation แล้วเลือกลงแบบ New SQL Server Stand-alone...





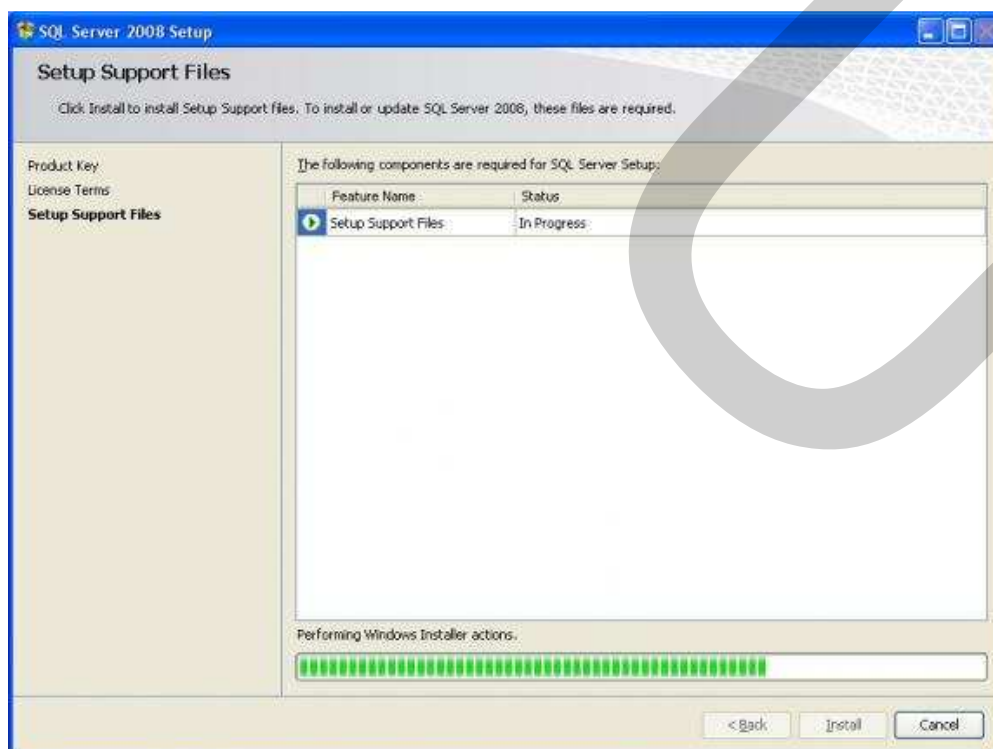
เลือก Enter the Product key เรียบร้อย กด Next แต่ส่วนใหญ่มันจะใส่ให้เลย

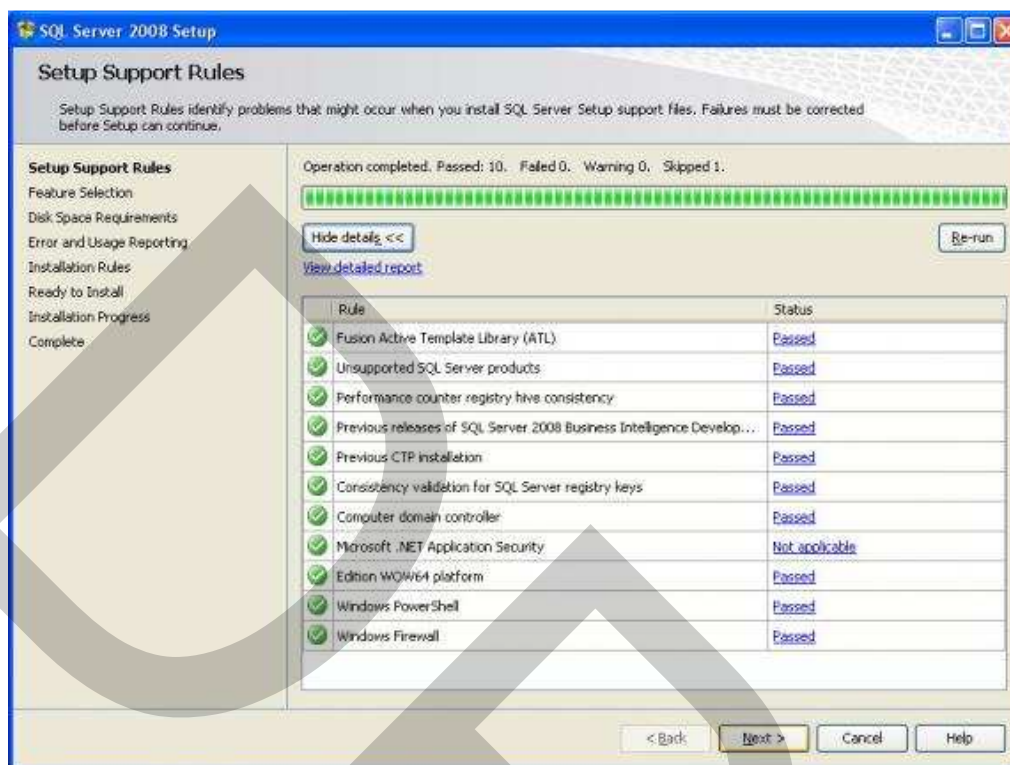


ใส่ product key

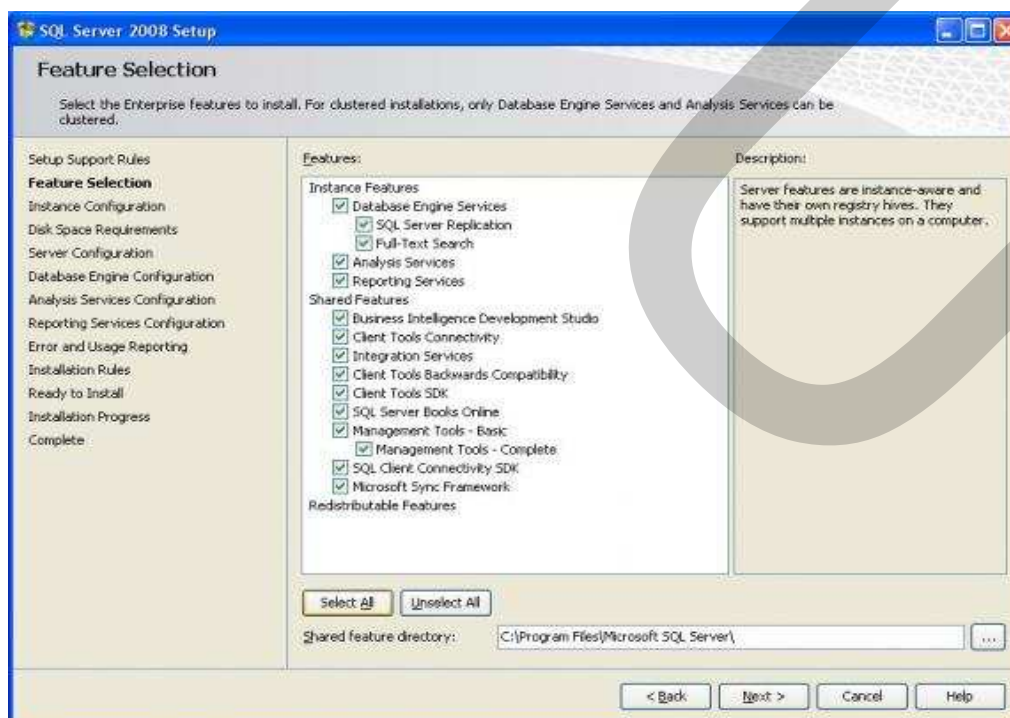


เลือก I accept.... แล้วกด Next

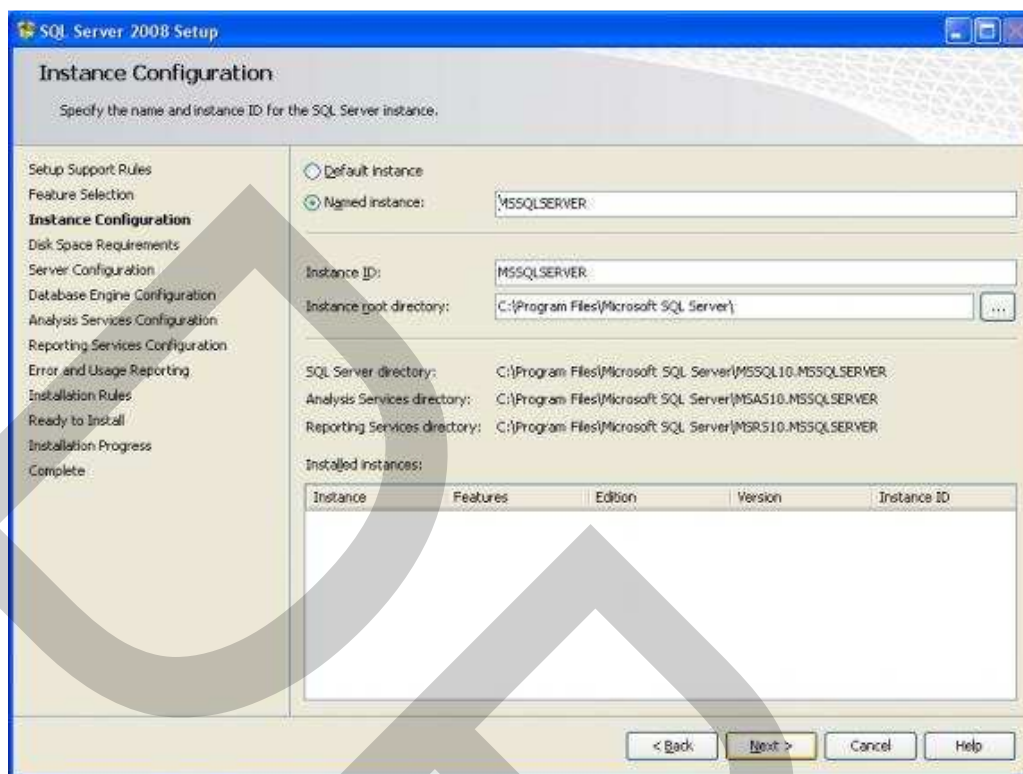




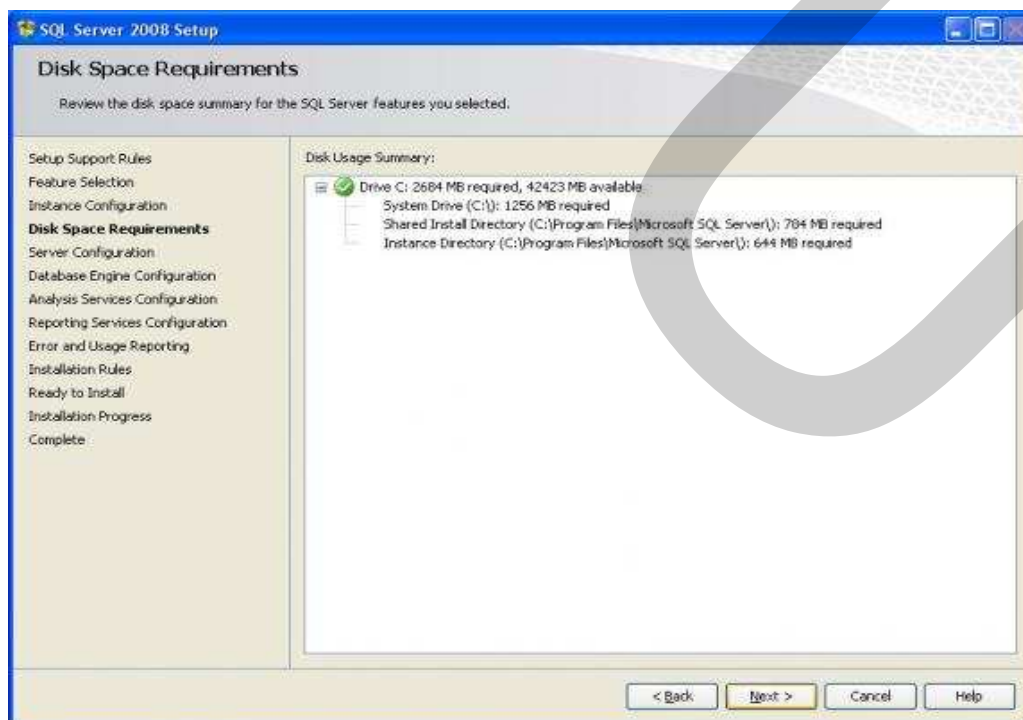
เลือก Select All แล้วกด Next

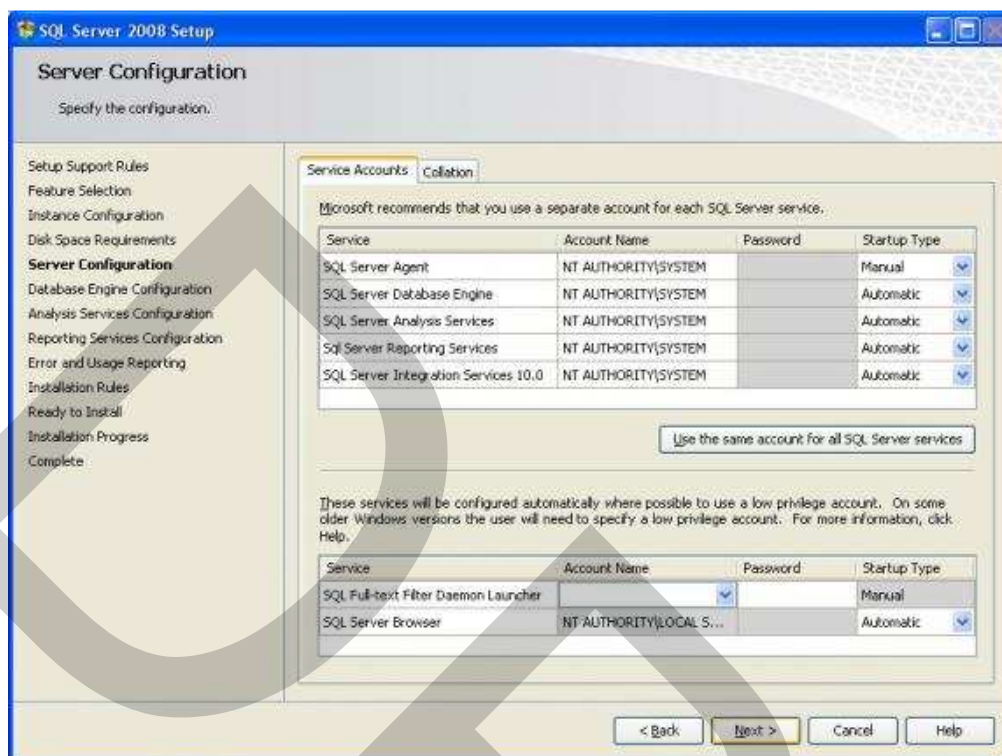


เลือก Named instance แล้วกด Next

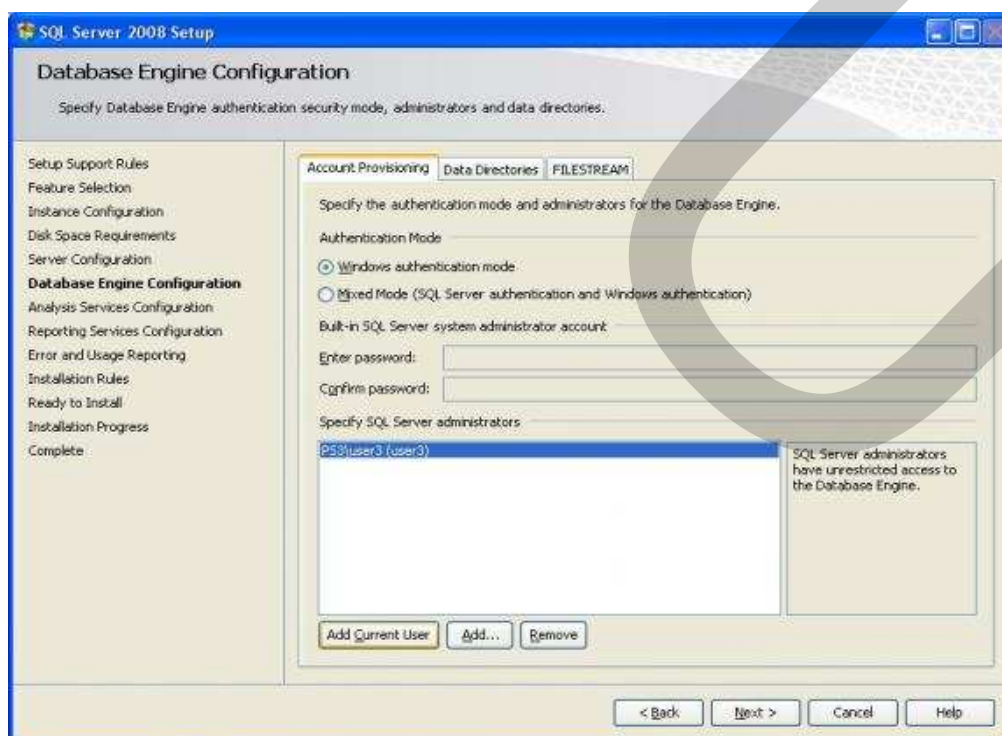


ส่วนนี้เป็นการตรวจสอบเนื้อที่ว่างพอหรือไม่ ถ้ามีเครื่องหมายถูกตามภาพแสดงว่าพอ ให้กด Next

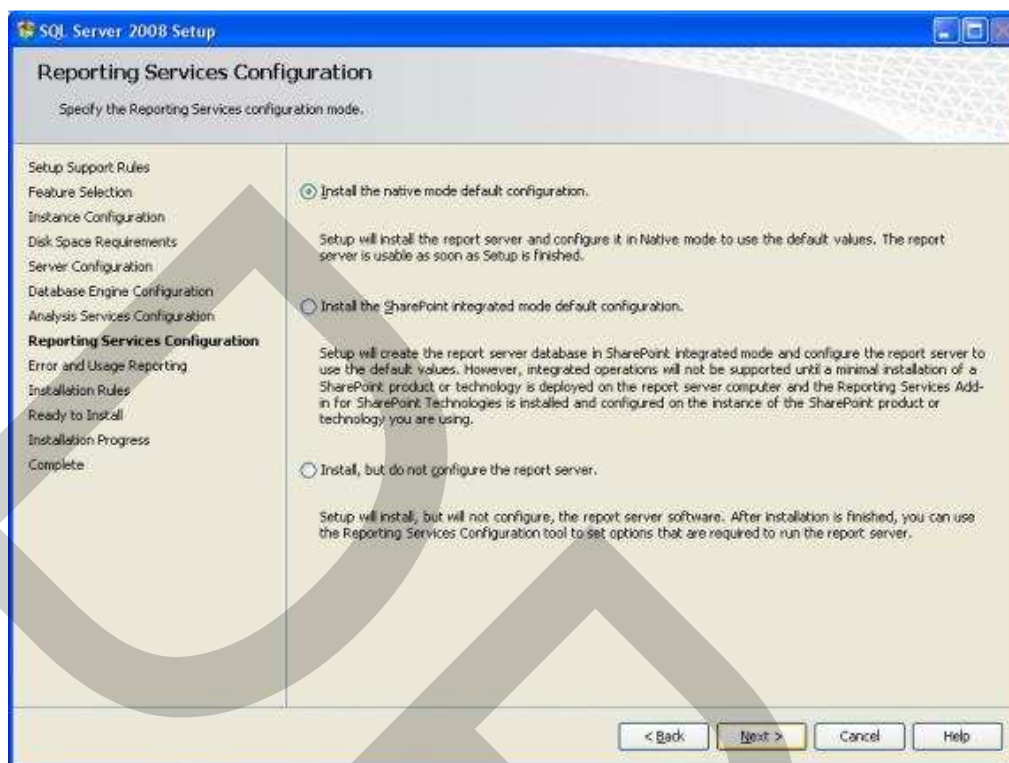




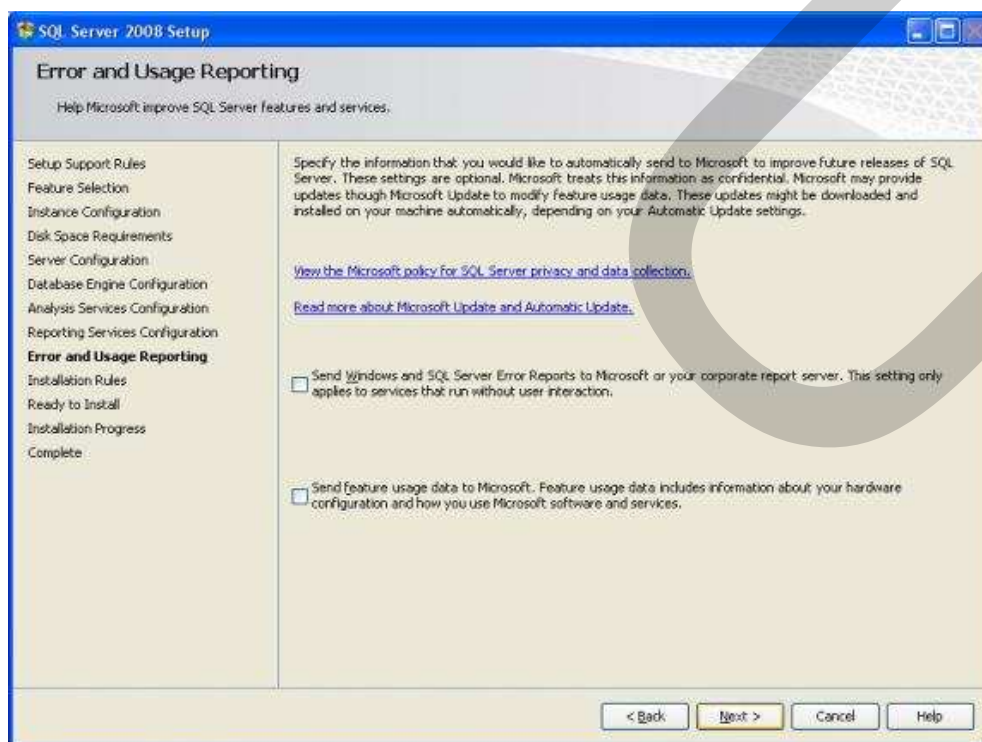
เลือกให้ครบตามภาพ NT AUTHORITY/SYSTEM แล้วเลือก Automatic ถ้าเลือกหมดแล้วกด Next



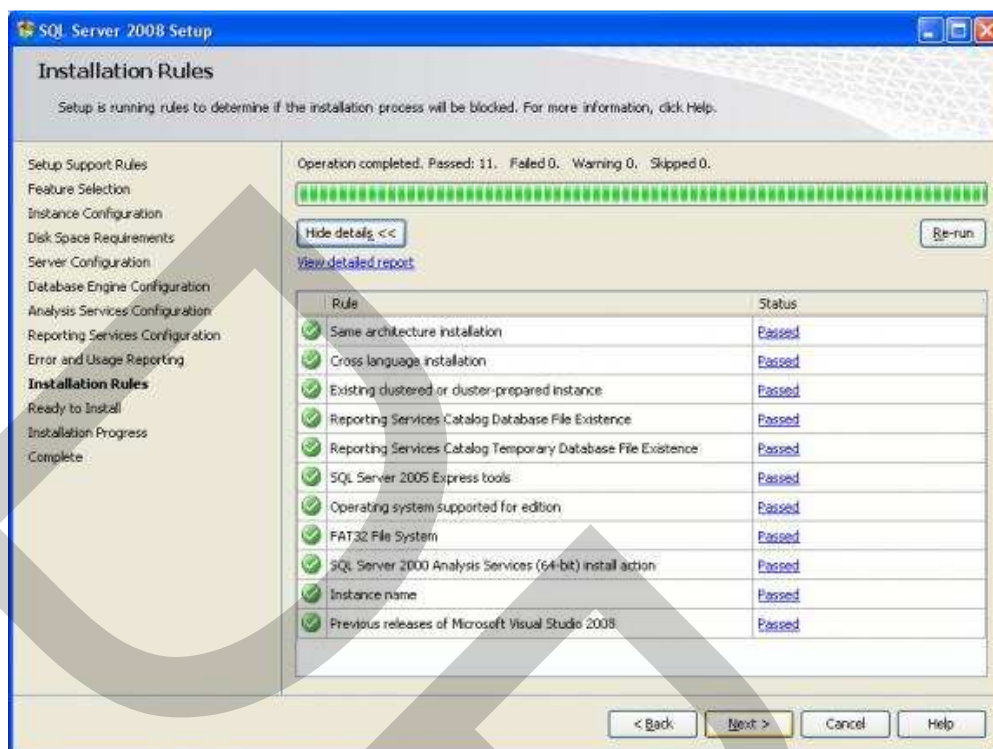
กด Add Current User ให้ได้ชื่อเครื่องและ User ของเครื่องคุณ แล้วกด Next



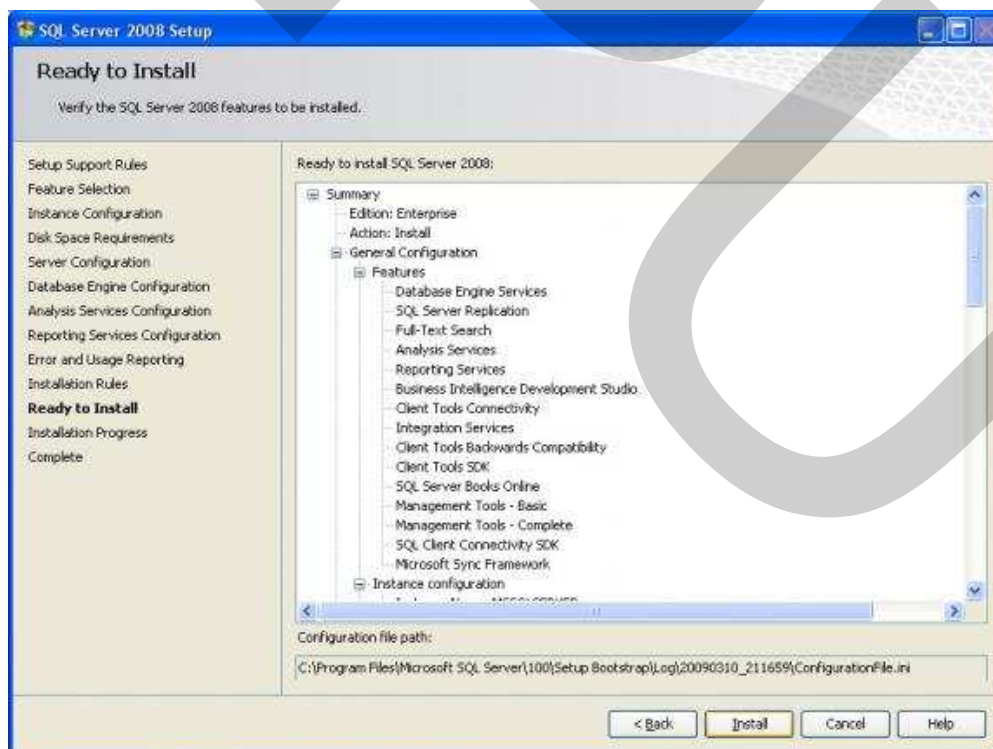
กด Next



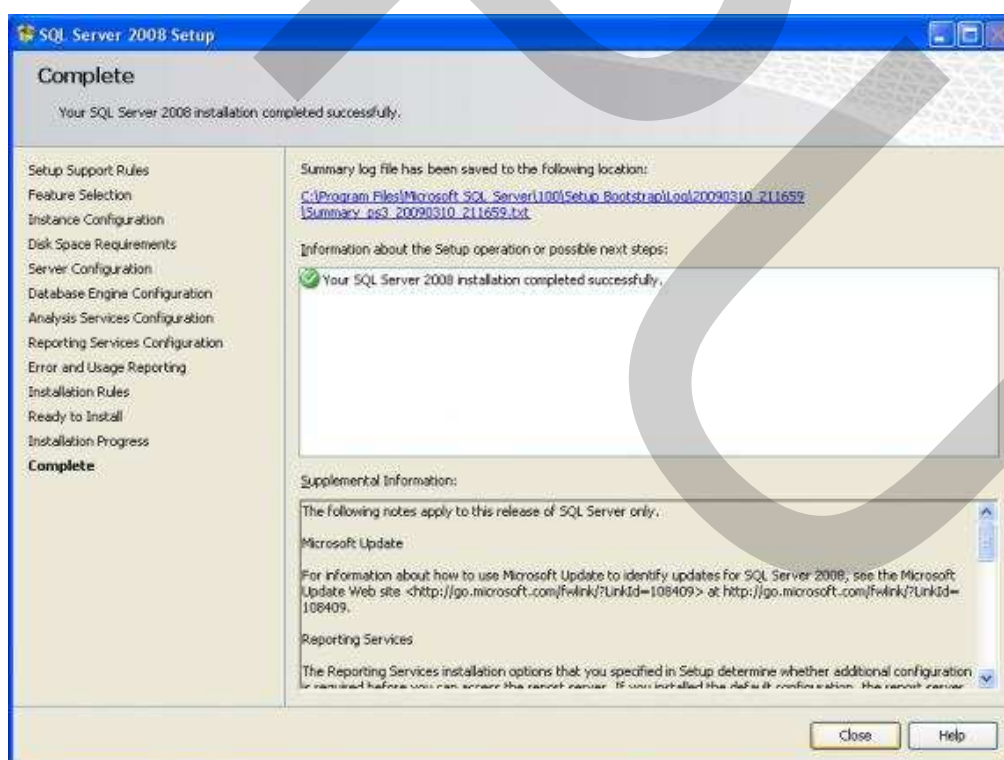
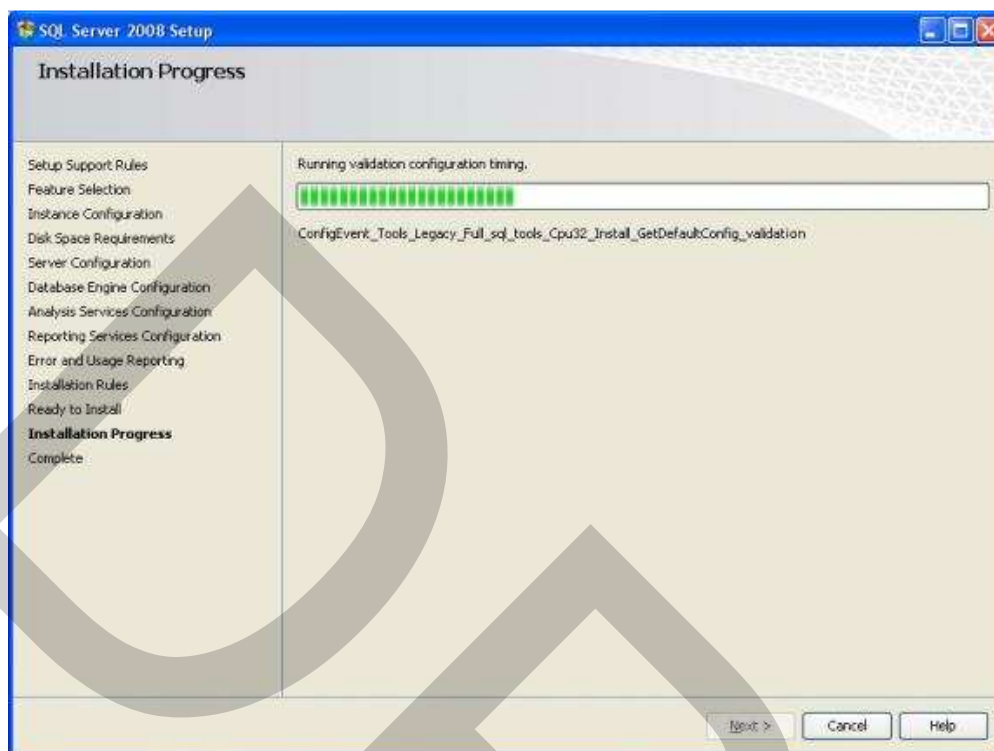
กด Show Details เพื่อดูว่าผ่านหมด ถ้าได้ตามภาพก็ผ่าน



กด Install เริ่มขั้นตอนการติดตั้ง



กำลังติดตั้ง รอ.....



การติดตั้งเสร็จสิ้น กด Close แล้ว รีสตาร์ทเครื่อง จบเป็นอันใช้ได้

ประวัติผู้เขียน

ชื่อ-นามสกุล

นายเสถียร วิชาเรือง

ประวัติการศึกษา

บริหารธุรกิจบัณฑิต คณะบริหารธุรกิจ

มหาวิทยาลัยราชภัฏสวนสุนันทา พ.ศ.2549

ตำแหน่งและสถานที่ทำงานปัจจุบัน

เจ้าหน้าที่บริหารลูกค้าหนี้

บริษัทเอกชนหลวงลิซซิ่ง จำกัด

ตั้งอยู่ที่ 2170 อาคารกรุงเทพทาวเวอร์ ชั้น 5 ถนน

เพชรบุรีตัดใหม่ แขวงบางกะปิ เขตห้วยขวาง

กรุงเทพมหานคร

ประสบการณ์ทำงาน

- เจ้าหน้าที่บริหารลูกค้าหนี้ (10 ปี)

- ติดตั้งและบำรุงรักษาอุปกรณ์คอมพิวเตอร์ (3 ปี)

ทุนการศึกษา

ทุนข้าราชการ เจ้าหน้าที่ของรัฐ พนักงานรัฐวิสาหกิจ
และวิสาหกิจเอกชน