# การพัฒนาระบบดาต้าไมนิ่งในการจำแนกกลุ่มเครดิตลูกค้าโดยใช้ดิซิชั่นทรี

# System Development of Data Mining for Classification Credit Types of Clients Using Decision Tree

โดย

นางสาวมณฑิรา ไวพ่อค้า

รหัส 45066060

อาจารย์ที่ปรึกษา

ผศ.ดร.อาริต ธรรมโน



รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2546
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

| | |
|---|---|
| **ชื่อหัวข้อ** | การพัฒนาระบบดาต้าไมนิ่งในการจำแนกกลุ่มเครดิตลูกค้า โดยใช้ดิซิชั่นทรี |
| **นักศึกษา** | นางสาวมณฑิรา ไวพ่อค้า |
| **อาจารย์ที่ปรึกษา** | ผศ.ดร.อาริต ธรรมโน |
| **ระดับการศึกษา** | วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ |
| **แขนงวิชา** | วิทยาการสารสนเทศ |
| **ปีการศึกษา** | 2546 |

# บทคัดย่อ

ในธุรกิจการให้บริการสินเชื่อนั้น เป็นธุรกิจที่มีความเสี่ยงสูง เนื่องจากมีโอกาสที่จะเกิดกลุ่มลูกค้าที่มีเครดิตไม่ดีได้ ได้ซึ่งเทคนิคดาต้าไมนิ่งสามารถเข้ามาช่วยในการวิเคราะห์ลักษณะข้อมูลและจำแนกกลุ่มลูกค้าได้ เพื่อช่วยลดความเสี่ยงในการอนุมัติสินเชื่อ ดังนั้นโครงการนี้จะทำการพัฒนาระบบในการจัดแบ่งกลุ่มเครดิตลูกค้า เพื่อทำนายกลุ่มลูกค้าที่มีโอกาสค้างชำระได้ โดยใช้ Decision Tree ด้วย C4.5 อัลกอริทึมซึ่ง Decision Tree ที่ได้สามารถเป็นประโยชน์ในการนำไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อเพื่อลดความเสี่ยงได้

| | |
|---|---|
| **Title** | System Development of Data Mining for classification credit types of clients Using Decision Tree |
| **Student** | Miss Montira Waiporka |
| **Advisor** | Dr. Arit Thammano |
| **Level of Study** | Master of Science in Information Science |
| **Major** | Information Science |
| **Academic Year** | 2003 |

# ABSTRACT

A financial institution such as bank or credit union has high risk from loan defaults. Data mining technique can solve this problem. To predict whether or not an applicant will be a good or poor credit risk. This project is to develop system for classificatin credit types of clients with C4.5 algorithm in decision tree technique. Decision Tree, which is result from system, can help in coming to a decision in order to decrease risk.

# กิตติกรรมประกาศ

การจัดทำโครงงานพัฒนาระบบนี้สำเร็จลุล่วงไปด้วยดี เนื่องจากคำแนะนำ และความช่วยเหลือจากบุคคลต่างๆดังต่อไปนี้

บิดา และมารดา ที่ให้การสนับสนุนการศึกษา ช่วยเหลือ และให้กำลังใจในการฝ่าฟันอุปสรรคต่าง ๆ จนสำเร็จการศึกษา

ท่าน ผศ.ดร. อาริต ธรรมโน อาจารย์ที่ปรึกษาที่ให้คำปรึกษา และแนะนำแนวทางในการศึกษา และพัฒนาโครงการ ตั้งแต่เริ่มต้น จนกระทั่งสำเร็จ

เพื่อน ๆ IS13 ที่ให้คำแนะนำ และความช่วยเหลือ พร้อมทั้งกำลังใจในการพัฒนาโครงงานนี้มาโดยตลอด

ด้วยความขอบคุณเป็นอย่างสูง

มณฑิรา ไวพ่อค้า

กุมภาพันธ์ 2546

# สารบัญ

หน้า

# สารบัญ (ต่อ)

# สารบัญตาราง

หน้า

# สารบัญภาพ

หน้า

ภาพที่

# บทที่ 1
# บทนำ

## 1.1 หลักการและเหตุผล

ทุกองค์กรจำเป็นต้องพัฒนาตนเองอยู่ตลอดเวลาและส่วนหนึ่งก็เพื่อที่จะสามารถแข่งขันกับคู่แข่งได้ ซึ่งข้อมูลก็เป็นส่วนหนึ่งที่สำคัญ การที่เรามีข้อมูลอยู่จำนวนมากและสามารถดึงส่วนที่เป็นประโยชน์ออกมาใช้ได้ ก็จะเป็นการสร้างโอกาสในการเพิ่มศักยภาพให้กับองค์กรมากขึ้นแต่ข้อมูลที่มีอยู่จำนวนมากนั้น เกินความสามารถของบุคคลที่จะวิเคราะห์ออกมาได้อย่างมีประสิทธิภาพและถูกต้อง จึงได้นำเอาเทคนิคของดาต้าไมนิ่ง มาช่วยในการวิเคราะห์และทำการดึงข้อมูลที่มีประโยชน์ออกจากฐานข้อมูลขนาดใหญ่ได้ ซึ่งข้อมูลที่ได้มานี้จะมีประโยชน์ในการเพิ่มมูลค่าให้กับองค์กรหรือลดความเสียหายที่อาจเกิดขึ้นได้ ตัวอย่างเช่นในธุรกิจการให้บริการสินเชื่อนั้น เป็นธุรกิจที่มีความเสี่ยงสูง ดังนั้นในการตัดสินใจอนุมัติสินเชื่อจึงต้องผ่านการพิจารณาและวิเคราะห์อย่างถี่ถ้วน ซึ่งสามารถนำเทคนิค Data Mining เข้ามาช่วยในการวิเคราะห์ลักษณะข้อมูลและจำแนกกลุ่มลูกค้าได้

## 1.2 วัตถุประสงค์

เพื่อนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการวิเคราะห์ลักษณะของลูกค้าทั้งกลุ่มที่มีเครดิตที่ดีและไม่ดี เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อ เพื่อลดความเสี่ยงในการอนุมัติสินเชื่อให้แก่ลูกค้า รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้เพื่อพิจารณาปัจจัยอื่น ๆ ที่มีผลต่อการดำเนินธุรกิจ

## 1.3 ขอบเขตการดำเนินงาน

โครงการพัฒนาระบบงานนี้ เป็นการนำข้อมูลลูกค้าในธุรกิจการให้บริการสินเชื่อที่ได้ผ่านขั้นตอนต่าง ๆ ของ ดาต้าไมนิ่ง คือการคัดเลือกข้อมูล, การเตรียมข้อมูลก่อนประมวลผล และการแปลงข้อมูล มาทำการสร้างแบบจำลอง โดยอาศัยหลักการของ C4.5 อัลกอริทึม โดยจะนำเสนอในรูปของกฎและ Decision Tree เพื่อนำมาวิเคราะห์ลักษณะของลูกค้าในกลุ่มที่มีเครดิตที่ดีและไม่ดี

**1.4 ขั้นตอนการดำเนินงาน**

1. ศึกษาและเก็บรวบรวมข้อมูล
2. ศึกษาขั้นตอนและวิธีการทางดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้
3. ศึกษาอัลกอริทึม C4.5 เพื่อนำมาประยุกต์ใช้กับระบบ
4. ออกแบบและพัฒนาระบบงานเพื่อใช้ในการจัดแบ่งกลุ่มเครดิตลูกค้า
5. สรุปผลการศึกษา

**1.5 ประโยชน์ที่คาดว่าจะได้รับ**

จากการศึกษาเทคนิคของดาต้าไมนิ่ง และนำมาประยุกต์ใช้กับธุรกิจการให้บริการสินเชื่อ ในโครงงานนี้ คาดว่าจะทำให้เข้าใจหลักการและขั้นตอนของการทำดาต้าไมนิ่ง รวมทั้งเทคนิคการจัดหมวดหมู่ โดยการสร้างแบบจำลองด้วยดิซิชั่นทรี อีกทั้งยังเป็นแนวทางในการนำดาต้าไมนิ่งมาประยุกต์ใช้กับข้อมูลทางธุรกิจด้านอื่น ๆ หรือ เป็นแนวทางในการออกแบบและพัฒนาโปรแกรมวิเคราะห์ข้อมูลโดยใช้วิธีการอื่น ๆ ต่อไป

# บทที่ 2
# ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

## 2.1 ดาต้าไมนิ่ง (Data Mining)

Data mining เป็นกระบวนการในการค้นหาความรู้ (Knowledge) ที่สนใจ เช่นรูปแบบ (Patterns), การเปลี่ยนแปลง (Changes) จากข้อมูลขนาดใหญ่ที่ถูกเก็บอยู่ใน Database, Data warehouse หรือที่เก็บสารสนเทศอื่น ๆ ซึ่ง ดาต้าไมนิ่งสามารถมองความสัมพันธ์ในหลายมิติในข้อมูลขนาดใหญ่ได้และจะเน้นเฉพาะจุดเด่นหรือความพิเศษของข้อมูลนั้น โดยใช้การวิเคราะห์ทางสถิติและเทคนิคแบบจำลองในการหารูปแบบและความสัมพันธ์ของข้อมูลขององค์กร ซึ่งการใช้วิธีธรรมดาอาจไม่สามารถมองเห็นได้ ทำให้เกิดศักยภาพในการใช้ข้อมูล

## 2.2 กระบวนการของดาต้าไมนิ่ง (Data Mining Process)

กระบวนการของดาต้าไมนิ่ง เป็นกระบวนการของการสร้างแบบจำลอง (Model) โดยสร้างแบบจำลองของกลุ่มข้อมูลเพื่อสร้างความเข้าใจในแนวโน้ม รูปแบบ และความเกี่ยวข้องกันของกลุ่มข้อมูลเพื่อใช้ในการทำนายบนข้อมูลนั้น ๆ โดยสรุปแล้วกระบวนการของดาต้าไมนิ่งประกอบด้วย 5 ขั้นตอน ดังนี้

1. กำหนดจุดประสงค์ทางธุรกิจ (Business Objective Determination)
2. การเตรียมข้อมูล (Data Preparation)
3. การทำดาต้าไมนิ่ง (Data Mining)
4. การวิเคราะห์ผลลัพธ์ที่ได้ (Analysis of Result)
5. การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of knowledge)

**ขั้นตอนที่1 : กำหนดจุดประสงค์ทางธุรกิจ** ทำความเข้าใจกับข้อมูล และความต้องการทางธุรกิจขององค์กรก่อน ถ้าขาดพื้นฐานความเข้าใจในเรื่องเหล่านี้แล้วจะทำให้ไม่สามารถ ระบุปัญหาขององค์กรที่เราต้องการแก้ไขได้, ไม่สามารถเตรียมข้อมูลเพื่อทำ Mining หรือไม่สามารถตีความผลลัพธ์ที่ออกมาได้และเพื่อให้การทำ Data mining เกิดประโยชน์สูงสุดเราจำเป็นจะต้องเข้าใจในวัตถุประสงค์ในการทำอย่างชัดเจนซึ่ง

วัตถุประสงค์เหล่านั้นขึ้นอยู่กับ จุดมุ่งหมายของแต่ละองค์กรซึ่งในขั้นตอนนี้จะสามารถมองเห็นถึงอัลกอริทึมและข้อมูลที่จะใช้ในเบื้องต้นที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้

**ขั้นตอนที่2: การเตรียมข้อมูล** เป็นขั้นตอนที่ต้องใช้เวลาและความพยายามมากกว่าขั้นตอนอื่น ๆ ขั้นตอนการจัดเตรียมข้อมูลนี้ใช้เวลาถึง 60 เปอร์เซ็นต์ ซึ่งประกอบด้วย 3 ขั้นตอนย่อยคือ

**2.1 การเลือกข้อมูล (Data Selection)** คือการคัดเลือกข้อมูลสำหรับดาต้าไมนิ่งจากข้อมูลทั้งหมดขององค์กรโดยต้องคำนึงถึง วัถตุประสงค์ในการนำข้อมูลมาใช้งาน นอกจากนี้ต้องทำความเข้าใจกับข้อมูลและประเภทข้อมูลที่จะนำมาใช้ด้วย โดยประเภทข้อมูลแบ่งเป็น

- ข้อมูลตัวเลข (Quantitative) แบ่งเป็น
  - Discrete เป็นค่าจำนวนเต็ม เช่น จำนวนพนักงาน
  - Continuous เป็นเลขจำนวนจริงเช่น รายได้
- ข้อมูลที่ไม่ใช่ตัวเลข (Categorical Data) แบ่งเป็น
  - Nominal Categorical ลำดับไม่มีความสำคัญเช่น เพศ
  - Ordinal Categorical ลำดับมีความสำคัญเช่น เกรด

และสิ่งที่สำคัญอีกอย่างคือ อายุการใช้งานของข้อมูลที่เลือกเช่นข้อมูลใดมีอัตราการเปลี่ยนแปลงสูง (Variation) ก็ต้องทำการตรวจสอบให้แน่ชัดว่าข้อมูลนั้นถูกต้องหรือไม่ก่อนนำมาใช้งาน นอกจากนี้ยังมีหลักเกณฑ์ที่ต้องพิจารณาเพิ่มเติมเกี่ยวกับข้อมูลที่จะนำมาใช้อยู่ 4 ประเด็นคือ

1. ระดับของข้อมูลที่พิจารณา สิ่งที่นำมาช่วยตัดสินใจว่าข้อมูลที่นำมาใช้ควรเป็นข้อมูลระดับรายการ (Item) หรือ ข้อมูลที่สรุปแล้ว คือวัตถุประสงค์ในการทำดาต้าไมนิ่ง เช่น

- การทำไมนิ่งเกี่ยวกับการโทรศัพท์ ถ้าจุดประสงค์ของเราต้องการเน้นไปที่พฤติกรรมการใช้โทรศัพท์ของลูกค้า ข้อมูลที่จัดเก็บโดยปกติแล้วจะมีการจัดเก็บเป็นลักษณะรายละเอียดของแต่ละชุมสาย การเคลื่อนย้ายของอิเล็กตรอนไปยังสวิตชิ่ง ข้อมูลเหล่านี้จะไม่มีประโยชน์เลย เพราะจุดประสงค์ของเราสนใจสิ่งที่อยู่ภายใต้การควบคุมของลูกค้าและมีผลต่อการตลาด ดังนั้น ข้อมูลที่เราสนใจจะเป็น เบอร์โทรศัพท์ของผู้โทร, เวลาเริ่มต้นที่ใช้โทร และเวลาที่ใช้ในการโทรศัพท์แต่ละครั้ง

- ข้อมูลที่ยังไม่สรุป ทำให้จัดการได้ยาก รวมทั้งเกิดจำนวนการ Combination สูง เมื่อใช้เทคนิคของ Association Discovery เพราะข้อมูลของร้านค้าปลีก ย่อมมีรายการสินค้ามาก ดังนั้นการนำเอาหน่วยวัดในการจัดเก็บสินค้าในคลัง (Stock Keeping Unit) เข้ามาช่วยจะสามารถลดจำนวนการคอมไบเนชั่นลงได้

2. ลักษณะของข้อมูลที่จัดเก็บ การจัดเก็บข้อมูลด้วยภาษาคอมพิวเตอร์ที่แต่ละระบบปฏิบัติการเลือกใช้แตกต่างกัน ทำให้ข้อมูลที่นำมาวิเคราะห์มีผลกระทบ เช่น ข้อมูลที่นำมาวิเคราะห์ส่วนมากจัดเก็บด้วยภาษา COBOL และ RPG ข้อมูลที่เป็น Text จะถูกเก็บเป็น EBCDIC และข้อมูลตัวเลขจะเก็บเป็น Packed Decimal ขณะที่ภาษาที่เลือกใช้ในการสร้างระบบ ดาต้าไมนิ่ง เช่น ภาษาซี ข้อมูลชนิด Text จะมีรูปแบบเป็น ASCII และข้อมูลตัวเลขเก็บเป็น Integer หรือ Floating Point

3. ความแตกต่างของข้อมูลแต่ละแหล่ง เมื่อข้อมูลที่นำมาวิเคราะห์มาจากหลายแหล่ง ซึ่งแต่ละแหล่งมีรูปแบบการจัดเก็บข้อมูลที่ต่างกัน เช่น การวิเคราะห์ข้อมูลการโทรศัพท์ เพื่อหาเบอร์โทรศัพท์ที่ใช้ฝากข้อความเข้า Voice Mailbox ในแต่ละเมือง จะมีวิธีการจัดเก็บข้อมูลที่ต่างกัน เช่น เมือง ๆ หนึ่งอาจเก็บเบอร์โทรศัพท์ที่ใช้โทรเข้า Voice Mailbox ด้วยต้นทางและปลายทาง แต่อีกเมืองหนึ่ง อาจเก็บเบอร์โทรศัพท์ที่ไม่รู้ด้วยเบอร์ปลายทาง อีกเมืองหนึ่งอาจเก็บเบอร์โทรศัพท์ที่โทรเข้า Voice Mailbox จริง ๆ ดังนั้น จึงจำเป็นต้องทำข้อมูลเหล่านี้ให้ออกมาในรูปแบบมาตรฐานเดียวกันก่อน เพื่อที่จะได้เข้าใจถึงความแตกต่างในการเก็บข้อมูลของแต่ละแหล่งได้

4. ข้อมูลที่เป็นข้อความ ข้อมูลที่จัดเก็บแบบ Text อาจก่อให้เกิดความสับสน เช่น '_no' กับ 'no_' ซอฟท์แวร์ที่ใช้ในการทำดาต้าไมนิ่งย่อมมองข้อมูลเหล่านี้ไม่เหมือนกัน ในทางแก้ไขคือสร้างตารางเก็บค่าที่ถูกต้อง และแทนที่ข้อมูลที่นำมาวิเคราะห์ด้วย index ตัวอย่างที่เห็นได้ชัดเจนคือ Relational Database มีการแทนที่ข้อมูลที่เป็น Product_Name ด้วย Product_code ซึ่งมีความเป็น unique มากกว่า

**2.2 การตรวจสอบข้อมูล (Data Preprocess)**

ในกระบวนการนี้จะมีปริมาณข้อมูลจำนวนหนึ่งที่ถูกเลือกเข้ามาจากกระบวนการ Data Selection ซึ่งข้อมูลเหล่านี้เราจะต้องนำมาตรวจสอบว่าเป็นข้อมูลที่ถูกต้อง เหมาะสมหรือไม่โดยใช้หลักการทางสถิติเช่น การวัดการกระจายของข้อมูล หรือนำข้อมูลนั้นมาสร้างเป็นกราฟ เพื่อช่วยให้เห็นความโน้มเอียงของข้อมูลและข้อมูลที่ผิดปกติได้ ส่วนข้อมูลที่อยู่ในลักษณะที่เป็นตัวเลขก็สามารถวิเคราะห์ได้โดย การหาค่าสูงสุด, ต่ำสุด, ค่าเฉลี่ย, ค่ากลาง ซึ่งสิ่งที่อาจจะปรากฏให้เห็นได้คือ

- Noisy Data คือ ค่าของข้อมูลผิดไปจากค่าที่ควรจะเป็นเช่น ข้อมูลอายุ 200 ปี ซึ่งอาจเกิดจากความผิดพลาดในการป้อนข้อมูล
- มีข้อมูลบางส่วนหายไป แก้ไขโดยทำการตัดข้อมูลนั้นทิ้งไปทั้งรายการ หรือแทนส่วนที่หายไปด้วยค่าเฉลี่ย (Mean) หรือค่าที่ปรากฏบ่อย (Mode) หรือบันทึกเป็น 'UNKNOWN'

**2.3 การแปลงข้อมูล (Data Transformation)**

เป็นการปรับเปลี่ยนรูปแบบข้อมูลให้เหมาะสมกับ อัลกอริทึมที่เลือกใช้ เช่นแปลงข้อมูลตัวเลขให้เป็นช่วงเพื่อใช้กับ Decision Tree

**ขั้นตอนที่ 3 : การทำดาต้าไมนิ่ง** ในขั้นนี้เป็นการเลือก เทคนิคการสร้าง Model และนำมาประยุกต์ใช้ ซึ่งเป็นกระบวนการที่กลับมาทำซ้ำในขั้นตอนที่ทำไปแล้วได้ (Iterative Process) จะต้องเลือก Operation ที่เหมาะสม และมีประโยชน์ในการแก้ปัญหาทางธุรกิจขององค์กรซึ่งสิ่งที่เราได้เรียนรู้จากการพิจารณาค้นหา Operation ที่เหมาะสมนั้นจะนำไปสู่การย้อนกลับไปทำขั้นตอนที่ผ่านมาและทำการเปลี่ยนแปลงข้อมูลที่ใช้หรือดัดแปลงแก้ไขหัวข้อปัญหานั้น (Problem Statement) โดยจุดมุ่งหมายในการทำ ดาต้าไมนิ่ง มี 2 ประเภทคือ การทำนาย (Prediction) และการหาลักษณะของข้อมูล (Description) Prediction คือการทำนายค่าในอนาคตหรือค่าที่เราไม่รู้ของ Attribute ที่สนใจ โดยใช้ Attribute ในฐานข้อมูล ในขณะที่ Description คือการหารูปแบบ (Pattern) เพื่ออธิบายข้อมูลในรูปแบบที่บุคลากรสามารถเข้าใจได้ง่าย Data mining มี Operation หลัก ๆ 4 Operation ด้วยกัน คือ

1) Predictive Modeling มีคุณสมบัติบางอย่างเหมือนกับประสบการณ์การเรียนรู้ของมนุษย์ โดยใช้การสังเกตสิ่ง ๆ หนึ่ง และนำมาสร้างรูปแบบโมเดลตามลักษณะที่สำคัญของสิ่งนั้น ๆ ในดาต้าไมนิ่งนั้นใช้ Predictive Model ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อบ่งชี้ลักษณะที่จำเป็นเกี่ยวกับข้อมูลนั้น ซึ่งข้อมูลจะต้องมีความสมบูรณ์ โดยจะสังเกตได้จากความสามารถในการทำนายได้อย่างถูกต้องของโมเดล โมเดลจะต้องให้คำตอบที่ถูกต้องตรงกับคำตอบของสิ่งที่ได้พิสูจน์แล้ว ก่อนที่จะเริ่มนำมาใช้ในการทำนายจริง ซึ่งวิธีการแบบนี้เรียกว่า Supervised Learning

Prediction Model เป็นโมเดลในการวิเคราะห์ข้อมูลที่มีอยู่เพื่อทำนายแนวโน้มของข้อมูลที่จะเกิดขึ้นในอนาคต โดยในการสร้างโมเดลประกอบด้วย 2 ช่วงคือ Training และ Testing ซึ่ง Training เป็นช่วงของการสร้างโมเดลใหม่โดยใช้ข้อมูลเก่าที่มีอยู่แล้ว และ Testing คือการใช้ข้อมูลที่ไม่เคยใช้ในการสร้างโมเดลมาทำการทดสอบความถูกต้องของโมเดล และโมเดลสามารถเป็นได้ทั้งกลุ่มคำสั่งของ SQL, เงื่อนไข IF THEN, หรือกลุ่มคำสั่งภาษาซี และ Predictive Modeling แบ่งเป็น 2 แบบดังนี้คือ

- Classification การแบ่งข้อมูลออกเป็นกลุ่ม และใช้ Predictive Model ทำนายว่าข้อมูลควรอยู่ในกลุ่มใด เช่น ในการให้เงินกู้ จะทำนายว่าลูกค้าควรอยู่ในกลุ่มลูกค้าชั้นใด

- Value Prediction (Forecasting) ใช้ Predictive Model ในทำนายค่าที่สัมพันธ์กับข้อมูลที่มีอยู่ เช่นการพยากรณ์อากาศหรือการทำนายหุ้นเป็นต้น

2) Database Segmentation (Clustering) เป็นวิธีการนำข้อมูลมาแบ่งเป็นกลุ่มที่มีความสัมพันธ์กัน เพื่อวิเคราะห์หาลักษณะที่เหมือน หรือความแตกต่างของข้อมูลในแต่ละกลุ่ม อัลกอริทึมการแบ่งกลุ่ม (Segmentation Algorithm) สามารถทำการแบ่งกลุ่มของข้อมูลในลักษณะใด การแบ่งกลุ่มของข้อมูล (Segmentation) สนับสนุนแอปพลิเคชั่นทางธุรกิจต่าง ๆ เช่น ข้อมูลประวัติของลูกค้า (customer profile) หรือกลุ่มตลาดเป้าหมาย (target marketing) และรักษาลูกค้าให้ใช้บริการต่อไป (customer retention)

3) Link Analysis เป็นการค้นหาความสัมพันธ์ (Associations) ระหว่างข้อมูล หรือกลุ่มข้อมูล เช่น การหาความสัมพันธ์ระหว่างสินค้า หรือบริการที่ลูกค้าชอบซื้อร่วมกัน, การซื้อสินค้าประเภทหนึ่ง แล้วจะซื้อสินค้าอีกประเภทหนึ่งต่อเนื่องกัน เป็นต้น แบ่งเป็น 3 แบบคือ

- Association Discovery ใช้ในการวิเคราะห์การซื้อสินค้าเพื่อหาความสัมพันธ์ที่ซ่อนอยู่ระหว่างผลิตภัณฑ์ซึ่งจะขายได้ดีเมื่อขายคู่กัน การวิเคราะห์ในลักษณะนี้ถูกเรียกว่า Market Basket Analysis

- Sequential Pattern Discovery ใช้ในการกำหนดความสัมพันธ์ในการซื้อสินค้าที่ไม่มีความเกี่ยวข้องกันในช่วงเวลาหนึ่ง ๆ ที่มีการแสดงข้อมูลเป็นลำดับในการซื้อสินค้าและใช้บริการ เช่นเมื่อซื้อพัดลมแล้วต่อมาอาจจะซื้อแอร์มาใช้ เป็นต้น ช่วยทำให้เข้าใจพฤติกรรมการซื้อของลูกค้าและนำมาจัดรายการส่งเสริมการขาย

- Similar Time Sequence Discovery เป็นการค้นหาความสัมพันธ์ระหว่างข้อมูลสองกลุ่มในช่วงระยะเวลาหนึ่งเช่น รายเดือน, รายปี โดยเทียบระดับความเหมือนกันระหว่างแบบ 2 แบบ (Patterns) ในช่วงระยะเวลาที่ทำการทดลองเดียวกัน

4) Deviation Detection เป็นการวิเคราะห์ว่ามีอะไรแตกต่างจากกลุ่มอื่น โดยใช้กราฟหรือรูปภาพ เพื่อแสดงให้เห็นความแตกต่างจากกลุ่มอื่น แอปพลิเคชั่นที่อัลกอริทึมนี้สนับสนุนมีทั้ง การป้องกันการโกง (Fraud detection) ในการใช้บัตรเครดิต ในสินไหมการประกัน หรือในการใช้บัตรโทรศัพท์ และการควบคุมคุณภาพ

**ขั้นตอนที่ 4 : การวิเคราะห์ผลลัพธ์ที่ได้** เป็นการวิเคราะห์และตีความผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง การทำงานในขั้นตอนนี้ต้องใช้ทักษะในการวิเคราะห์ข้อมูล และการวิเคราะห์ทางธุรกิจ ซึ่งทำโดยการนำเอาแบบจำลองที่ได้ไปทดสอบกับข้อมูลชุดอื่น ที่ไม่ใช่ข้อมูลที่ใช้ในการสร้างแบบจำลอง เพื่อนำเอาผลลัพธ์ที่ได้มาเปรียบเทียบกับผลตามแบบจำลอง ว่ามีความแม่นยำและยอมรับได้หรือไม่ ซึ่งถ้าไม่สามารถยอมรับได้ก็ทำการแก้ไข โดยการเพิ่มจำนวนของข้อมูลให้มากขึ้นหรือเปลี่ยนไปใช้อัลกอริทึมอื่นแทน

**ขั้นตอนที่ 5 : การปรับความรู้ที่ได้เข้ากับธุรกิจ** เป็นการรวบรวมความเข้าใจทางธุรกิจที่เป็นผลมาจากขั้นตอนที่ 4 มารวมเข้ากับส่วนความรู้เพื่อนำไปใช้ในโอกาสต่อไป ในขั้นตอนนี้มีหลักอยู่ 2 ประการคือ การเสนอแนวคิดทางธุรกิจที่ค้นพบใหม่ และหาแนวทางที่จะใช้กฎเกณฑ์ใหม่ที่ค้นพบเพื่อให้เกิดประโยชน์สูงสุด

# บทที่ 3

## การจัดกลุ่ม (Classification)

การจัดหมวดหมู่ของข้อมูล เป็นการเรียนรู้ฟังก์ชั่นซึ่งแบ่งประเภทข้อมูลแต่ละหน่วยออกตาม Class ที่ได้กำหนดไว้ก่อนแล้ว งานของการทำ Classification คือการวิเคราะห์ Training data และพัฒนาเป็น Model ของแต่ละ Class เพื่อแสดงถึงลักษณะของข้อมูล ส่วนข้อมูลที่จะนำมาทดสอบในอนาคตนั้นจะถูกแบ่งประเภทโดยใช้ Model ของ Class

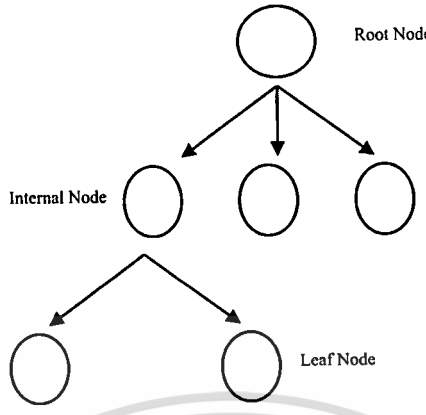
เทคนิคที่ใช้ในการจัดหมวดหมู่แบ่งเป็น 2 แบบ ได้แก่ Tree Induction และ Neural Induction โดยในที่นี้จะนำเสนอเทคนิคของ Tree Induction ในการประยุกต์ใช้งานในธุรกิจการให้บริการสินเชื่อ โดย Tree Induction เป็นการนำข้อมูลมาสร้างแบบจำลองพยากรณ์ในรูปของ Decision Tree

### 3.1 ขั้นตอนพื้นฐานในการสร้าง Tree

Decision tree เป็น Flow-chart ที่มีโครงสร้างเป็น Tree ซึ่งแต่ละ Internal node จะแสดงถึง Attribute ที่จะใช้ Test, แต่ละกิ่ง (Branch) ของ Tree จะแสดงผลของการทดสอบ และ Leaf node จะแสดงถึง Class ส่วน Node สูงสุดของ Tree คือ Root node ลักษณะของ Tree แสดงดังรูป ที่ 3.1 การนำข้อมูลมาสร้าง Decision Tree มีขั้นตอนพื้นฐานคือ

- หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกนำมาสร้างเป็น Root node
- นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกแตกออกมาเป็นกลุ่ม
- แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root Node
- นำข้อมูลแต่ละกลุ่มมาทำซ้ำขั้นตอนแรกคือ หา Attribute ที่สำคัญที่สุด

สำหรับอัลกอริทึมที่ใช้สร้าง Decision Tree มีหลายอัลกอริทึม เช่น CHAID, CART, ID3 และ C4.5 เป็นต้น และในแต่ละอัลกอริทึมจะมีวิธีการที่แตกต่างกันในการหา Attribute ที่จะนำมาใช้แบ่งข้อมูล ซึ่งในที่นี้ได้เลือกอัลกอริทึม C4.5 มาประยุกต์ใช้ในการพัฒนาระบบเพื่อช่วยในการจัดแบ่งกลุ่มเครดิตลูกค้า และเนื่องจาก C4.5 เป็นเวอร์ชันที่พัฒนาจาก ID3 จึงขอกล่าวถึงการทำงานของ ID3 และ C4.5




รูปที่ 3.1 รูปแบบ Decision Tree

**3.2 ID3 Algorithm**

ID3 Algorithm มีการสร้าง Tree ตามอัลกอริทึมดังนี้

1. Tree เริ่มต้นด้วย Node หนึ่ง Node แสดงถึง ข้อมูลที่ใช้ Train (Sample)
2. ถ้า Sample เป็น Class เดียวกัน Node ก็จะกลายเป็น Leaf และมีชื่อตาม class
3. ถ้าไม่เช่นนั้น จะใช้การวัดแบบ Entropy-based ที่เรียกว่า Information gain เพื่อเลือก Attribute ที่เหมาะที่สุดที่จะเป็นตัวแยก Sample ออกเป็นแต่ละ Class ซึ่ง Attribute นี้จะกลายมาเป็น Attribute ที่ใช้ในการทดสอบหรือใช้ในการตัดสินใจที่ node กิ่งของ Tree ถูกสร้างตามแต่ละค่าของ Attribute และ Sample จะถูกแบ่งตามค่าของ Attribute นั้น แล้วจะทำซ้ำกระบวนการเดิมโดยใช้ Sample จากแต่ละส่วนที่ถูกแบ่งออกมา กระบวนการทำซ้ำจะหยุดก็ต่อเมื่อ เงื่อนไขเหล่านี้เป็นจริง

1. Sample ทั้งหมดอยู่ใน Class เดียวกัน
2. Sample ไม่มี Attribute เหลืออยู่อีก
3. ไม่มี Sample แล้ว

ต่อไปนี้จะเป็นการแสดงถึง การเลือก Attribute ที่มีความสำคัญที่สุดเพื่อใช้แบ่งข้อมูล โดยกำหนดให้

| | | |
|---|---|---|
| T | แทน | Training Set |
| S | แทน | Set ของข้อมูลใด ๆ |
| p | แทน | ความน่าจะเป็นที่ Sample นั้นจะเป็นของ Class $C_i$ |
| m | แทน | จำนวน Class |

$$\text{Info}(S) = -\sum_{i=0}^{m} p_i \log_2(p_i)$$

และเมื่อนำสูตรนี้ไปประยุกต์ใช้กับ Training Set จะได้ Info(T), $\text{Info}_x(T)$ เป็นการวัดค่าของ information เพื่อแบ่ง T โดยใช้ค่าที่เป็นไปได้ของ Attribute X

$$\text{Info}_x(T) = \sum_{j=1}^{n} \frac{|T_i|}{|T|} \times \text{Info}(T_i)$$

Gain (X) เป็นการวัดค่าของ information ที่ได้รับถ้าเลือก Attribute X

$$\text{Gain}(X) = \text{Info}(X) - \text{Info}_x(T)$$

ต่อไปจะเป็นการอธิบายการทำงานของ ID3 โดยใช้ข้อมูลตัวอย่างจากตารางที่ 6 ดังนี้

| Outlook | Temp(°F) | Humidity(%) | Windy | Class |
|---|---|---|---|---|
| Sunny | 75 | 70 | True | Play |
| Sunny | 80 | 90 | True | Don't play |
| Sunny | 85 | 85 | False | Don't play |
| Sunny | 72 | 95 | False | Don't play |
| Sunny | 69 | 70 | False | Play |
| Overcast | 72 | 90 | True | Play |
| Overcast | 83 | 78 | False | Play |
| Overcast | 64 | 65 | True | Play |
| Overcast | 81 | 75 | False | Play |
| Rain | 71 | 80 | True | Don't play |
| Rain | 65 | 70 | True | Don't play |
| Rain | 75 | 80 | False | Play |
| Rain | 68 | 80 | False | Play |
| Rain | 70 | 96 | False | Play |

**ตารางที่ 3.1** Training Set

จากตารางที่ 6 จะเห็นว่าประกอบด้วย class 2 class คือ Play และ Don't Play โดยข้อมูลจำนวน 9 record อยู่ใน class Play และ 5 record อยู่ใน class Don't Play จะได้ว่า

$$\text{Info}(T) = -9/14 \times \log_2(9/14) - 5/14 \times \log_2(5/14)$$
$$= 0.940$$

พิจารณา Attribute ต่าง ๆ โดยหาค่า Gain ของแต่ละ Attribute ที่มีค่า Gain สูงสุดมาเป็นตัวแบ่งข้อมูล หรือ Root Node พิจารณาที่ Attribute Outlook ซึ่งสามารถแบ่งข้อมูลได้เป็น 3 subset จะได้ว่า

$$\text{Info}_x(T) = 5/14 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$$
$$+ 4/14 \times (-4/4 \times \log_2(4/4) - 0/4 \times \log_2(0/4))$$
$$+ 5/14 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))$$
$$= 0.694$$

ดังนั้นค่า Gain ของ Attribute Outlook จะมีค่าดังนี้

$$\text{gain}(x) = 0.940 - 0.694$$
$$= 0.048$$

จะพบว่าค่า Gain ที่ได้จากการแบ่ง Training Set โดยใช้ Attribute Outlook มากกว่า Windy ดังนั้นควรใช้ Attribute Outlook ในการแบ่ง Training Set แล้วนำ Sample ในแต่ละ กิ่งของ Attribute ที่ใช้ทดสอบมาทำซ้ำขั้นตอนตั้งแต่ต้นคือหา Attribute ที่มีค่าสูงสุดมาแบ่งข้อมูลต่อไป

ตามหลักการของ ID3 ต้องคำนวณหาค่า Gain ของทุก Attribute แล้วเลือก Attribute ที่ค่า Gain สูงสุด แต่จากข้อมูลตัวอย่างพบว่า ค่าใน Attribute Temp และ Humidity เป็นค่าชนิด Continuous ซึ่งกรณีนี้ ID3 ไม่สามารถจัดการได้ ต้องใช้ C4.5 ซึ่งจะกล่าวถึงต่อไป

### 3.3 C4.5 Algoithm

พัฒนามาจาก ID3 โดยเพิ่ม Feature ต่าง ๆ ขึ้นมาดังนี้

- **Gain ratio criterion** พัฒนาขึ้นเพื่อแก้ปัญหาของ Gain Criterion กรณีที่ Attribute มีค่าที่ unique การแบ่งข้อมูลโดยใช้ Attribute นี้จะทำให้เกิด subset จำนวนมากซึ่งแต่ละ subset จะประกอบด้วยข้อมูลเพียง 1 record เท่านั้น ทำให้ $\text{info}_x(T) = 0$ ซึ่งจะมีผลให้ค่า information Gain ของ Attribute นี้มีค่าสูงมาก และการแบ่งข้อมูลโดยใช้ Attribute นี้ไม่ก่อให้เกิดประโยชน์ใด ๆ ต่อ

การทำนาย C4.5 แก้ไขได้โดยใช้ค่า Gain ratio ซึ่งคำนวณโดยใช้ split info(x) และ gain ratio(x) โดย split info(x) เป็นค่า information ที่ได้จากการแบ่ง T ออกเป็น n subset

$$\text{split info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

gain ratio(X) เป็นการวัดว่าการแบ่งข้อมูลโดยใช้ Attribute นั้น ๆ ก่อให้เกิดประโยชน์ต่อการทำนายหรือไม่

$$\text{gain ratio}(X) = \text{gain}(X) / \text{split info}(X)$$

ซึ่งการใช้ Gain ratio criterion ทำให้ Tree ที่ได้มีขนาดเล็กกว่าการใช้ Gain criterion

● **Unknown attribute value** การหา Attribute เพื่อใช้แบ่งข้อมูล ทำโดย

- หาค่า info(T) และ $\text{info}_x(T)$ โดยพิจารณาเฉพาะข้อมูลที่รู้ค่าของ A
- หาค่า gain(X) โดย

$$\text{Gain}(X) = \text{probability A is know} \times (\text{info}(T) - \text{info}_x(T))$$

- หาค่า split info(X) โดยพิจารณาจากกลุ่มของข้อมูลที่ไม่รู้ค่าของ A เป็นอีก 1 subset เช่น ถ้า Attribute ที่จะนำมาทดสอบมีค่าที่เป็นไปได้ n ค่า split info(X) จะถูกคำนวณโดยแบ่งข้อมูลออกเป็น n+1 subsets

การแบ่ง Training Set สมมุติ Attribute ที่เลือกจากขั้นตอนแรกมีค่าที่เป็นไปได้คือ O1, O2, …., $O_n$ เมื่อข้อมูล 1 record ใน T ซึ่งมีค่า $O_i$ ถูกกำหนดให้ subset $T_i$ ค่าความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset $T_i$ เท่ากับ 1 และความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset อื่น ๆ เท่ากับ 0 แต่ถ้าค่าใน Attribute ไม่ทราบค่า ความน่าจะเป็นจะมีค่าน้อยลง สำหรับข้อมูลแต่ละ record ในแต่ละ subset $T_i$ weight จะเท่ากับความน่าจะเป็นของ $O_i$ ที่จุดนั้นๆ ทำให้ $|T_i|$ เป็นผลรวมของค่า weight w ซึ่งค่าใน Attribute ไม่ทราบค่าจะถูกกำหนดให้แต่ละ subset $T_i$ ด้วย weight

$$W \times \text{probability of outcome } O_i$$

โดยความน่าจะเป็นคือ ผลรวมของ weight ของข้อมูลทั้งหมดใน T ซึ่งมีค่า $O_i$ หารด้วยผลรวมของ weight ของข้อมูลทั้งหมดใน T ซึ่งค่าใน Attribute เป็นค่าที่ทราบค่า

การใช้ decision tree ที่ได้มาทำนายกลุ่มของข้อมูล ในกรณีที่ค่าใน attribute ที่จะทดสอบที่ decision node เป็นค่าที่ไม่ทราบค่า ทำให้ไม่สามารถแบ่งข้อมูลได้ กรณีนี้ระบบจะสำรวจทุกเส้นทางที่เป็นไปได้ และรวมผลที่ได้จากการ classification ด้วยวิธีการทางคณิตศาสตร์ โดยผลที่ได้จะเกิดได้จากหลายเส้นทางจาก root ของ tree หรือ subtree ไปยัง leaf node และ class ที่ได้จากการทำนายจะเป็น class ที่มีความน่าจะเป็นสูงสุด

ต่อไปจะนำเสนอตัวอย่างโดยใช้ตารางที่ 6 โดยแบ่งเป็น 3 ขั้นตอนดังนี้

1. การหา Attribute เพื่อใช้แบ่งข้อมูล สมมุติว่าค่าใน Attribute outlook ใน record ที่ 6 เป็นค่าที่ไม่ทราบค่า ซึ่งแทนโดย "?" ซึ่งเราจะพิจารณาเฉพาะข้อมูล 13 record ที่เหลือจะได้ความถี่ดังแสดงในตารางที่ 7 ทำการคำนวณค่าต่าง ๆ โดยพิจารณา Attribute Outlook ดังนี้

| | Play | Don't Play | Total |
|---|---|---|---|
| Outlook = sunny | 2 | 3 | 5 |
| Overcast | 3 | 0 | 3 |
| Rain | 3 | 2 | 5 |
| Total | 8 | 5 | 13 |

**ตารางที่ 3.2** แสดงความถี่ของข้อมูล

$$\text{Info}(T) = -8/13\times\log_2(8/13)-5/13\times\log_2(5/13)$$
$$= 0.9691$$

$$\text{info}_x(T) = 5/13\times(-2/5\times\log_2(2/5)-3/5\times\log_2(3/5))$$
$$+3/13\times(-3/3\times\log_2(3/3)-0/3\times\log_2(0/3))$$
$$+5/13\times(-3/5\times\log_2(3/5)-2/5\times\log_2(2/5))$$
$$= 0.747$$

$$\text{gain}(X) = 13/14\times(0.961-0.747)$$
$$= 0.199$$

จะพบว่าค่า gain ที่ได้จะลดลงเล็กน้อยจากเดิม 0.246 เป็น 0.199 bits ส่วนค่า split information จะพิจารณาจากข้อมูลใน training set ทั้งหมด จึงทำให้ค่าที่ได้เพิ่มขึ้นจาก 1.577 เป็น 1.809 ดังนี้

$-5/14 \times \log_2(5/14)$ (for sunny)

$-3/14 \times \log_2(3/14)$ (for overcast)

$-5/14 \times \log_2(5/14)$ (for rain)

$-1/14 \times \log_2(1/14)$ (for "?")

และค่า gain ratio ลดลงจาก 0.156 เป็น 0.110

2. การแบ่ง Training Set เมื่อข้อมูลใน Training Set ทั้ง 14 record ถูกแบ่งออกโดยใช้ค่าใน Attribute outlook record ที่มีค่าใน Attribute outlook เป็นค่าที่ไม่ทราบค่า จะถูกกำหนดให้ในทุก subset คือ sunny, overcast และ rain ด้วยค่า weight เท่ากับ 5/13, 3/13 และ 5/13 ตามลำดับ พิจารณาที่ subset แรกดังนี้

| Outlook | Temp(°F) | Humidity (%) | Windy | Class | Weight |
|---|---|---|---|---|---|
| Sunny | 75 | 70 | True | Play | 1 |
| Sunny | 80 | 90 | True | Don't play | 1 |
| Sunny | 85 | 85 | False | Don't play | 1 |
| Sunny | 72 | 95 | False | Don't play | 1 |
| Sunny | 69 | 70 | False | Play | 1 |
| ? | 72 | 90 | True | Play | 5/13 |

**ตารางที่ 3.3** แสดง subset ของ outlook = sunny

ถ้า subset นี้ถูกแบ่งต่อไปโดยใช้ Attribute humidity การกระจายของ class ใน subset จะเป็นดังนี้

Humidity $\leq$ 75 2 class Play, 0 class Don't Play

Humidity > 75 5/13 class Play, 3 class Don't Play

Decision Tree ที่ได้จะมีโครงสร้างดังนี้

Outlook = sunny

| Humidity $\leq$ 75:Play(2.0)

| Humidity > 75:Don't Play (3.4/0.4)

Outlook = overcast: Play (3.2)

Outlook = rain

| Windy = true: Don't play (2.4/0.4)
| Windy = false: Play (3.0)

โดยค่าของตัวเลขที่ leaf node จะอยู่ในรูป (N) หรือ (N/E) N เป็นจำนวนข้อมูลทั้งหมดที่มาถึง leaf node นั้น ๆ และ E เป็นจำนวนข้อมูลที่ไม่อยู่ใน class ที่ระบุไว้ เช่น Don't Play (3.4/0.4) หมายความว่า จำนวนข้อมูลที่มาถึงที่ leaf node นี้เท่ากับ 3.4 และ 0.4 ในจำนวนนี้ไม่อยู่ใน class Don't Play

3.การใช้ decision tree ที่ได้ มาทำนายกลุ่มของข้อมูล สมมุติข้อมูลคือ Sunny outlook, temperature $70^{\circ}$, unknown humidity, windy false
จากค่าใน outlook พบว่าต้อง move ไปยัง subset แรกแต่เนื่องจากไม่สามารถตรวจสอบค่า humidity ได้ จึงทำการพิจารณาดังนี้

- ถ้า humidity $\leq$ 75% จะได้ class Play และ
- ถ้า humidity > 75% จะได้ class Don't Play ด้วย ความน่าจะเป็นเท่ากับ 3/3.4(88%) และ class Play ด้วยความน่าจะเป็นเท่ากับ 0.4/3.4(12%)

จะพบว่าการกระจายของ class สุดท้ายสำหรับข้อมูลนี้เท่ากับ

Play: 2.0/5.4×100% +3.4/5.4 × 12% = 44%

Don't Play 3.4/5.4 ×88% = 56%

● **Continuous attribute values** สมมติว่า A เป็น Attribute ชนิด continuous numeric value การทดสอบค่าที่ Attribute นี้จะแบ่งเป็น A $\leq$ Z และ A > Z โดยทำการเปรียบเทียบค่าของ A กับค่า Threshold value Z โดยการหาค่า Threshold ที่เหมาะสมมีขั้นตอนดังนี้

1. เรียงลำดับ Training Set ด้วยค่าใน Attribute A จากน้อยไปมาก และเลือกเฉพาะค่าไม่ซ้ำกันมาพิจารณาจะได้ {v1,v2,…,vn}

2. หาค่า Threshold ใด ๆ ซึ่งค่า Threshold ใด ๆ จะอยู่ระหว่าง Vi และ Vi+1 โดยคำนวณจาก Midpoint ของแต่ละช่วงดังนี้ Vi+Vi+1/2 โดย C4.5 จะเลือกค่าที่มากที่สุดใน Attribute A แต่ต้องไม่เกินค่า Midpoint นั้น ๆ จาก Training Set เป็นค่า Threshold ของแต่ละช่วง เพื่อที่ว่าค่า Threshold ทั้งหมดที่ปรากฏอยู่ใน Tree หรือ Rule จะเป็นค่าที่เกิดขึ้นจริงในข้อมูล

3. หาค่า Threshold ที่เหมาะสม โดยพิจารณาจากค่า Threshold ที่มีค่า Information Gain สูงสุด

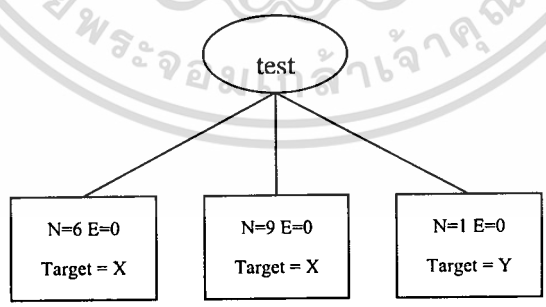
● **Pruning decision tree** การแบ่งข้อมูลใน Training set เพื่อสร้างดิซิชั่นทรีจะทำไปจนกระทั่งข้อมูลในแต่ละ Subset อยู่ใน Class เดียวกัน ซึ่งผลลัพธ์ที่ได้อาจทำให้ Tree มีความซับซ้อนมากเกินไป ที่เรียกว่า "Overfits the data" ซึ่งปัญหานี้สามารถทำการแก้ไขได้โดยทำการ Pruning จะทำให้แต่ละ Leaf node ที่ได้ไม่จำเป็นที่จะต้องประกอบด้วยข้อมูลที่อยู่ใน class เดียวกันทั้งหมด โดยแต่ละ Leaf node จะมีการระบุการกระจายของข้อมูลแต่ละ Class ไว้ ซึ่งจะบอกถึงความน่าจะเป็นที่ข้อมูลจะอยู่ใน Class นั้น ๆ อัลกอริทึมของ C4.5 จะทำการ Pruning โดยการตัด Subtree ที่ทำให้เกิดความผิดพลาดในการทำนายออกไป แล้วทำการแทนที่ Subtree นั้นด้วย Leaf node โดยเทคนิคนี้จะใช้เพียงข้อมูลใน Training set ที่ใช้ในการสร้าง tree เท่านั้น และการคำนวณความผิดพลาดที่เกิดจากการทำนาย ของแต่ละ Leaf node และ Subtree จะทำโดยสมมติว่าจะทำการแบ่งกลุ่ม set ของข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Training set โดยการคำนวณจะใช้ Function ทางสถิติ ซึ่งอยู่บนพื้นฐานของการกระจายแบบ Binomial จำนวน Error ที่เกิดขึ้นเมื่อข้อมูลมีขนาดเท่ากับ N

$$= N \times U_{CF}(E,N)$$

โดย N แทน ขนาดของข้อมูลที่ Leaf node ใด ๆ

E แทน จำนวนของ Error ที่เกิดขึ้นใน Set ของข้อมูลที่ Leaf Node ใด ๆ

$U_{CF}(E,N)$ แทนความน่าจะเป็นสูงสุดที่จะเกิด Error และ C4.5 ใช้ Confidence level เท่ากับ 0.25 หรือ 25% ต่อไปจะอธิบายการ Pruning โดยพิจารณา subtree ดังรูป




**รูปที่ 3.2** Subtree ก่อนทำการ Pruning

จากรูปจะพบว่าค่าที่เป็นไปได้ที่เกิดจากการทดสอบมี 3 ค่าคือ A, B และ C และ Target attribute มี 2 ค่าคือ X และ Y ซึ่งในกรณีนี้ไม่พบ error ที่เกิดขึ้นใน Training set ใน leaf node ที่1 พบว่า N = 6 และ E = 0 ดังนั้น

$$U_{25\%}(0,6) = 0.206$$

ถ้าเราใช้ Leaf node นี้ใน การแบ่งข้อมูลจำนวน 6 record จำนวน Error ที่เกิดขึ้นในการทำนายจะเท่ากับ 6 x 0.206 สำหรับ Leaf node ที่ 2 และ 3 จะได้ $U_{25\%}(0,9) = 0.143$ และ $U_{25\%}(0,1) = 0.750$ ตามลำดับ ดังนั้นจำนวน Error ที่เกิดจากการทำนายของ Subtree นี้เท่ากับ

$$6 \times 0.206 + 9 \times 0.143 + 1 \times 0.750 = 3.273$$

ถ้าทำการแทนที่ Subtree นี้ด้วย Leaf node ที่มี Target = X เมื่อ X เป็นค่าที่มีความถี่มากที่สุดของ Target Attribute ของ Training subset จำนวน 16 record จะเกิด Error 1 record และจำนวน Error ที่เกิดจากการทำนายเท่ากับ

$$16 \times U_{25\%}(1,16) = 16 \times 0.157 = 2.512$$

จะพบว่า subtree นี้มีจำนวนของ Error ที่เกิดจากการทำนายสูงกว่า ดังนั้นจึงทำการ Pruning โดยแทนที่ด้วย Leaf node

# บทที่ 4
# การประยุกต์ใช้ดาต้าไมนิ่งเพื่อช่วยในการจัดแบ่งกลุ่มเครดิตลูกค้า

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงกำหนดขั้นตอนในการศึกษาโดยอิงตามกระบวนการทำงานของ Data Mining ซึ่งมีขั้นตอนดังนี้

## 4.1 กำหนดวัตถุประสงค์

ในการตัดสินใจอนุมัติสินเชื่อในธุรกิจการให้บริการสินเชื่อนั้น จำเป็นต้องพิจารณาความเสี่ยงที่จะเกิดขึ้นจากการอนุมัติสินเชื่อด้วย เพราะถ้าลูกค้าไม่สามารถชำระเงินคืนได้ จะทำให้ธนาคารสูญเสียกำไร และความสามารถในการให้บริการแก่ลูกค้ารายใหม่ อาจส่งผลให้ธนาคารไม่สามารถคงอยู่ในธุรกิจต่อไปได้

จึงมีความคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้ในธุรกิจการธนาคาร โดยมีวัตถุประสงค์เพื่อจำแนกลักษณะ ของลูกค้าทั้งกลุ่มที่มีเครดิตดีและไม่ดี เพื่อนำมาประกอบการตัดสินใจพิจารณาอนุมัติสินเชื่อต่อไป

## 4.2 การคัดเลือกข้อมูล

คัดเลือกข้อมูลโดยคำนึงถึงวัตถุประสงค์ในการนำข้อมูลมาใช้งาน ซึ่งข้อมูลที่ได้อาจได้มาจากข้อมูลหลาย ๆ แหล่ง โดยจะต้องทำให้ข้อมูลเหล่านั้นอยู่ในรูปแบบเดียวกันเสียก่อน ข้อมูลที่นำมาใช้นี้เป็นข้อมูลลูกค้าสินเชื่อเพื่อซื้อห้องชุด โดยได้ทำการคัดเลือกข้อมูลมาดังนี้

| ชื่อข้อมูล | ประเภทข้อมูล |
|---|---|
| สถานภาพสมรส | Text |
| จำนวนบุตร | Number |
| จำนวนบุตรที่ยังไม่มีรายได้ | Number |
| การศึกษา | Text |
| รายได้ต่อเดือน | Number |
| อาชีพ | Text |
| ประสบการณ์ทำงาน | Number |
| อายุ | Number |
| จำนวนผู้กู้ร่วม | Number |
| เงินกู้ | Number |
| ระยะเวลากู้ | Number |
| ค่างวด | Number |
| สถานะบัญชี | Text |

**ตารางที่ 4.1** ตารางข้อมูลลูกค้า

| รหัส | ความหมาย |
|---|---|
| A1 | โสด |
| A2 | หย่า/หม้าย |
| A3 | สมรส |
| A4 | สมรสไม่จดทะเบียน |

**ตารางที่ 4.2** รายการสถานภาพสมรส

| รหัส | ความหมาย |
|---|---|
| B1 | ต่ำกว่าหรือเท่ากับระดับประถม |
| B2 | ตั้งแต่ระดับมัธยมขึ้นไปแต่ไม่ถึงระดับปริญญาตรี |
| B3 | ตั้งแต่ระดับปริญญาตรีขึ้นไป |

**ตารางที่ 4.3** รายการการศึกษา

| รหัส | อาชีพ |
|---|---|
| C1 | รับราชการ |
| C2 | พนักงานรัฐวิสาหกิจ |
| C3 | พนักงานบริษัทหรือลูกจ้าง |
| C4 | อิสระ |

**ตารางที่ 4.4** รายการอาชีพ

| สถานะบัญชี |
|---|
| ปกติ |
| ค้างชำระ |

**ตารางที่ 4.5** สถานะบัญชี

Target Attribute ที่จะใช้ในการสร้างแบบจำลองการพยากรณ์คือ สถานะบัญชี ซึ่งมีค่าที่เป็นไปได้ดังแสดงในตารางที่ 4.5

**4.3 การเตรียมข้อมูล**

เป็นการนำข้อมูลที่ถูกเลือกเข้ามาจากกระบวนการ Data Selection ซึ่งข้อมูลเหล่านี้เราจะต้องนำมาทำความสะอาดเพื่อจัดการกับ Noisy Data คือ ค่าของข้อมูลผิดไปจากค่าที่มันควรจะเป็น แก้ไขโดยค้นหาค่าที่ถูกต้องนำมาแก้ไข และเพื่อจัดการกับ Missing Data แก้ไขโดยทำการตัดข้อมูลนั้นทิ้งไปทั้งรายการ หรือแทนส่วนที่หายไปด้วยค่าเฉลี่ย (Mean) หรือค่าที่ปรากฏบ่อย (Mode)

ถึงแม้ว่าอัลกอริทึมของ Classification ส่วนมากจะมีกลไกในการจัดการกับ Noisy Data และ Missing Data แต่ขั้นตอนนี้จะช่วยลดความไม่ชัดเจนในระหว่างการเรียนรู้ได้ เนื่องจากตัวอย่างข้อมูลที่ได้เลือกมามีความสมบูรณ์ ไม่พบทั้ง Noisy Data และ Missing Data จึงไม่มีการกระทำใด ๆ กับข้อมูล

**4.4 การแปลงข้อมูล (Data Transformation)**

เป็นการปรับเปลี่ยนรูปแบบของข้อมูลให้เหมาะสมกับอัลกอริทึมที่เลือกใช้เช่นแปลงข้อมูลตัวเลขให้เป็นช่วงเพื่อใช้กับ ID3 Algorithm แต่อัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมที่พัฒนาปรับปรุงมาจาก ID3 นั้นมีกลไกในการจัดการกับข้อมูลที่มีลักษณะเป็น Continuous Value

จากนั้นทำการแบ่งข้อมูลออกเป็น 2 ชุด โดยชุดแรกใช้สำหรับทำการฝึกสอนเพื่อสร้างแบบจำลองพยากรณ์ จำนวน 800 รายการ และชุดที่ 2 ใช้สำหรับทดสอบความถูกต้องของแบบจำลองที่ได้จำนวน 200 รายการ

เมื่อได้ทำการจัดเตรียมข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดไปเป็นการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น ได้เลือกใช้ภาษาจาวา (Java Language) ในการพัฒนาโปรแกรม

**4.5 การจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น**

ในส่วนของการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้นประกอบด้วย 3 ขั้นตอน ดังนี้

4.5.1 การฝึกสอนระบบเพื่อสร้างแบบจำลองพยากรณ์ ในส่วนนี้จะประกอบด้วยขั้นตอนย่อยอีก 4 ขั้นตอน คือ

4.5.1.1 การนำข้อมูลเข้าระบบ

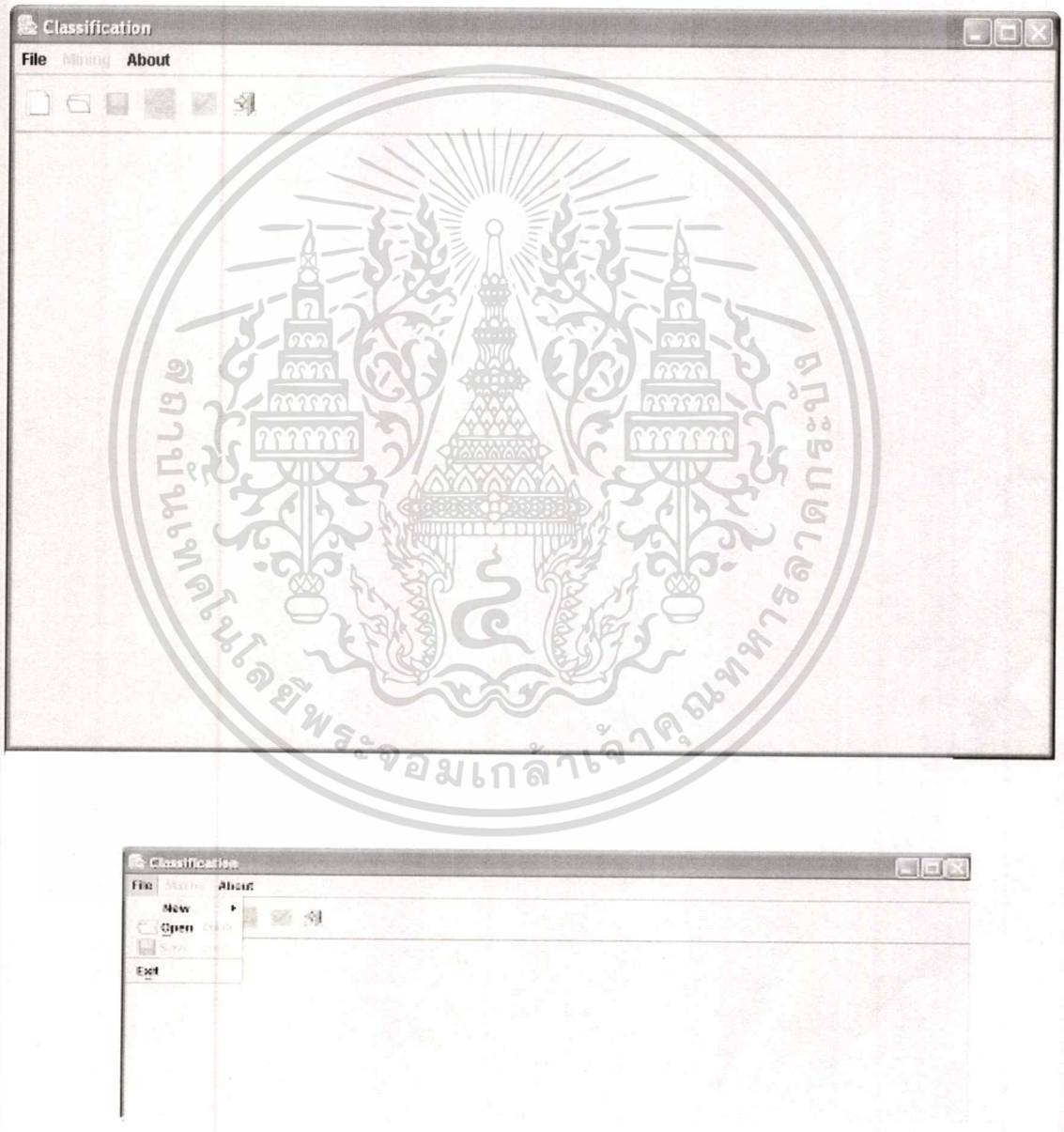
4.5.1.2 การตรวจสอบคุณภาพของข้อมูล

4.5.1.3 การกำหนดเงื่อนไขในการสร้างแบบจำลอง

4.5.1.4 การแสดงผล

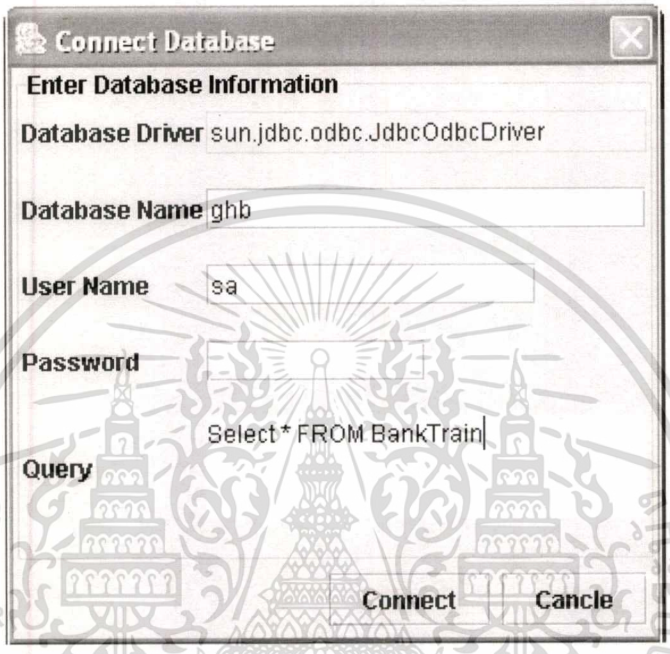
4.5.1.1 การนำข้อมูลเข้าระบบ

ในส่วนของการใช้งานโปรแกรมเพื่อนำข้อมูลเข้าระบบมีขั้นตอนการใช้งานคือ เมื่อเข้าสู่โปรแกรมจะปรากฏเมนูหลัก โดยในส่วนของเมนูหลัก จะประกอบด้วย เมนู File ถ้าต้องการสร้างแบบจำลองใหม่เลือก เมนูย่อย New หรือ ต้องการเปิดแบบจำลองที่มีอยู่แล้วเลือก Open แสดงดังรูปที่ 4.1




**รูปที่ 4.1** หน้าจอหลักและเมนูการทำงานของระบบ

สำหรับกรณีที่ต้องการสร้างแบบจำลองใหม่ให้นำเข้าข้อมูลโดยเลือกที่เมนู New แล้วเลือกเมนูย่อย Connect Database จากนั้นจะปรากฏหน้าจอเพื่อการติดต่อกับฐานข้อมูล ดังรูปที่ 4.2 โดยผู้ใช้ต้องกรอก ชื่อ Database, user name, password และทำการเลือกข้อมูลด้วยคำสั่ง sql



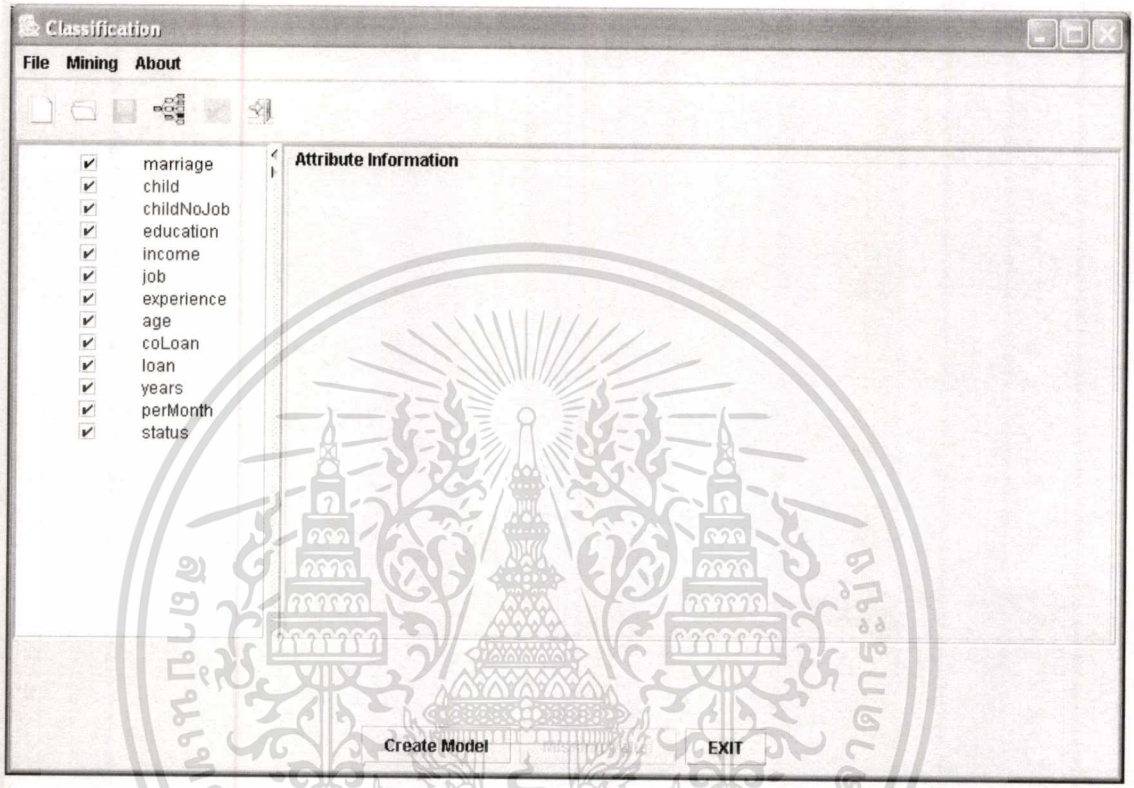

**รูปที่ 4.2** แสดงหน้าจอการติดต่อกับฐานข้อมูล

4.5.1.2 การตรวจสอบคุณภาพของข้อมูล

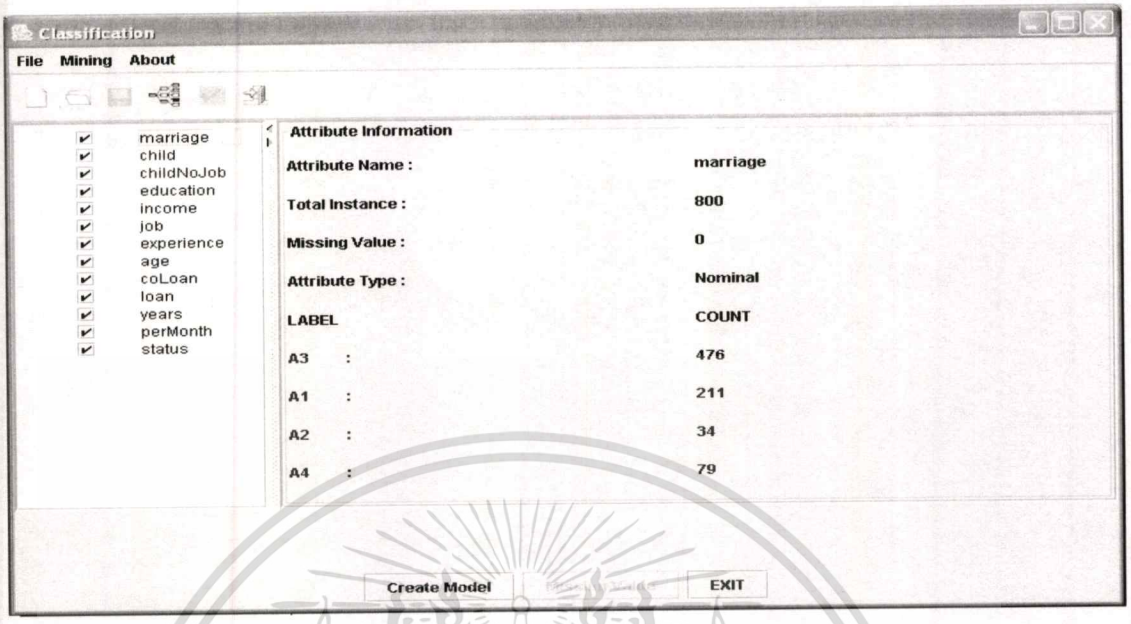
หลังจากเลือกฐานข้อมูลที่จะใช้วิเคราะห์แล้ว จะปรากฏหน้าจอแสดงแอททริบิวต์ทั้งหมดที่เลือกมา แสดงดังรูปที่ 4.3 ซึ่งสามารถเลือกได้ว่าต้องการแอททริบิวต์ใดในการนำไปสร้างแบบจำลอง และถ้าต้องการทราบรายละเอียดของแต่ละแอททริบิวต์สามารถทำได้โดยการเลือกที่ชื่อของแต่ละ แอททริบิวต์ ผลแสดงดังรูปที่ 4.4 ซึ่งรายละเอียดจะประกอบด้วย ชื่อแอททริบิวต์, จำนวนข้อมูล, จำนวนข้อมูลของแอททริบิวต์นี้ที่มีค่าที่หายไป (Missing Value), ประเภทของ แอททริบิวต์ และในกรณีที่เป็นข้อมูลประเภท Nominal จะแสดงค่าที่เป็นไปได้ทั้งหมดและจำนวนของข้อมูลของแต่ละค่าที่เป็นไปได้ แต่ถ้าเป็นข้อมูลประเภทตัวเลข (Numeric) จะแสดงค่าทางสถิติ ซึ่งได้แก่ ค่าต่ำสุด, ค่าสูงสุด, ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของข้อมูล

ในกรณีที่แอททริบิวต์ใดมีค่าของข้อมูลที่ขาดหายไป ก็จะต้องทำการจัดการกับข้อมูลเหล่านี้ก่อนจึงจะทำการสร้างแบบจำลองได้ โดยเลือกปุ่ม Missing Value หลังจากนั้นจะปรากฏหน้าจอ

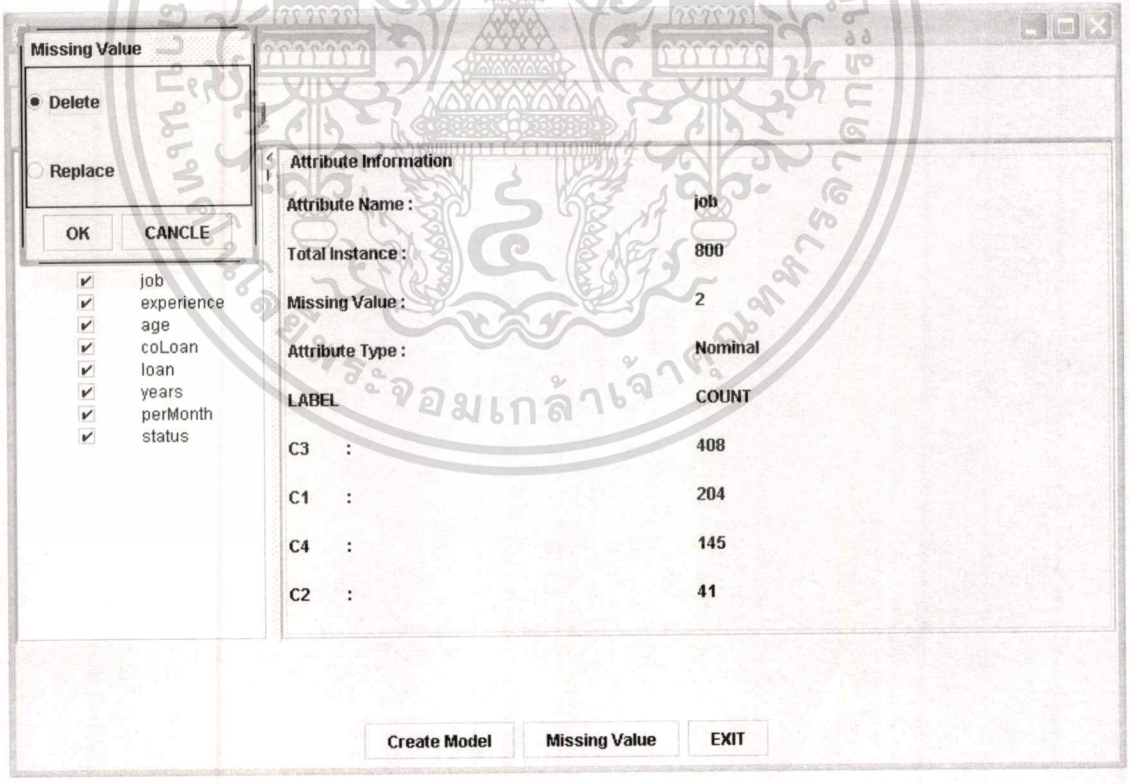
ในการจัดการดังรูปที่ 4.5 โดยสามารถเลือกจัดการได้สองวิธีคือ ลบ record นั้นทิ้ง หรือแทนค่าที่ขาดหายไปด้วยค่าเฉลี่ยของแอททริบิวต์นั้น




**รูปที่ 4.3** หน้าจอแสดง Attribute ทั้งหมด



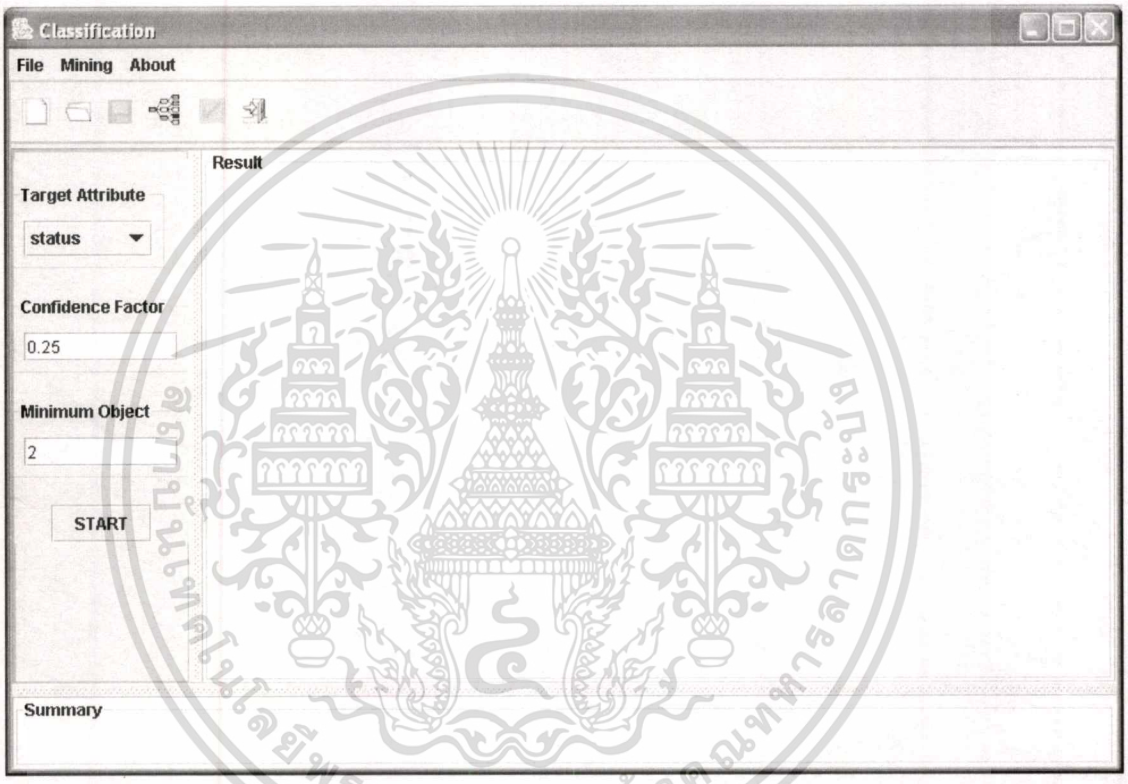

**รูปที่ 4.4** หน้าจอแสดงรายละเอียดของแต่ละ Attribute




**รูปที่ 4.5** หน้าจอแสดงการจัดการกับ Missing Value

4.5.1.3 การกำหนดเงื่อนไขในการสร้างแบบจำลอง

หลังจากที่ทำการตรวจสอบคุณภาพของข้อมูลเสร็จแล้ว จะเข้าสู่ขั้นตอนของการกำหนดแอททริบิวต์เป้าหมาย (target Attribute), ค่า Confidence Factor และ Minimum Example เพื่อกำหนดเงื่อนไขในการสร้างดิซิชั่นทรีและกฎ แสดงดังรูปที่ 4.6 หลังจากนั้นทำการสร้างแบบจำลองโดยการเลือกปุ่ม START

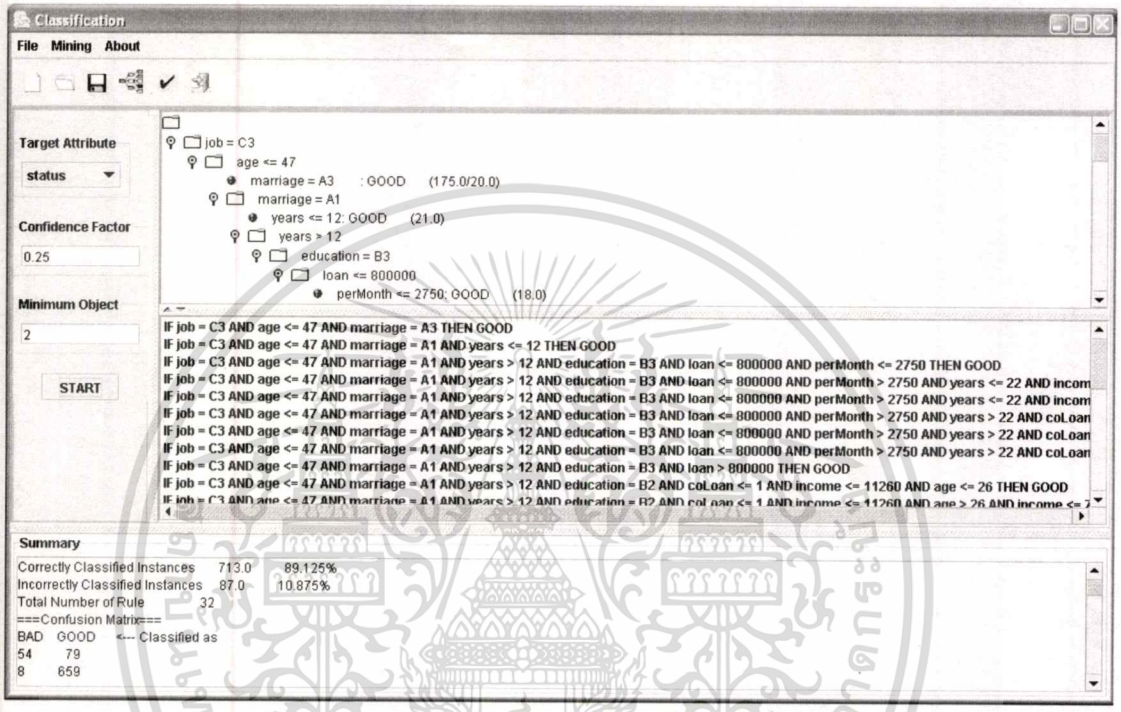



**รูปที่ 4.6** หน้าจอแสดงการกำหนดเงื่อนไขในการสร้างดิซิชั่นทรี

4.5.1.4 การแสดงผล

เมื่อโปรแกรมทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงดังรูปที่ 4.7 โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใดและข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภท (Class). ที่ข้อมูลส่วนใหญ่ในโหนดนั้นตกอยู่ และมีสรุปผลของการสร้างแบบจำลองอยู่ด้านล่างในส่วนของ Summary ซึ่งประกอบด้วยจำนวนข้อมูลที่สามารถจำแนกได้อย่างถูกต้อง, จำนวนข้อมูลที่จำแนกผิดพลาด, จำนวนกฎที่ได้ และ Confusion Matrix ซึ่งแสดงถึงการกระจายของ class ที่สามารถทำนายได้ถูกต้องตามความเป็น

จริง และที่ไม่ตรงตามความเป็นจริง อีกทั้งยังสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้ โดยการเลือก เมนูย่อย Save หรือปุ่มสัญลักษณ์ Save

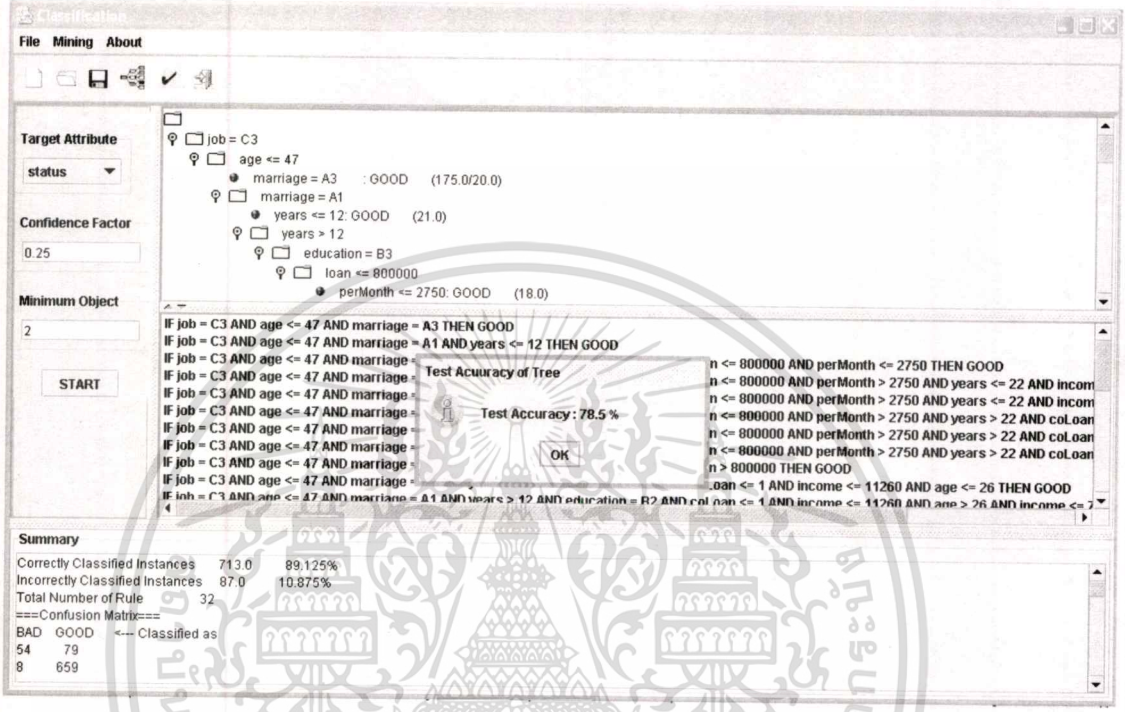



**รูปที่ 4.7** หน้าจอแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ

4.5.2 การทดสอบความถูกต้องของแบบจำลองพยากรณ์

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูลที่ใช้ฝึกสอนแล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากเพียงใด โดยการนำข้อมูลอีกชุดหนึ่งมาทำการทดสอบกับแบบจำลองพยากรณ์ที่ได้ โดยเลือกเมนู Mining และ เมนูย่อย Test Model หลังจากนั้นจะมีหน้าจอสำหรับการติดต่อกับฐานข้อมูลแสดงขึ้นมา ซึ่งมีลักษณะเดียวกันกับการ

ติดต่อฐานข้อมูลเพื่อนำข้อมูลมาสร้างแบบจำลอง ซึ่งระบบจะแสดงหน้าจอไดอะล็อกบ๊อกซ์ของค่าความถูกต้อง (Accuracy) ของโมเดลนั้น ๆ ดังแสดงในรูปที่ 4.8

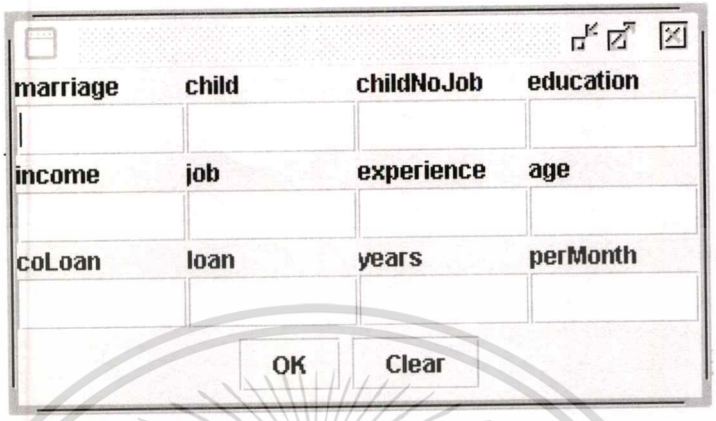



**รูปที่ 4.8** หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

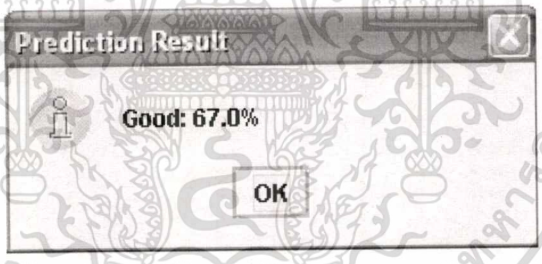
4.5.3 การนำแบบจำลองพยากรณ์มาใช้จัดกลุ่มข้อมูล

เมื่อทำการสร้างแบบจำลองและทดสอบความถูกต้องจนได้ผลอยู่ในระดับที่พึงพอใจแล้ว ผู้ใช้สามารถสอบถามเกี่ยวกับข้อมูลของผู้ใช้ว่าจัดอยู่ในกลุ่มใดได้ โดยเลือกเมนูย่อย Predict Class ในหน้าจอหลักของระบบ จากนั้นจะปรากฏหน้าต่างให้ใส่ข้อมูลในแต่ละแอททริบิวต์ดังรูปที่ 4.9

ในกรณีที่ผู้ใช้ไม่ทราบค่าในแอททริบิวต์ใดก็สามารถเว้นว่างไว้ได้ ระบบจะแสดงผลการทำนายว่าข้อมูลมีแนวโน้มที่จะอยู่ในกลุ่มใด โดยแสดงออกมาในรูปของเปอร์เซ็นต์ ดังแสดงในรูปที่ 4.10




**รูปที่ 4.9** แสดงหน้าจอสำหรับใส่ข้อมูลเพื่อสอบถามกลุ่มของข้อมูล




**รูปที่ 4.10** หน้าจอแสดงผลการทำนาย

### 4.6 สรุปผลการดำเนินงาน

ผลจากการทดสอบแบบจำลองพยากรณ์ที่ได้โดยใช้ข้อมูลทดสอบจำนวน 200 รายการ สามารถวัดความถูกต้องของแบบจำลองการจัดหมวดหมู่ สำหรับการทำนายข้อมูลในแต่ละกลุ่ม โดยสรุปรายการที่แบบจำลองสามารถทำนายได้ถูกต้องคิดเป็น 78.5 % ของข้อมูลทั้งหมด ซึ่งความผิดพลาดจำนวน 21.5% นี้อาจเกิดจากข้อมูลที่ใช้สร้างแบบจำลองมีจำนวนไม่มากพอหรืออาจเกิดจากการกระจายของข้อมูลที่นำมาทดสอบอาจจะมีการเกาะกลุ่มในบางกรณีมากเกินไป ถ้าข้อมูลมากขึ้นก็ความถูกต้องก็จะมากขึ้นด้วย ซึ่งค่าความถูกต้องของแบบจำลองนั้นขึ้นอยู่กับปัจจัยต่าง ๆ

มากมาย หากเลือกและกำหนดค่าต่าง ๆ อย่างเหมาะสมแล้ว ก็จะได้แบบจำลองที่มีค่าความถูกต้องที่ยอมรับได้

# บทที่ 5
# สรุปผลการศึกษาและข้อเสนอแนะ

## 5.1 สรุปผลการดำเนินงาน

โครงการพัฒนาระบบนี้มีวัตถุประสงค์หลักเพื่อที่จะนำเสนอและประยุกต์ใช้ดาต้าไมนิ่งในธุรกิจ ซึ่งดาต้าไมนิ่งนั้นเป็นกระบวนการที่ใช้เพื่อค้นหาความรู้ (Knowledge) ที่สนใจจากฐานข้อมูลเพื่อนำมาช่วยในการตัดสินใจ ซึ่งวิธีการแก้ปัญหาด้วยดาต้าไมนิ่งนั้นมีหลายรูปแบบขึ้นอยู่กับวัตถุประสงค์ของการใช้งาน โดยในโครงงานนี้ได้เสนอเทคนิคการจัดหมวดหมู่ข้อมูล (Classification) เพื่อจำแนกเครดิตของลูกค้าที่ทำการขอสินเชื่อจากธนาคาร โดยมีวัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อให้แก่ลูกค้าเพื่อลดความเสี่ยงขององค์กร รวมทั้งเป็นแนวทางในการนำไปประยุกต์เพื่อใช้นำไปพิจารณาปัจจัยอื่น ๆ ที่มีผลต่อการดำเนินธุรกิจ โดยใช้อัลกอริทึม C4.5 ซึ่งมีประสิทธิภาพในการแก้ปัญหา, มีความยืดหยุ่นและกฎที่ได้สามารถเข้าใจได้ง่าย

ผลจากการศึกษาทำให้ได้ระบบที่ใช้สำหรับจัดแบ่งกลุ่มของข้อมูล ซึ่งสามารถนำไปประยุกต์ใช้กับธุรกิจอื่นได้ และจากการนำข้อมูลเข้าไปสร้างแบบจำลองพยากรณ์และทำการทดสอบพบว่า แบบจำลองพยากรณ์ที่ได้มีความถูกต้องคิดเป็น 78.5 % ของข้อมูลทั้งหมด ซึ่งความผิดพลาดจำนวน 21.5% นี้อาจเกิดจากข้อมูลที่ใช้สร้างแบบจำลองมีจำนวนไม่มากพอหรืออาจเกิดจากการกระจายของข้อมูลที่นำมาทดสอบอาจจะมีการเกาะกลุ่มในบางกรณีมากเกินไป

จากอัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมของ Classificaton พบว่ากฎที่ได้สามารถเข้าใจได้ง่าย และสามารถบอกได้ว่าปัจจัยใดที่มีอิทธิพลต่อการทำนายมากที่สุด แต่วิธีนี้ก็มีข้อเสียคือ ถ้าข้อมูลมีจำนวนน้อย ความผิดพลาดในการทำนายจะสูงขึ้น เนื่องจากเมื่อทำการแตก Tree ไปเรื่อย ๆ จำนวนข้อมูลจะลดลง ยิ่ง level ของ tree มากขึ้นความน่าเชื่อถือจะยิ่งน้อยลง

## 5.2 ข้อเสนอแนะ

ระบบที่พัฒนาขึ้นนี้สามารถนำไปใช้กับข้อมูลในธุรกิจอื่นได้ เนื่องจากไม่จำกัดขอบเขตกับธุรกิจธนาคารเท่านั้น สามารถนำไปพัฒนาให้มีประสิทธิภาพ และตรงกับความต้องการเฉพาะด้านทางธุรกิจมากขึ้น หรือทำการพัฒนาระบบเป็น เว็บแอพพลิเคชั่น เพื่อความสะดวกและความยืดหยุ่น

ในการใช้งานมากยิ่งขึ้น หรือพัฒนาระบบให้สามารถแปลงผลการทำนายให้อยู่ในรูปคำสั่ง sql เพื่อดึงข้อมูลจากฐานข้อมูลได้เลย โดยผู้ใช้ไม่ต้องแปลงเป็นคำสั่ง SQL เอง

# บรรณานุกรม


Bill Wilson. 2003. **Induction of Decision Trees.** [Online] Available: http://www.cse.unsw.edu.au/~billw/cs9414/notes/ml/06prop/id3/id3.html

Howard Hamilton et al. 2002. **Knowledge Discovery in Databases.** [Online] Available: http://www2.cs.uregina.ca/~hamilton/courses/831/index.html

J R Quinlan. 1993. **C4.5: Programs for Machine Learning.** Morgan Kaufmann: San Mateo. CA.

James P. Ignizio.1991. **Introduction to Expert Systems: The Development and Implementation of Rule-Based Expert System.** McGraw-Hill. Singapore.

Karuna Pande Joshi. 1997. **Analysis of Data Mining Algorithms.** [Online] Available: http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm

Peter Cabena et al. 1998. **Discovering Data Mining: From Concept to Implementation.** New Jersey. Prentice Hall PTR.

# ภาคผนวก ก

# คู่มือการใช้ระบบดาต้าไมนิ่งในการจำแนกกลุ่มเครดิตลูกค้าโดยใช้ดิซิชั่นทรี

**ก.1 ความต้องการของระบบ**

1) ความต้องการทางด้านซอฟต์แวร์

ซอฟต์แวร์ที่ต้องการใช้ มีดังนี้

(1.1) ระบบปฏิบัติการวินโดวส์ 95 ขึ้นไป

(1.2) J2SDK Version 1.4.0 ขึ้นไป

(1.3) โปรแกรม Microsoft SQL Server Version 7.0

2) ความต้องการทางด้านฮาร์ดแวร์

ความต้องการทางด้านฮาร์ดแวร์ มีดังนี้

(2.1) เครื่องคอมพิวเตอร์ที่มีโปรเซสเซอร์เพนเทียมทรีขึ้นไป

(2.2) เนื้อที่ว่างบนฮาร์ดดิสก์อย่างน้อย 2 เมกะไบต์

(2.3) หน่วยความจำสำรอง อย่างน้อย 128 เมกะไบต์

(2.4) จอภาพชนิด Super VGA ความละเอียดของจอภาพอย่างน้อย 256 สี

**ก.2 การทำงานของระบบจัดแบ่งกลุ่มเครดิตลูกค้าโดยใช้อัลกอริทึม C4.5**

เมื่อเริ่มต้นทำงาน โปรแกรมจะแสดงหน้าจอแรกดังภาพที่ ก.1 โดยประกอบด้วยเมนู (Menu) ดังนี้

1) เมนู File

ประกอบด้วยเมนูย่อย 4 เมนู คือ

(1.1) เมนู New ใช้สำหรับเริ่มการสร้างแบบจำลองใหม่โดยจะมีเมนูย่อยคือ Connect Database เพื่อทำการติดต่อกับฐานข้อมูลที่จะใช้สร้างแบบจำลอง

(1.2) เมนู Open ใช้สำหรับการเปิดแบบจำลองที่มีอยู่แล้ว

(1.3) เมนู Save ใช้สำหรับการเก็บบันทึกแบบจำลองที่สร้างไว้

(1.4) เมนู Exit ใช้สำหรับการออกจากโปรแกรม

2) เมนู Mining

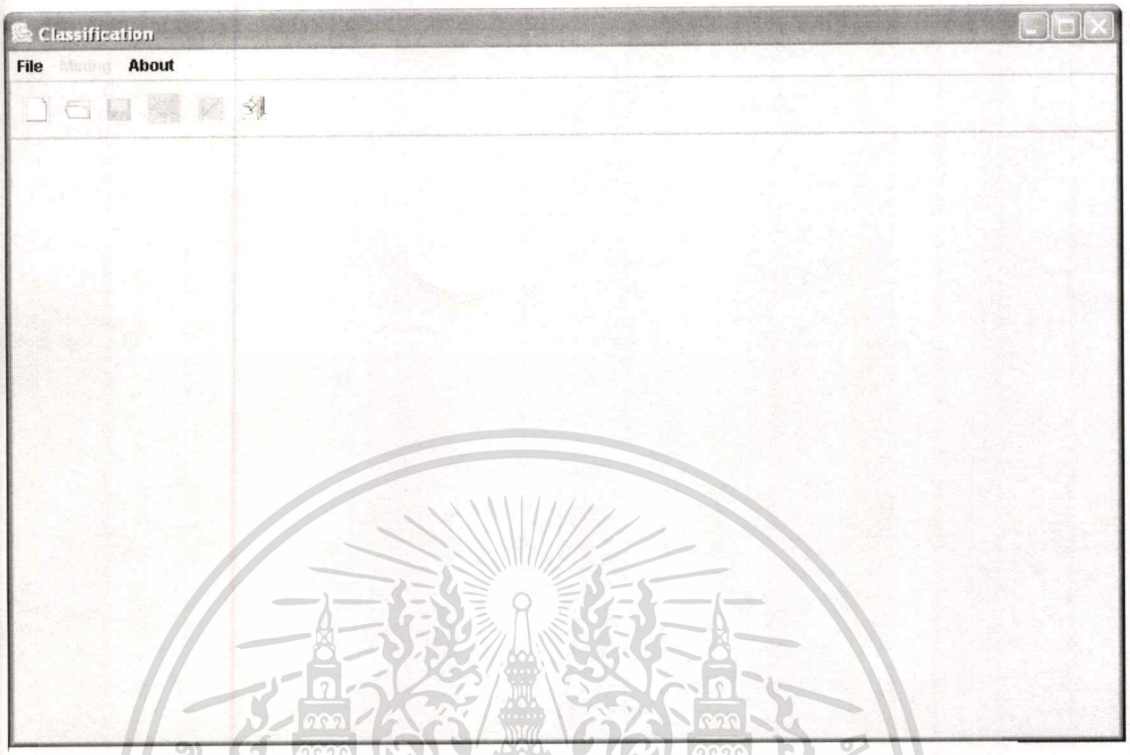
ประกอบด้วย เมนูย่อย

(2.1) เมนู Create Model ใช้เพื่อทำการสร้างแบบจำลอง

(2.2) เมนู Test Model ใช้เพื่อทำการทดสอบแบบจำลอง

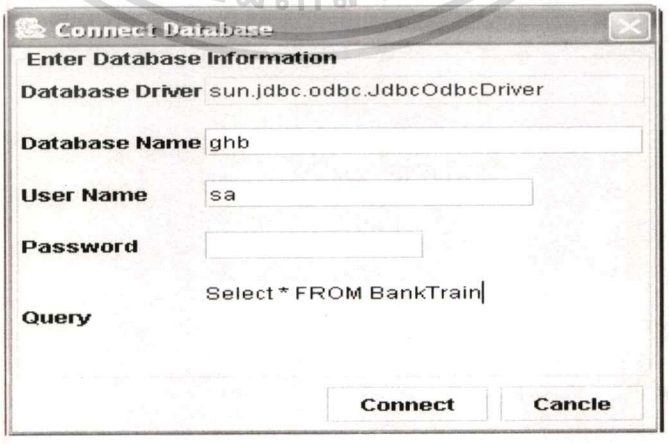
(2.3) เมนู Predict Class ใช้เพื่อทำการทำนายกลุ่มของข้อมูล

3) เมนู About




**รูปที่ ก.1** หน้าจอแรกของระบบ

ในการนำข้อมูลจากฐานข้อมูลเข้าสู่โปรแกรมเพื่อสร้างแบบจำลอง นั้น ทำได้โดยการเลือก เมนู New และเมนูย่อย Connect Database จากหน้าจอแรกของระบบ แสดงดังรูปที่ ก.1 ซึ่งจะแสดงไดอะล็อกซ์บ็อกเพื่อให้ผู้ใช้ป้อนข้อมูล ชื่อฐานข้อมูล, user name, password ซึ่งแสดงดังรูปที่ ก.2

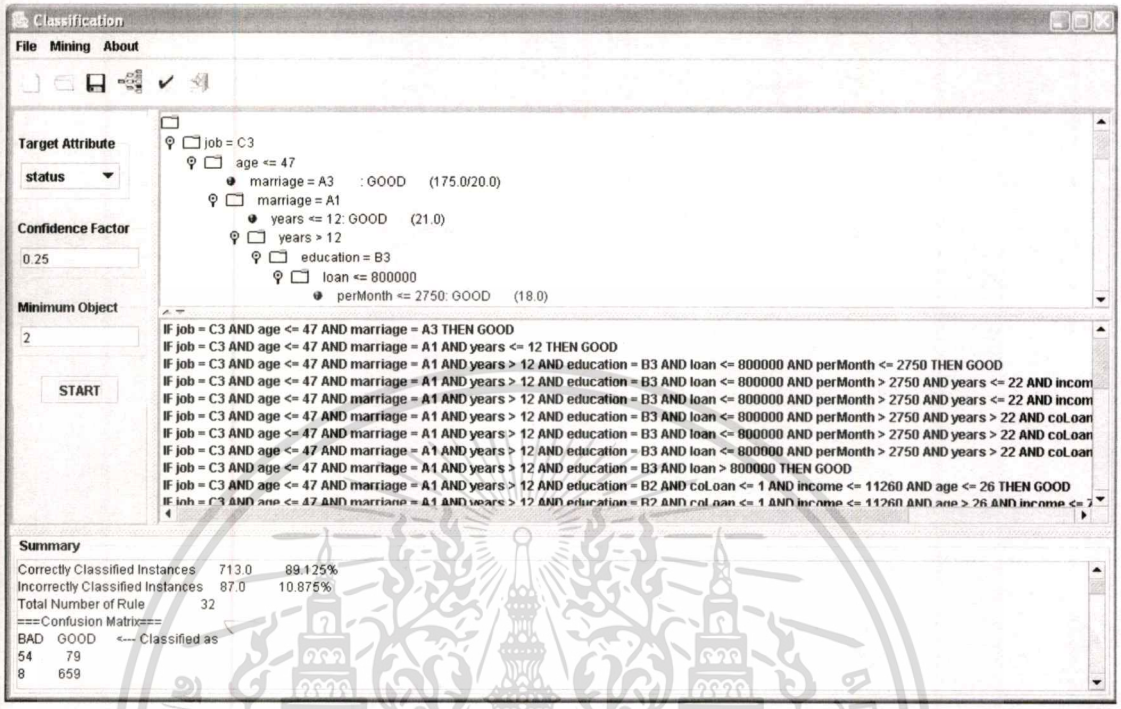



**รูปที่ ก.2** หน้าจอติดต่อกับฐานข้อมูล

หลังจากนั้นถ้ากรณีที่ข้อมูลมีค่าที่ขาดหายไป (Missing Value) ก็จัดการได้โดยเลือกปุ่ม Missing Value แต่ถ้าไม่มีก็ทำการสร้างแบบจำลองได้โดยเลือกปุ่ม Create Model หลังจากนั้นจะแสดงหน้าจอเพื่อให้ทำการกำหนด Target Attribute, Confidence Factor และ Minimum Example และให้เลือกปุ่ม Start เพื่อให้ระบบสร้างดิซิชั่นทรีและกฎ แสดงดังรูปที่ ก.3

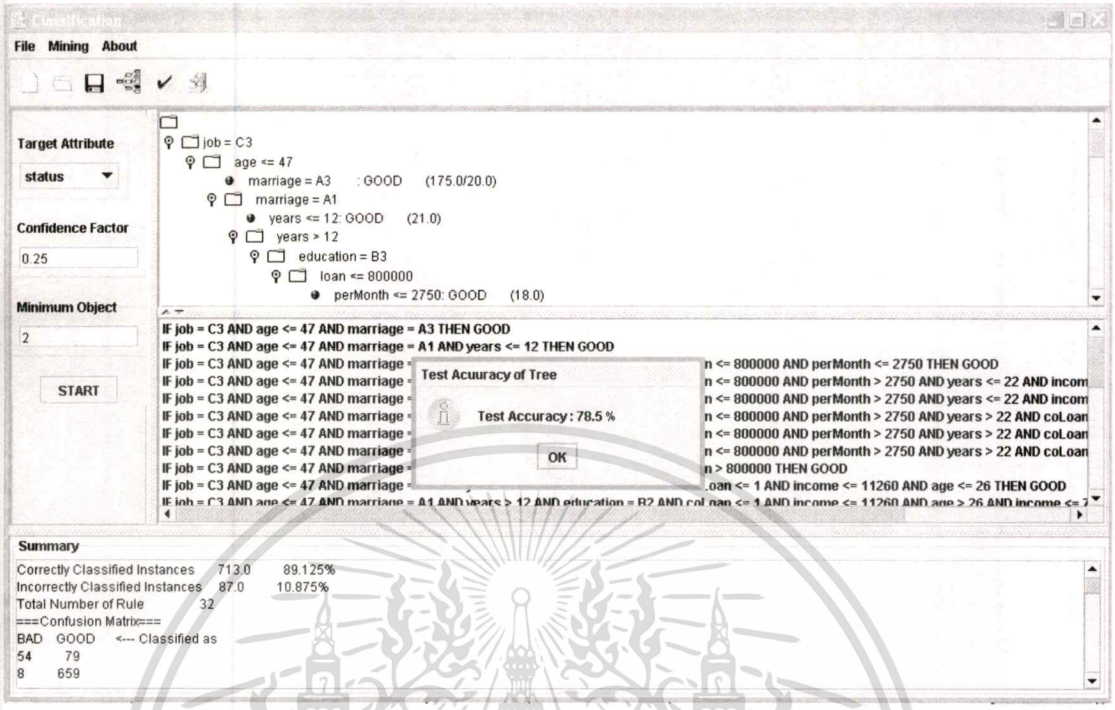

**รูปที่ ก.3** หน้าจอการกำหนดเงื่อนไขในการสร้างแบบจำลอง

เมื่อโปรแกรมทำการสร้างแบบจำลองเรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงในรูปที่ ก.4 โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใดและข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภท (Class) ที่ข้อมูลส่วนใหญ่ในโหนดนั้นตกอยู่ รวมทั้งสรุปว่าแบบจำลองนี้สามารถแบ่งแยกข้อมูลชุดฝึกสอนได้ถูกต้องเท่าไหร่ โดยสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้ โดยการเลือกปุ่ม Save



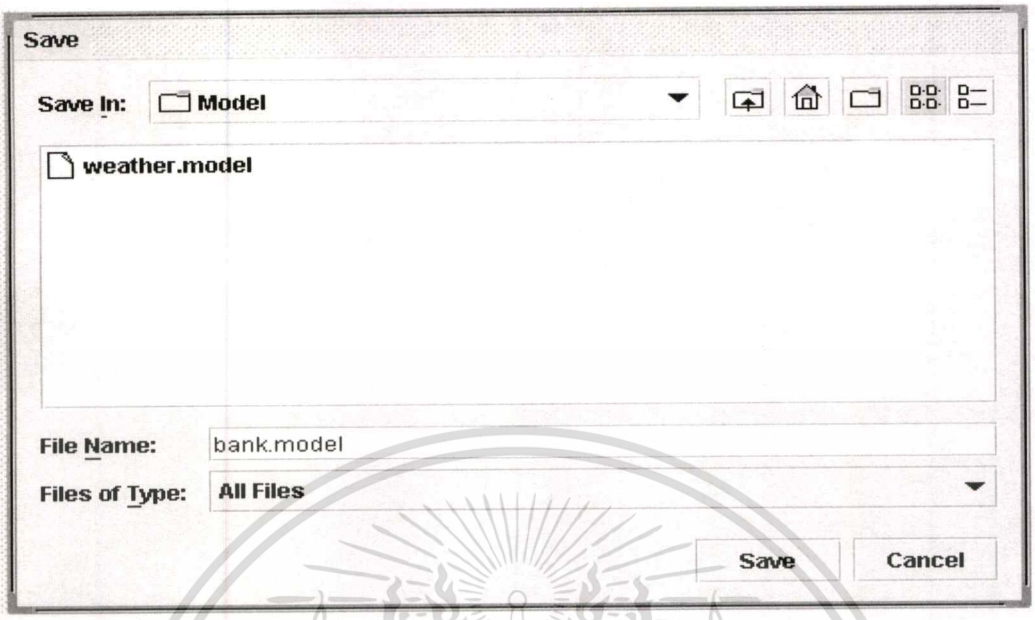

**รูปที่ ก.4** หน้าจอแสดงผลลัพธ์ในรูปดิซิชั่นทรีและกฎ

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูลที่ใช้ฝึกสอนแล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากเพียงใด โดยการนำข้อมูลอีกชุดหนึ่งมาทำการทดสอบกับแบบจำลองพยากรณ์ที่ได้ โดยเลือกเมนูย่อย Test Model เพื่อให้ระบบแสดงความถูกต้องของแบบจำลอง โดยใช้ข้อมูลชุดทดสอบ โดยระบบจะให้ดึงข้อมูลจากฐานข้อมูล และระบบจะแสดงผลการทดสอบ ดังรูปที่ ก.5

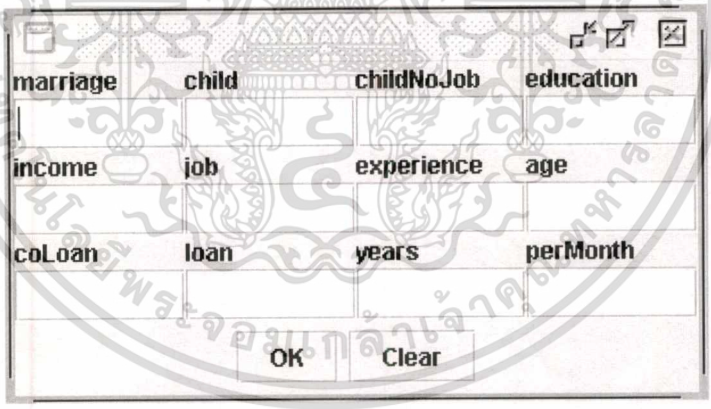



**รูปที่ ก.5** หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

เมื่อทำการสร้างแบบจำลองและทดสอบความถูกต้องจนได้ผลอยู่ในระดับที่พึงพอใจแล้ว ผู้ใช้สามารถเก็บบันทึกแบบจำลองนั้นได้โดยการเลือกปุ่ม Save โดยชนิดที่จะทำการบันทึกคือ " .model " แสดงดังรูปที่ ก.6 อีกทั้งผู้ใช้สามารถสอบถามเกี่ยวกับข้อมูลของผู้ใช้ว่าจัดอยู่ในกลุ่มใดได้ โดยเลือกเมนูย่อย Predict Class ในหน้าจอหลักของระบบ จากนั้นจะปรากฏหน้าต่างให้ใส่ข้อมูลในแต่ละแอททริบิวต์ดังรูปที่ ก.7

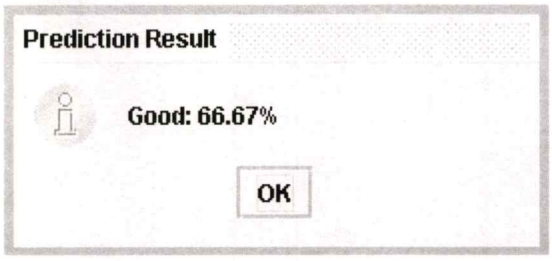



**รูปที่ ก.6** แสดงหน้าจอสำหรับการบันทึกแบบจำลอง




**รูปที่ ก.7** แสดงหน้าจอสำหรับใส่ข้อมูลในแต่ละแอททริบิวต์

โดยผู้ใช้ไม่จำเป็นต้องใส่ข้อมูลทุกแอททริบิวต์ และในกรณีที่ผู้ใช้ไม่ทราบค่าในแอททริบิวต์ใดก็สามารถเว้นว่างไว้ได้ ระบบจะแสดงผลการทำนายว่าข้อมูลมีแนวโน้มที่จะอยู่ในกลุ่มใด โดยแสดงออกมาในรูปของเปอร์เซ็นต์ แสดงดังรูปที่ ก.8




**รูปที่ ก.8** หน้าจอแสดงผลการทำนาย

และผู้ใช้สามารถนำแบบจำลองที่มีอยู่แล้วมาใช้ในการทำนายกลุ่มข้อมูลได้โดยการเลือกเมนูย่อย Open หลังจากนั้นจะแสดงหน้าจอสำหรับการเปิดไฟล์ขึ้นมาซึ่งไฟล์ที่ใช้มีชนิดเป็น ".model" แสดงดังรูปที่ ก.9


**รูปที่ ก.9** แสดงหน้าจอสำหรับการเปิดแบบจำลอง

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ นามสกุล** | นางสาว มณฑิรา ไวพ่อค้า |
| **วัน เดือน ปีเกิด** | 4 ธันวาคม พ.ศ. 2522 |
| **สถานที่เกิด** | จังหวัดกรุงเทพมหานคร |
| **วุฒิการศึกษา** | ปริญญาตรีวิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์ประยุกต์ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ |