การพัฒนาระบบงานเพื่ออนุมัติสินเชื่อเบื้องต้นโดยใช้อัลกอริทึม C4.5

C4.5 Classification Tree for Credit Approval

โดย

นางสาวนฤมล สมบูรณ์เงิน

รหัส 43067053

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2544
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อ	การพัฒนาระบบงานเพื่ออนุมัติสินเชื่อเบื้องต้นโดยใช้อัลกอริทึม C4.5
นักศึกษา	นางสาวนฤมล สมบูรณ์เงิน
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2544

บทคัดย่อ

ในภาคธุรกิจปัจจุบันมีการแข่งขันกันสูง จึงมีความพยายามที่จะคิดค้นและพัฒนาเทคนิคต่างๆ ขึ้น เพื่อใช้ช่วยในการตัดสินใจทางธุรกิจ และเป็นที่ยอมรับกันโดยทั่วไปว่า ข้อมูลคือหัวใจสำคัญในการทำธุรกิจ การที่เรามีข้อมูลจำนวนมาก เข้าใจถึงพฤติกรรมของข้อมูลเหล่านั้น และสามารถนำไปใช้ได้อย่างถูกต้อง ก็จะสร้างโอกาสให้กับธุรกิจมากขึ้น กระบวนการทางด้าน Data Mining จึงเข้ามามีบทบาทในธุรกิจ เนื่องจากสามารถใช้วิธีการอันชาญฉลาดมาช่วยในการตัดสินใจทางธุรกิจได้เป็นอย่างดี โครงการนี้จะนำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่ออนุมัติสินเชื่อเบื้องต้น โดยระบบจะทำการจำแนกลักษณะของลูกค้าที่อยู่ในกลุ่มที่มีเครดิตที่ดีและไม่ดี เพื่อลดความเสี่ยงในการอนุมัติสินเชื่อให้แก่ลูกค้า รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้เพื่อพิจารณาปัจจัยอื่นๆ ที่มีผลต่อการดำเนินธุรกิจต่อไป โดยใช้ C4.5 Algorithm ซึ่งเป็นอัลกอริทึมหนึ่งใน Classification ซึ่งเป็นเทคนิคหนึ่งในดาต้าไมนิ่ง (Data Mining) ในการแก้ปัญหา

Title	C4.5 Classification Tree for Credit Approval
Student	Miss Narumon Somboonngearn
Advisor	Asst. Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2001

ABSTRACT

Since today's business is very competitive, people attempt to develop different techniques to assist in business decision making. As we all know, data is the heart of any business. Once we have information, understand the nature of it, and know the way to apply it effectively, there will be a window of opportunity in our business. Data Mining, therefore, plays an important role in business because of its intelligent methodology applications for business decision making.

This project will present steps and system development methods to efficiently assist business decision making. It will identify the types of clients into good and bad credit. This helps identify credit loan risks. Additionally it is a guideline to apply to other factors, which will effect the business. C4.5, one of the algorithms in classification, is one of the problem solving techniques used by Data Mining.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ผศ.ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานที่ได้กรุณาสละเวลาให้ความรู้ คำปรึกษาและคำแนะนำต่าง ๆ อันเป็นประโยชน์ต่อการพัฒนาระบบ

นอกจากนี้ข้าพเจ้าขอกราบขอบพระคุณบุพการี และบุคคลในครอบครัว ที่ได้ให้การสนับสนุนส่งเสริมเป็นกำลังใจในการเรียนตลอดมา ตลอดจนขอขอบคุณเพื่อน ๆ IS9 ภาคสมทบ และเพื่อนๆ ภาควิชาวิทยาการคอมพิวเตอร์ รุ่นที่ 7 คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่มีส่วนให้ความช่วยเหลือ เป็นกำลังใจ และสนับสนุนให้ผลงานนี้สำเร็จลุล่วงด้วยดี

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าโครงการพัฒนาระบบงานนี้ จะเป็นประโยชน์แก่ผู้ที่สนใจ สำหรับข้อบกพร่องของระบบนี้ ข้าพเจ้าขอรับไว้ เพื่อนำไปปรับปรุงแก้ไขในคราวต่อไป สำหรับความดีที่ได้รับจากโครงการพัฒนาระบบงานนี้ ข้าพเจ้าขอมอบให้แก่บุพการี

นฤมล สมบูรณ์เงิน

กุมภาพันธ์ 2545

สารบัญ

สารบัญ(ต่อ)

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

ภาพที่

สารบัญภาพ(ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

เนื่องมาจากในภาคธุรกิจปัจจุบันมีการแข่งขันกันสูง ไม่ว่าจะเป็นการแข่งขันทั้งจากภายในและภายนอกประเทศ ทำให้องค์กรต้องปรับตัวเพื่อให้มีความได้เปรียบในเชิงการแข่งขันมากที่สุด เพื่อที่จะสามารถครองส่วนแบ่งทางการตลาดและทำกำไรสูงสุด รวมถึงการสร้างความพึงพอใจสูงสุดให้กับผู้บริโภค จึงมีความพยายามที่จะคิดค้นและพัฒนาเทคนิคต่างๆ ขึ้น เพื่อใช้ช่วยในการตัดสินใจทางธุรกิจ ส่งผลให้ธุรกิจสามารถแข่งขันได้ในตลาด และเป็นที่ยอมรับกันโดยทั่วไปว่า ข้อมูลคือหัวใจสำคัญในการทำธุรกิจ การที่เรามีข้อมูลจำนวนมาก เข้าใจถึงพฤติกรรมของข้อมูลเหล่านั้น และสามารถนำไปใช้ได้อย่างถูกต้อง ก็จะสร้างโอกาสให้กับธุรกิจมากขึ้น จึงได้นำเอาเทคนิคของดาต้าไมนิ่ง (Data Mining) เข้ามาช่วยในการวิเคราะห์ข้อมูลเพื่อให้ทราบถึงความสัมพันธ์ในรูปแบบต่าง ๆ ที่ซ่อนอยู่ในคลังข้อมูล

1.2 วัตถุประสงค์

เพื่อนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการวิเคราะห์ลักษณะของลูกค้าทั้งกลุ่มที่มีเครดิตที่ดีและไม่ดี วัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อ เพื่อลดความเสี่ยงในการอนุมัติสินเชื่อให้แก่ลูกค้า รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้เพื่อพิจารณาปัจจัยอื่นๆ ที่มีผลต่อการดำเนินธุรกิจ

1.3 ขอบเขตการดำเนินงาน

โครงการนี้เป็นการศึกษาถึงการนำเอาเทคนิคของดาต้าไมนิ่งมาประยุกต์ใช้ โดยอาศัยหลักการของอัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมหนึ่งใน Classification ในการแบ่งกลุ่มลูกค้า โดยจะนำเสนอผลลัพธ์ในรูปของกฎและ Decision Tree เพื่อนำมาวิเคราะห์ลักษณะของลูกค้าในกลุ่มที่มีเครดิตที่ดีและไม่ดี

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ภายใต้ขอบเขตของการศึกษา จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1) ศึกษาและเก็บรวบรวมข้อมูล
2) ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องของดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้
3) ศึกษาอัลกอริทึม C4.5 เพื่อนำมาประยุกต์ใช้กับระบบ
4) ออกแบบและพัฒนาระบบงานเพื่อใช้ในการแบ่งกลุ่มลูกค้า
5) สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนาระบบงานเพื่ออนุมัติสินเชื่อเบื้องต้น คาดว่าจะให้ประโยชน์แก่เจ้าของงานและผู้ค้นคว้าดังนี้

1) เพื่อให้ขบวนการของการทำดาต้าไมนิ่งทำการแบ่งกลุ่มลูกค้าด้วยเงื่อนไขและความสัมพันธ์ของข้อมูลที่ไม่สามารถคาดการณ์ได้จากข้อมูลเก่า ๆ
2) เป็นแนวทางในการนำดาต้าไมนิ่งมาประยุกต์ใช้กับข้อมูลทางธุรกิจ
3) เข้าใจหลักการและขั้นตอนของการทำดาต้าไมนิ่ง
4) เป็นแนวทางในการออกแบบและพัฒนาโปรแกรมวิเคราะห์ข้อมูลโดยใช้วิธีการอื่นๆ ต่อไป

ในบทนี้เป็นการกล่าวถึงวัตถุประสงค์และขอบเขตของการทำงานในเบื้องต้นของระบบ ในบทต่อไปจะกล่าวถึงรายละเอียดของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

ดาต้าไมนิ่งเป็นเครื่องมือที่มีประสิทธิภาพในการค้นหาสารสนเทศที่มีประโยชน์จากฐานข้อมูลซึ่งเป็นที่รู้จักกันอย่างแพร่หลาย และในปัจจุบันได้มีการนำดาต้าไมนิ่งมาประยุกต์ใช้ในหลายธุรกิจ โดยมีจุดประสงค์เพื่อความได้เปรียบในเชิงธุรกิจ ในบทนี้จะกล่าวถึงคำจำกัดความของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง

ดาต้าไมนิ่ง เป็นกระบวนการที่ใช้ Data analysis และ modeling techniques ที่หลากหลายเพื่อค้นหารูปแบบและความสัมพันธ์ในข้อมูล ที่อาจนำไปใช้สร้างการทำนายที่ถูกต้องแม่นยำ เพื่อให้สามารถนำมาช่วยในการตัดสินใจทางธุรกิจ และสามารถปรับเปลี่ยนกลยุทธ์ในการดำเนินธุรกิจได้อย่างทันต่อสถานการณ์การแข่งขันและการเปลี่ยนแปลงได้อย่างมีประสิทธิภาพยิ่งขึ้น ซึ่งกระบวนการค้นหาสารสนเทศจากคลังข้อมูลนี้ต้องผ่านกระบวนการจัดเตรียมข้อมูล (Preprocess Data) การค้นหาและจัดรูปแบบ (Search for pattern) จนกระทั่งได้ข้อมูลตามต้องการ การค้นหานี้อาจทำได้โดย

- ผู้ใช้เป็นผู้กำหนดคำถาม และระบบจะเป็นผู้ตอบคำถามเหล่านั้น เช่น อาจใช้การซักถาม (Query) และการรายงาน (Reporting) ซึ่งข้อบกพร่องจากการค้นหาแบบนี้คือ ผู้ใช้มักจะไม่ได้คิดถึงสิ่งที่สัมพันธ์กันหรือสิ่งที่ต้องการถามได้อย่างครอบคลุมทั้งหมด ทำให้ข้อมูลส่วนที่สำคัญหลายส่วนอาจไม่ได้ถูกคัดเลือก
- โปรแกรมทางด้านดาต้าไมนิ่ง จะค้นหาข้อมูลอย่างอัตโนมัติโดยโปรแกรมจะคิดคำถามที่น่าสนใจด้วยตัวเอง เมื่อพบข่าวสารแล้วจะแสดงในรูปแบบที่เหมาะสม เช่น กราฟ รายงาน หรือตัวอักษร

เดิมทีแม้ว่าข้อมูลในคลังข้อมูลจะผ่านกระบวนการในการจัดเก็บอย่างเป็นระบบ และมีประสิทธิภาพสูง แต่ถ้าขาดซึ่งกระบวนการในการทำสารสนเทศจากคลังข้อมูลมาใช้อย่างมีประสิทธิภาพและถูกวิธีแล้ว ข้อมูลต่าง ๆ ที่ถูกจัดเก็บไว้จะไม่มีประโยชน์เลย ปัจจุบันเราจึงเริ่มนำดาต้าไมนิ่งมาใช้ในการค้นหาข้อมูลควบคู่ไปกับการพัฒนาเครื่องมือเครื่องใช้ในการที่จะอำนวยความสะดวกด่างๆ เนื่องจากมองเห็นความสำคัญของดาต้าไมนิ่งในการค้นหาความรู้ในฐานข้อมูลเพื่อให้

เข้าใจถึงความสัมพันธ์ต่าง ๆ ในฐานข้อมูลได้เป็นอย่างดี และสามารถนำความรู้ที่ได้ไปประยุกต์ใช้ในธุรกิจสาขาต่าง ๆ ตลอดจนใช้ในชีวิตประจำวัน เทคโนโลยีดาต้าไมนิ่งจึงเป็นเทคโนโลยีในการค้นหาความรู้ในฐานข้อมูลโดยไม่ต้องตั้งสมมติฐานไว้ล่วงหน้า. แต่เป็นการนำความรู้ที่ได้มาทดสอบสมมติฐานภายหลัง สารสนเทศที่ได้มาจากการทำดาต้าไมนิ่งต้องมีลักษณะไม่รู้มาก่อนล่วงหน้า(Unknown) เป็นข้อมูลที่มีความถูกต้อง(Valid) และสามารถนำไปใช้ประโยชน์ได้จริง (Actionable) กล่าวคือ

- ข้อมูลที่ไม่รู้มาก่อนล่วงหน้า(Unknown) เป็นข้อมูลที่ผู้ใช้งานไม่รู้มาก่อนและไม่ชัดเจน ไม่สามารถตั้งสมมติฐานล่วงหน้าว่าควรเป็นแบบใด เช่น เจ้าของห้างสรรพสินค้าแห่งหนึ่งเพิ่งค้นพบพฤติกรรมของผู้บริโภคใหม่ว่าผู้บริโภคที่เป็นพ่อบ้านมักจะซื้อสินค้าเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น จากข้อมูลที่ได้เป็นสัญญาณให้เจ้าของกิจการเตรียมสินค้าไว้เพื่อจำหน่าย ขณะเดียวกันห้างสรรพสินค้าคู่แข่งอาจไม่รู้เรื่องนี้เลยก็ได้
- ข้อมูลที่มีความถูกต้อง(Valid) เป็นข้อมูลที่มีความถูกต้อง เนื่องจากเมื่อผู้ใช้ใช้เทคนิคดาต้าไมนิ่งจะค้นพบสิ่งที่น่าสนใจตลอดเวลา แต่ต้องพิจารณาด้วยว่าสิ่งนั้นถูกต้องหรือไม่ เช่น ผู้ใช้มักพบว่าเมื่อจำนวนความหลากหลายของสินค้ามากขึ้นจะมีความสัมพันธ์ของการซื้อของ 2 สิ่งเสมอ แต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้าเก็บสินค้ามากขึ้น เพราะข้อมูลที่ได้อาจเกิดจากความคลาดเคลื่อน
- ข้อมูลที่สามารถนำไปใช้ประโยชน์ได้จริง(Actionable) คือ ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจให้เป็นความได้เปรียบเชิงธุรกิจ บางครั้งข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้วหรือเป็นสิ่งผิดกฎหมาย ข้อมูลดังกล่าวจะไม่มีประโยชน์อะไร ดังนั้น จำเป็นต้องใช้วิจารณญานในการเลือกใช้ข้อมูลด้วย

2.2 กระบวนการทำงานของดาต้าไมนิ่ง

กระบวนการของดาต้าไมนิ่งเป็นกระบวนการของการสร้างแบบจำลอง (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)

ขั้นตอนที่ 1 : กำหนดจุดประสงค์ทางธุรกิจ

ในกระบวนการทำงานของดาต้าไมนิ่งนั้น การกำหนดวัตถุประสงค์ทางธุรกิจเป็นพื้นฐานหลักในการทำดาต้าไมนิ่ง ซึ่งส่วนนี้เป็นส่วนที่บอกถึงวัตถุประสงค์และสิ่งที่ต้องการจากการทำดาต้าไมนิ่ง ดังนั้นจึงต้องกำหนดปัญหาและเป้าหมายให้ชัดเจน ซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจ และการวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลใดอยู่บ้างและต้องการอะไรจากข้อมูล ซึ่งเป้าหมายทางธุรกิจนี้จะนำไปสู่การสร้างแบบจำลองที่เหมาะสม ซึ่งแบบจำลองที่สร้างขึ้นจะแตกต่างกัน โดยขึ้นอยู่กับเป้าหมายทางธุรกิจ เช่นการเพิ่มอัตราการตอบสนอง และการเพิ่มอัตราการตอบสนองจะได้แบบจำลองที่แตกต่างกัน

ขั้นตอนที่ 2 : การเตรียมข้อมูล

เป็นขั้นตอนที่สำคัญที่สุด ซึ่งต้องใช้เวลาและความพยายามมากกว่าขั้นตอนอื่นๆ ทั้งหมดรวมกัน โดยจะต้องมีการย้อนกลับมาทำซ้ำในขั้นตอนการเตรียมข้อมูล และขั้นตอนการสร้าง Model เนื่องจากการเรียนรู้บางสิ่งจาก Model อาจนำไปสู่การแก้ไขข้อมูล ขั้นตอนการจัดเตรียมข้อมูลนี้ใช้เวลาถึง 60 เปอร์เซ็นต์ ซึ่งประกอบด้วย 3 ขั้นตอนย่อยคือ

1) การเลือกข้อมูล (Data selection)

จุดประสงค์หลักคือการระบุลักษณะและเลือกข้อมูลที่ต้องการและนำข้อมูลที่ไม่ต้องการออกไปซึ่งเป็นการเริ่มต้นของการเตรียมการไมนิ่ง การเลือกข้อมูลนั้นจะแตกต่างไปตามจุดประสงค์ของแต่ละธุรกิจที่ได้กำหนดไว้แต่ต้น

การเลือกข้อมูลจำเป็นต้องมีความเข้าใจกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูลและลักษณะอื่น ๆ ตัวแปรข้อมูลมี 2 ลักษณะ

- ตัวแปรแบบ Categorical แบ่งเป็น
  - Nominal เป็นตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น เพศ(ชาย, หญิง), ระดับการศึกษา (ปริญญาโท, ปริญญาตรี, ม.ปลาย, ปวช)
  - Ordinal เป็นตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น เช่น เกรด (A, B, C, D, F) ถ้าแปลงให้อยู่ในรูปตัวเลขต้องให้ได้ความหมายเดิม
- ตัวแปรแบบ Quantitative แบ่งเป็น
  - Continuous ค่าที่เก็บเป็นเลขจำนวนจริง (Real number) หรือเป็นค่าที่ต่อเนื่อง เช่น รายได้
  - Discrete ค่าที่เก็บเป็นเลขจำนวนเต็ม (Integer) เช่น ข้อมูลจำนวนพนักงาน

ในการเลือกข้อมูลต้องคำนึงถึงอายุของข้อมูลด้วย เช่น ข้อมูลของอาชีพลูกค้า ซึ่งจะมีการเปลี่ยนแปลงบ่อยเมื่อเวลาผ่านไป เพราะฉะนั้นการนำเอาข้อมูลอาชีพของลูกค้ามาใช้นั้นต้องตรวจสอบให้

แน่ชัดว่าข้อมูลนั้นถูกต้องหรือไม่ นอกจากนี้ ยังมีหลักเกณฑ์ที่ต้องพิจารณาเพิ่มเติมเกี่ยวกับข้อมูลที่จะนำมาใช้อยู่ 4 ประเด็นคือ

1. ระดับของข้อมูลที่พิจารณา

สิ่งที่นำมาช่วยตัดสินใจว่าข้อมูลที่นำมาใช้ควรเป็นข้อมูลระดับรายการ (Item) หรือ ข้อมูลที่สรุปแล้ว คือวัตถุประสงค์ในการทำดาต้าไมนิ่ง เช่น

- การทำไมนิ่งเกี่ยวกับการโทรศัพท์ ถ้าจุดประสงค์ของเราต้องการเน้นไปที่พฤติกรรมการใช้โทรศัพท์ของลูกค้า ข้อมูลที่จัดเก็บโดยปกติแล้วจะมีการจัดเก็บเป็นลักษณะรายละเอียดของแต่ละชุมสาย การเคลื่อนย้ายของอิเล็กตรอนไปยังสวิตชิง ข้อมูลเหล่านี้จะไม่มีประโยชน์เลย เพราะจุดประสงค์ของเราสนใจสิ่งที่อยู่ภายใต้การควบคุมของลูกค้าและมีผลต่อการตลาด ดังนั้น ข้อมูลที่เราสนใจจะเป็น เบอร์โทรศัพท์ของผู้โทร, เวลาเริ่มต้นที่ใช้โทร และเวลาที่ใช้ในการโทรศัพท์แต่ละครั้ง
- ข้อมูลที่ยังไม่สรุป ทำให้จัดการได้ยาก รวมทั้งเกิดจำนวนการคอมไบเนชั่น(Combination) สูง เมื่อใช้เทคนิคของ Association Discovery เพราะข้อมูลของร้านค้าปลีกย่อมมีรายการสินค้าเยอะ ดังนั้น การนำเอาหน่วยวัดในการจัดเก็บสินค้าในคลัง หรือ SKU (Stock Keeping Unit) เข้ามาช่วยจะสามารถลดจำนวนการคอมไบเนชั่นลงได้

2. ลักษณะของข้อมูลที่จัดเก็บ

การจัดเก็บข้อมูลด้วยภาษาคอมพิวเตอร์ที่แต่ละระบบปฏิบัติการเลือกใช้แตกต่างกัน ทำให้ข้อมูลที่นำมาวิเคราะห์มีผลกระทบ เช่น ข้อมูลที่นำมาวิเคราะห์ส่วนมากจัดเก็บด้วยภาษา COBOL และ RPG ข้อมูลที่เป็น Text จะถูกเก็บเป็น EBCDIC และข้อมูลตัวเลขจะเก็บเป็น Packed Decimal ขณะที่ภาษาที่เลือกใช้ในการสร้างระบบ ดาต้าไมนิ่งใช้ภาษา C และ C++ ซึ่งข้อมูลชนิด Text จะมีรูปแบบเป็น ASCII และข้อมูลตัวเลขเก็บเป็น Integer หรือ Floating Point

3. ความแตกต่างของข้อมูลแต่ละแหล่ง

เมื่อข้อมูลที่นำมาวิเคราะห์มาจากหลายแหล่ง ซึ่งแต่ละแหล่งมีรูปแบบการจัดเก็บข้อมูลที่ต่างกัน เช่น การวิเคราะห์ข้อมูลการโทรศัพท์ (Call Detail) เพื่อหาเบอร์โทรศัพท์ที่ใช้ฝากข้อความเข้า Voice Mailbox ในแต่ละเมือง จะมีวิธีการจัดเก็บข้อมูลที่ต่างกัน เช่น เมือง ๆ หนึ่งอาจเก็บเบอร์โทรศัพท์ที่ใช้โทรเข้า Voice Mailbox ด้วยเบอร์ต้นทางและปลายทาง แต่อีกเมืองหนึ่ง อาจเก็บเบอร์โทรศัพท์ที่ไม่รู้ด้วยเบอร์ปลายทาง อีกเมืองหนึ่งอาจเก็บเบอร์โทรศัพท์ที่โทรเข้า Voice Mailbox จริง ๆ ดังนั้น จึงจำเป็นต้องทำข้อมูลเหล่านี้ให้ออกมาในรูปแบบมาตรฐานเดียวกันก่อน เพื่อที่จะได้เข้าใจถึงความแตกต่างในการเก็บข้อมูลของแต่ละแหล่งได้

4. ข้อมูลที่เป็นข้อความ (Textual Data)

ข้อมูลที่จัดเก็บเป็นแบบ Text อาจก่อให้เกิดความสับสน เช่น '_no' *กับ* 'no_' หรือ 'VOR2J0' กับ 'VOR 2J0' กับ 'VOR-2J0' ซอฟท์แวร์ที่ใช้ในการทำดาต้าไมนิ่งย่อมมองข้อมูลเหล่านี้ไม่เหมือนกัน ในทางแก้ไขคือสร้างตารางเก็บค่าที่ถูกต้อง และแทนที่ข้อมูลที่นำมาวิเคราะห์ด้วย Index ตัวอย่างที่เห็นได้ชัดเจน คือ ฐานข้อมูลแบบสัมพันธ์ (Relational Database) มีการแทนที่ข้อมูลที่เป็น Product_Name ด้วย Product_Code ซึ่งมีการ Unique มากกว่า

2) การกลั่นกรองข้อมูล (Data Preprocessing)

จุดประสงค์เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นถูกต้องและ เหมาะสมที่จะนำไปทำดาต้าไมนิ่ง โดยใช้หลักการทางสถิติ และเทคนิคการนำเสนอข้อมูลที่น่าสนใจ (Data Visualization Techniques) สำหรับข้อมูลประเภท Categorical การจัดการกระจายข้อมูลจะทำให้เข้าใจข้อมูลที่มีอยู่ได้ดียิ่งขึ้น วิธีการที่ง่ายที่สุดคือ การนำเอาข้อมูลมาสร้างกราฟ ซึ่งจะช่วยให้เห็นความโน้มเอียงของข้อมูล และข้อมูลที่ผิดปกติได้ ส่วนข้อมูลประเภท Quantitative การวิเคราะห์ข้อมูลทำได้โดยการหาค่าสูงสุด(Max) , ค่าต่ำสุด(Min), ค่าเฉลี่ย (Mean), ค่าที่ปรากฏบ่อย (Mode), ค่ากลาง (Median) เป็นต้น ซึ่งสิ่งผิดปกติที่จะปรากฏให้เห็นในขั้นตอนนี้คือ

- Noisy Data คือค่าของข้อมูลที่ผิดไปจากที่ควรจะเป็น อาจเกิดจากความเลินเล่อในการบันทึกข้อมูล เช่น บันทึกอายุเป็น 650 ปี หรือบันทึกรายได้ติดลบ เป็นต้น ซึ่งข้อมูลที่ผิดนี้อาจไปรบกวนการวิเคราะห์ จึงต้องกำจัดข้อมูลที่ผิดนี้ออกไป
- Missing Data คือค่าของข้อมูลที่ขาดหายไปไม่ได้แสดงในข้อมูลที่เราได้เลือกแล้ว หรือค่าที่ไม่สมบูรณ์ที่เราลบออกไประหว่างการทำ Noise Detection ค่าอาจหายไปเพราะเกิดจากความเลินเล่อของมนุษย์ แก้ไขโดยการตัดข้อมูลนั้นทิ้งทั้งรายการ หรือบันทึกส่วนที่ขาดหายไปด้วยค่าเฉลี่ย (Mean) หรือค่าที่ปรากฏบ่อย (Mode) สำหรับข้อมูลประเภท Quantitative ส่วนข้อมูลประเภท Categorical อาจบันทึกด้วยค่าที่ปรากฏบ่อย (Mode) หรือบันทึกเป็น 'UNKNOW'

3) การแปลงข้อมูล (Data Transformation)

เป็นการแปลงข้อมูลให้อยู่ในรูปแบบของข้อมูลที่พร้อมที่จะนำไปวิเคราะห์ตามอัลกอริทึมของดาต้าไมนิ่งที่ใช้ เช่น การแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยแบ่งค่าของตัวแปรให้เป็นช่วง ๆ เช่น การแปลงข้อมูลเงินเดือน นอกจากนี้ยังมีเทคนิคของการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric เช่น ยี่ห้อรถ HONDA, TOYOTA และ NISSAN ให้เป็น 001, 010 และ 011

ขั้นตอนที่ 3 : การทำดาต้าไมนิ่ง

เป็นการนำข้อมูลที่จัดเตรียมไว้มาทำดาต้าไมนิ่ง เพื่อแปลงสภาพของข้อมูลดิบให้เป็นความรู้ ในลักษณะของรูปแบบและกฎเกณฑ์ (Pattern And Rule Finder) ขั้นตอนนี้จะมีความ

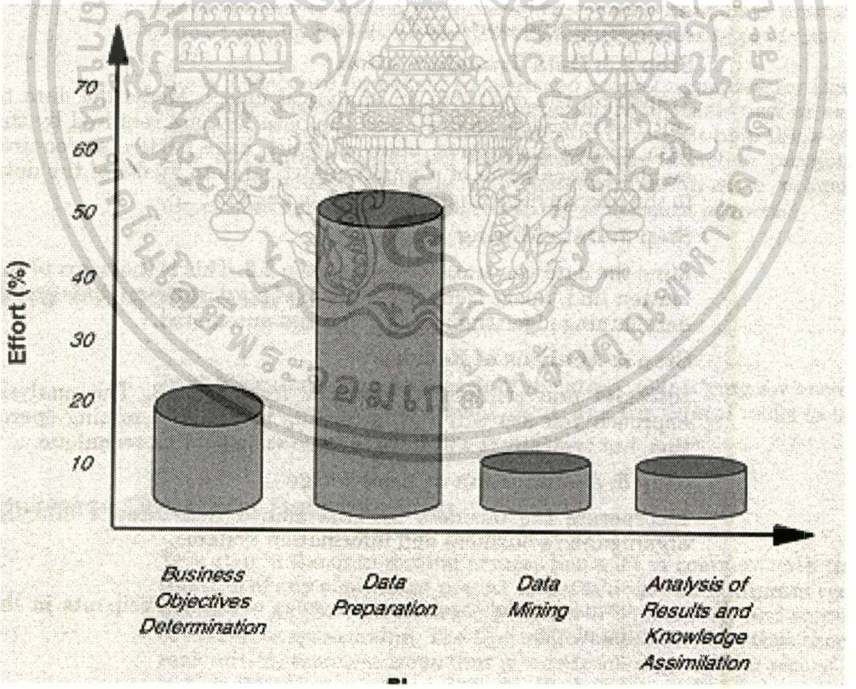
สัมพันธ์กับขั้นตอนที่ผ่านมา โดยเมื่อทำขั้นตอนนี้แล้วอาจต้องย้อนกลับไปทำขั้นตอนของการเตรียมข้อมูลใหม่ ขั้นตอนการทำดาต้าไมนิ่งนี้จะเกี่ยวข้องกับการใช้อัลกอริทึมหลายๆ แบบ

ขั้นตอนที่ 4 : การวิเคราะห์ผลลัพธ์ที่ได้

เป็นการวิเคราะห์และตีความผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง การทำงานในขั้นตอนนี้ต้องใช้ทักษะในการวิเคราะห์ข้อมูล และการวิเคราะห์ทางธุรกิจ ซึ่งทำโดยการนำเอาแบบจำลองที่ได้ไปทดสอบกับข้อมูลชุดอื่น ที่ไม่ใช่ข้อมูลที่ใช้ในการสร้างแบบจำลอง เพื่อนำเอาผลลัพธ์ที่ได้มาเปรียบเทียบกับผลตามแบบจำลอง ว่ามีความแม่นยำและยอมรับได้หรือไม่ ซึ่งถ้าไม่สามารถยอมรับได้ก็ทำการแก้ไข โดยการเพิ่มจำนวนของข้อมูลให้มากขึ้นหรือเปลี่ยนไปใช้อัลกอริทึมอื่นแทน

ขั้นตอนที่ 5 : การปรับความรู้ที่ได้เข้ากับธุรกิจ

เป็นการรวบรวมความเข้าใจทางธุรกิจที่เป็นผลมาจากขั้นตอนที่ 4 มารวมเข้ากับส่วนความรู้เพื่อนำไปใช้ในโอกาสต่อไป ในขั้นตอนนี้มีหลักอยู่ 2 ประการคือ การนำเสนอแนวคิดทางธุรกิจที่ค้นพบใหม่ และหาแนวทางที่จะใช้กฎเกณฑ์ใหม่ที่ค้นพบเพื่อให้เกิดประโยชน์สูงสุด



ภาพที่ 2.1 เปอร์เซ็นต์การทำงานแต่ละขั้นตอนในการทำ Data Mining

ในกระบวนการทำดาต้าไมนิ่งนั้นประกอบด้วยหลายขั้นตอน และในแต่ละขั้นตอนก็จะมีการทำซ้ำๆ ในขั้นตอนนั้นหรือต้องมีการวนกลับมาทำซ้ำใหม่ในหลายขั้นตอน ภาพที่ 2.1 แสดง

เปอร์เซ็นต์ของเวลาที่ใช้ในการทำงานแต่ละขั้นตอน จะเห็นว่าการจัดเตรียมข้อมูลเพื่อทำ Miningใช้เวลาถึง 60 เปอร์เซ็นต์ ดังนั้นจุดที่ต้องให้ความสำคัญคือ การทำความสะอาดข้อมูล (Data Cleaning) ในลักษณะของการกำจัดข้อมูลผิดที่จะไปรบกวนการวิเคราะห์ และการปรับรูปแบบของข้อมูล และการจัดการเกี่ยวกับความขัดแย้งกันของข้อมูล ส่วนเวลาที่ใช้ในการทำ Mining จริงๆมีเพียง 10% ของเวลาทั้งหมด

2.3 โอเปอเรชั่นของดาต้าไมนิ่ง (Data Mining Operation)

ดาต้าไมนิ่งประกอบด้วย 4 โมเดลหลักที่ใช้สำหรับประยุกต์ใช้งานทางธุรกิจ ได้แก่ การสร้างแบบจำลองพยากรณ์(Predictive Modeling), การแบ่งส่วนฐานข้อมูล(Database Segmentation), การวิเคราะห์ความสัมพันธ์ (Link Analysis) และการตรวจสอบค่าเบี่ยงเบน (Deviation detection)

1. การสร้างแบบจำลองพยากรณ์(Predictive Modeling) มีลักษณะคล้ายการเรียนรู้ของมนุษย์คือจะต้องเข้าใจลักษณะของสิ่งที่ศึกษาอย่างแท้จริง ในดาต้าไมนิ่งเราจะใช้โมเดลนี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่ เพื่อกำหนดคุณสมบัติที่สำคัญของข้อมูล ฉะนั้นข้อมูลที่มีอยู่จะต้องเป็นข้อมูลที่สมบูรณ์ จึงจะทำให้แบบจำลองให้คำทำนายที่ถูกต้อง โดยเริ่มต้นจะต้องให้คำตอบที่ถูกต้องกับแบบจำลองเพื่อแบบจำลองจะได้เห็นถึงข้อสังเกตใหม่ๆ วิธีนี้เรียกว่า “Supervised Learning” ซึ่งการทำงานจะมีลักษณะคล้ายกับ IF THEN การพัฒนาแบบจำลองพยากรณ์จะนำเอาข้อมูลในอดีตมาสร้างแบบจำลอง โดยแบ่งออกเป็น 2 ขั้นตอนคือ

- Training Phase เพื่อสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต
- Testing Phase เพื่อทดสอบแบบจำลองที่สร้างว่ามีความเหมาะสมหรือไม่ โดยใช้กับข้อมูลที่ถูกแบ่งเอาไว้สำหรับทำการทดสอบ

Predictive Modeling แบ่งเป็น 2 ลักษณะคือ

- Classification เป็นลักษณะการสร้างแบบจำลองพยากรณ์เพื่อทำนายกลุ่มของรายการที่เราสนใจ ซึ่งกลุ่มต่างๆ จะมีการกำหนดไว้ล่วงหน้าแล้ว เช่น ใช้ในการจัดกลุ่มลูกค้าเพื่อทำนายลักษณะของลูกค้าที่เปลี่ยนไปใช้บริการของคู่แข่ง เป็นต้น
- Value Prediction เป็นการทำนายค่าที่เป็นตัวเลข เช่น การทำนายราคาหุ้น เป็นต้น

2. การแบ่งส่วนฐานข้อมูล (Database Segmentation) เป็นการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยไม่รู้ล่วงหน้าว่าจะมีทั้งหมดกี่กลุ่ม โดยการจัดกลุ่มดังกล่าวได้จากการพิจารณาคุณสมบัติในหลายๆมิติของข้อมูล ถ้ารายการในข้อมูลมีลักษณะคล้ายคลึงเป็นกลุ่มเดียวกันได้ก็จะรวมเข้าด้วยกันเพื่อให้ง่ายต่อการวิเคราะห์ เช่น การแบ่งลูกค้าออกตามอายุ, เพศ, รายได้ เป็นต้น

3. การวิเคราะห์ความสัมพันธ์ (Link Analysis) เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างข้อมูลว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันหรือไม่ อย่างไร เช่น เก็บข้อมูลการซื้อสินค้าแต่ละครั้งของลูกค้าเพื่อศึกษาพฤติกรรมการซื้อสินค้า เพื่อนำมาทำนายการส่งเสริมการขายและการจัดชั้นวางสินค้าให้เหมาะสม

4. การตรวจสอบค่าเบี่ยงเบน (Deviation detection) เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) Operation นี้สามารถใช้ในการตรวจสอบลายเซ็นปลอม หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

โมเดลเหล่านี้นำไปประยุกต์ใช้ในงานทางธุรกิจได้ดังภาพที่ 2.2 แต่จะไม่สามารถเจาะจงได้ว่าธุรกิจ ประเภทใด ต้องใช้โมเดลไหน เพียงแต่บอกว่าลักษณะงานทางธุรกิจใดมีความเกี่ยวข้องกัน และลักษณะงานแบบไหน ควรใช้โมเดลแบบใด

Market Management	Risk Management	Fraud Management
Target Marketing *Customer Relationship* *Market basket analysis* *Cross selling* *Market segmentation*	*Forecasting* *Customer retention* *Improved underwriting* *Quality control* *Competitive analysis*	*Fraud detection*

Predictive Modeling	**Database Segmentation**	**Link Analysis**	**Deviation Detection**
Classification *Value prediction*	*Demographic clustering* *Neural clustering*	*Association discovery* *Sequential pattern discovery* *Similar time sequence discovery*	*Visualization* *Statistics*

ภาพที่ 2.2 โมเดลของดาต้าไมนิ่งกับการประยุกต์ใช้งาน

จากที่กล่าวมาแล้วข้างต้นว่าโอเปอเรชั่นของดาต้าไมนิ่งมีมากมาย สำหรับโครงการที่นำเสนอนี้ จะนำเสนอโอเปอเรชั่นของ Classification โดยการนำอัลกอริทึม C4.5 มาใช้เพื่อทำการแบ่งกลุ่มลูกค้า โดยจะกล่าวถึงรายละเอียดในบทถัดไป

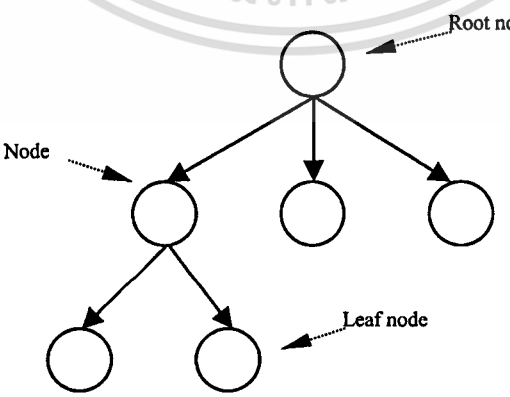
บทที่ 3

การจัดกลุ่ม (Classification)

Classification เป็นเทคนิคหนึ่งในดาต้าไมนิ่งที่ใช้สำหรับสร้างแบบจำลองการพยากรณ์ (Predictive Model) โดยจะทำการสร้างแบบจำลองจากกลุ่มข้อมูลตัวอย่างที่เลือกมาจากฐานข้อมูลขนาดใหญ่ และแบบจำลองนั้นสามารถพยากรณ์ผลลัพธ์ของข้อมูลที่ไม่เคยพบเห็นมาก่อน บนพื้นฐานความสัมพันธ์ของกลุ่มข้อมูลที่มีอยู่เดิม และแบบจำลองที่สามารถทำงานได้ตามลักษณะนี้เรียกว่า supervised learning สำหรับเทคนิคที่ใช้ใน Classification นั้นยังแบ่งได้เป็น 2 แบบ ได้แก่ Tree Induction และ Neural Induction โดยในที่นี้จะนำเสนอเทคนิคของ Tree Induction ในการประยุกต์ใช้งานในธุรกิจการธนาคาร โดยมีจุดมุ่งหมายเพื่อนำไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อ เพื่อลดความเสี่ยงในการอนุมัติสินเชื่อให้แก่ลูกค้า โดย Tree Induction เป็นการนำข้อมูลมาสร้างแบบจำลองพยากรณ์ในรูปของ Decision Tree

3.1 รูปแบบและขั้นตอนพื้นฐานในการสร้าง Tree

รูปแบบของ Tree ประกอบด้วย Node แรกสุดที่เรียกว่า Root node จาก Root node ก็จะแตกออกเป็น node ลูก และที่ node ลูกก็จะมี node ลูกของตัวเอง ซึ่ง node ลูกแต่ละระดับอาจมีมากกว่า 2 node ก็ได้ ส่วน node ที่ระดับสุดท้ายเรียกว่า Leaf node



ภาพที่ 3.1 รูปแบบ Decision Tree

จะเห็นว่า จาก Root node จนถึง Leaf node จะใช้เพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบายถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่ม ซึ่งในแต่ละ Leaf node นั้นอาจเป็นกลุ่มเดียวกันซึ่งเกิดจากเหตุผลที่แตกต่างกันได้

การนำข้อมูลมาสร้าง Tree นั้น มีขั้นตอนพื้นฐาน คือ

1. หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกนำมาสร้างเป็น Root node โดยจะมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf node ถูกกำหนดไว้ก่อน
2. นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกออกเป็นกลุ่มของตัวเอง
3. แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root node
4. วนกลับไปทำแบบขั้นตอนแรก คือ หา Attribute ที่สำคัญที่สุดจากข้อมูลที่เข้ามา เพื่อหา Attribute ที่ใช้แบ่งข้อมูลต่อไป

สำหรับอัลกอริทึมที่ใช้สร้าง Decision Tree มีหลายอัลกอริทึม เช่น CHAID, CART, ID3 และ C4.5 เป็นต้น และในแต่ละอัลกอริทึมจะมีวิธีการที่แตกต่างกันในการหา Attribute ที่จะนำมาใช้แบ่งข้อมูล ซึ่งในที่นี้ได้เลือกอัลกอริทึม C4.5 ของ J. Ross Quinlan มาประยุกต์ใช้ในการพัฒนาระบบเพื่ออนุมัติสินเชื่อเบื้องต้น และเนื่องจาก C4.5 เป็นเวอร์ชันที่พัฒนาจาก ID3 จึงขอกล่าวถึงการทำงานของ ID3 และ C4.5 ในส่วนที่เพิ่มเติมและปรับปรุงการทำงานของ ID3 ออกเป็น 2 หัวข้อดังนี้

1. ID3 Algorithm
2. C4.5 Algorithm

3.2 ID3 Algorithm

การเลือก Attribute ที่มีความสำคัญที่สุดเพื่อใช้แบ่งข้อมูลจะใช้หลักการของ Gain Criterion ในการวัด โดยกำหนดให้

T แทน Training Set

S แทน set ของข้อมูลใดๆ

freq(C_j,S) แทนจำนวนของข้อมูลใน S ซึ่งอยู่ใน Class C_j

|S| แทน จำนวนของข้อมูลใน S

info(S) หรือ entropy ของ set S เป็นการวัดค่าของ information

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \quad bits.$$

และเมื่อนำสูตรนี้ไปประยุกต์ใช้กับ Training Set จะได้ info(T)

$info_X(T)$ เป็นการวัดค่าของ Information เพื่อแบ่ง T โดยใช้ค่าที่เป็นไปได้ของ Attribute X

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i)$$

gain(X) เป็นการวัดค่าของ information ที่ได้รับถ้าเลือก Attribute X

$$gain(X) = info(T) - info_X(T)$$

ต่อไปจะอธิบายการทำงานของ ID3 โดยใช้ข้อมูลตัวอย่างจากตารางที่ 3.1 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 3.1 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

$$info(T) = -9/14 \times \log_2(9/14) - 5/14 \times \log_2(5/14)$$
$$= 0.940 \, bits.$$

พิจารณา Attribute ต่างๆ โดยหาค่า Gain ของแต่ละ Attribute ออกมา แล้วเลือก Attribute ที่มีค่า Gain สูงสุดมาเป็นตัวแบ่งข้อมูล หรือ Root node

พิจารณาที่ Attribute Outlook ซึ่งสามารถแบ่งข้อมูลได้เป็น 3 subset จะได้ว่า

$$info_X(T) = 5/14 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$$
$$+ 4/14 \times (-4/4 \times \log_2(4/4) - 0/4 \times \log_2(0/4))$$
$$+ 5/14 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))$$
$$= 0.640 \, bits.$$

ดังนั้นค่า Gain ของ Attribute Outlook หรือ gain(x) = 0.940 – 0.694 =0.246 bits

ในทำนองเดียวกัน ถ้าพิจารณาที่ Attribute windy จะแบ่งข้อมูลได้ 2 subset โดย set แรกมีข้อมูลจำนวน 6 record ซึ่งอยู่ใน class Play 3 record และอยู่ใน class Don't Play 3 record ส่วน set ที่ 2 มีข้อมูลทั้งหมด 8 record อยู่ใน class Play 6 record และอยู่ใน class Don't Play 2 record ดังนั้น

$$info_X(T) = 6/14 \times (-3/6 \times \log_2(6/6) - 3/6 \times \log_2(3/6))$$
$$+ 8/14 \times (-6/8 \times \log_2(6/8) - 2/8 \times \log_2(2/8))$$
$$= 0.892 \, bits.$$

$$gain(x) = 0.940 - 0.892 = 0.048 \, bits.$$

จะพบว่าค่า Gain ที่ได้จากการแบ่ง Training Set โดยใช้ Attribute Outlook มากกว่า Windy ดังนั้นควรใช้ Attribute Outlook ในการแบ่ง Training Set

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 Algorithm

พัฒนามาจาก ID3 โดยเพิ่ม Feature ต่างๆ ขึ้นมาดังนี้

- **Gain ratio criterion** พัฒนาขึ้นเพื่อแก้ปัญหาของ Gain Criterion กรณีที่ Attribute มีค่าที่ unique การแบ่งข้อมูลโดยใช้ Attribute นี้จะทำให้เกิด subset จำนวนมาก ซึ่งแต่ละ subset จะ

ประกอบด้วยข้อมูลเพียง 1 record เท่านั้น ทำให้ $info_X(T) = 0$ ซึ่งจะมีผลให้ค่า Information Gain ของ Attribute นี้มีค่าสูงมาก และการแบ่งข้อมูลโดยใช้ Atrribute นี้ไม่ก่อให้เกิดประโยชน์ใดๆ ต่อการทำนาย C4.5 แก้ไขโดยใช้ค่า Gain ratio ซึ่งคำนวณโดยใช้ split info (X) และ gain ratio(X) โดย

split info(X) เป็นค่า Information ที่ได้จากการแบ่ง T ออกเป็น n subset

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

gain ratio(X) เป็นการวัดว่าการแบ่งข้อมูลโดยใช้ Attribute นั้นๆ ก่อให้เกิดประโยชน์ต่อการทำนายหรือไม่

$$gain\ ratio(X) = gain(X) / split\ info(X)$$

จากตัวอย่างที่แล้วพบว่าการแบ่งข้อมูลโดยใช้ Attribute Outlook ทำให้เกิด subset ทั้งหมด 3 subset ซึ่งประกอบด้วยจำนวน record เท่ากับ 5, 4 และ 5 record ตามลำดับ Split Information คำนวณได้ดังนี้

$$split\ info(X) = -5/14 \times \log_2(5/14) - 4/14 \times \log_2(4/14) - 5/14 \times \log_2(5/14) = 1.577\ bits.$$

$$gain\ ratio(X) = 0.246 / 1.577 = 0.156\ bits.$$

ซึ่งพบว่าการใช้ Gain ratio criterion ทำให้ Tree ที่ได้มีขนาดเล็กกว่าการใช้ Gain criterion

- **Unknown attribute values**
  - การหา Attribute เพื่อใช้แบ่งข้อมูล ทำโดย
    1. หาค่า info(T) และ $info_X(T)$ โดยพิจารณาเฉพาะข้อมูลที่รู้ค่าของ A
    2. หาค่า gain(X) โดย

$$gain(X) = probability\ A\ is\ known \times (info(T) - info_X(T)) + probability\ A\ is\ not\ known \times 0 = F \times (info(T) - info_X(T))$$

3. หาค่า split info(X) โดยพิจารณากลุ่มของข้อมูลที่ไม่รู้ค่าของ A เป็นอีก 1 subset เช่น ถ้า Attribute ที่จะนำมาทดสอบมีค่าที่เป็นไปได้ n ค่า split info(X) จะถูกคำนวณโดยแบ่งข้อมูลออกเป็น n+1 subsets

- การแบ่งTraining Set สมมุติ Attribute ที่เลือกจากขั้นตอนแรกมีค่าที่เป็นไปได้คือ $O_1,O_2,...O_n$ 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 ที่จุดนั้น ๆ ทำให้ $|T_i|$ เป็นผลรวมของค่า weight ใน subset T_i และ record ใน Tree ซึ่งมีค่า weight w ซึ่งค่าใน Attribute ไม่ทราบค่าจะถูกกำหนดให้แต่ละ subset T_i ด้วย weight

$$w \times probability\ of\ outcome\ O_i$$

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 ที่ได้จากการทำนาย จะเป็น class ที่มีความน่าจะเป็นสูงสุด

ต่อไปจะนำเสนอตัวอย่างเพื่อให้เกิดความเข้าใจยิ่งขึ้น โดยใช้ตัวอย่างจากตารางที่ 3.1 โดยแบ่งเป็น 3 ขั้นตอนดังนี้

1. การหา Attribute เพื่อใช้แบ่งข้อมูล สมมุติว่าค่าในAttribute outlook ใน record ที่ 6 เป็นค่าที่ไม่ทราบค่า ซึ่งแทนโดย "?" ซึ่งเราจะพิจารณาเฉพาะข้อมูล 13 record ที่เหลือจะได้ความถี่ดังนี้

	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

ทำการคำนวณต่าง ๆ โดยพิจารณา Attribute Outlook ดังนี้

$$info(T) = -8/13 \times \log_2(8/13) - 5/13 \times \log_2(5/13)$$
$$= 0.961 bits$$

$$info_X(T) =$$
$$5/13 \times (-2/5 \times \log_2(2/5) - 3/5 \times \log_2(3/5))$$
$$+ 3/13 \times (-3/3 \times \log_2(3/3) - 0/3 \times \log_2(0/3))$$
$$+ 5/13 \times (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))$$
$$= 0.747 bits$$

$$gain(X) = 13/14 \times (0.961 - 0.747)$$
$$= 0.199 bits$$

จะพบว่าค่า gain ที่ได้จะลดลงเล็กน้อยจากเดิม 0.246 เป็น 0.199 bits ส่วนค่า split information จะพิจารณาจากข้อมูลใน training set ทั้งหมด จึงทำให้ค่าที่ได้มีเพิ่มขึ้นจาก 1.577 เป็น 1.809 bits 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 ($^{\circ}$F)	Humidity (%)	Windy?	Decision	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 ≤ 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) จะมีความสำคัญ

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 ที่ได้มาทำนายกลุ่มของข้อมูล สมมุติข้อมูลคือ

Sunny outlook, temperature 70°, unknown humidity, windy false

จากค่าใน Outlook พบว่าต้อง move ไปยัง subset แรก แต่เนื่องจากไม่สามารถตรวจสอบค่าที่ humidity ได้ จึงทำการพิจารณาดังนี้

- ถ้า humidity $\leq$ 75 % จะได้ class Play และ
- ถ้า humidity > 75% จะได้ class Don't Play ด้วยความน่าจะเป็นเท่ากับ 3/3.4(85%) และ class Play ด้วยความน่าจะเป็นเท่ากับ 0.4/3.4(12%)

จะพบว่าการกระจายของ class สุดท้ายสำหรับข้อมูลนี้เท่ากับ

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

Don't Play: 3.4/5.4 x 88% = 56%

- **Continuous attribute values** สมมุติว่า A เป็น Attribute ชนิด continuous numeric value การทดสอบค่าที่ Attribute นี้ จะแบ่งเป็น $A \leq Z$ และ $A > Z$ โดยทำการเปรียบเทียบค่าของ A กับค่า Threshold value Z โดยการหาค่า Threshold ที่เหมาะสมมีขั้นตอนดังนี้

1. เรียงลำดับ Training Set ด้วยค่าใน Attribute A จากน้อยไปมาก และเลือกเฉพาะค่าที่ไม่ซ้ำกันมาพิจารณาจะได้ $\{v_1, v_2, \ldots, v_n\}$
2. หาค่า Threshold ใดๆ ซึ่งค่า Threshold ใดๆ จะอยู่ระหว่าง v_i และ v_{i+1} โดยคำนวณจาก Midpoint ของแต่ละช่วงดังนี้ $v_i + v_{i+1} / 2$

   โดย C4.5 จะเลือกค่าที่มากที่สุดใน Attribute A แต่ต้องไม่เกินค่า Midpoint นั้นๆ 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 trees** การแบ่งข้อมูลใน Training set เพื่อสร้าง Decision tree จะกระทำไปจนกระทั่งข้อมูลในแต่ละ subset อยู่ใน class เดียวกัน ซึ่งอาจจะทำให้ Tree ที่ได้มีความซับซ้อนและขึ้นกับกลุ่มข้อมูลที่ใช้ในการฝึกสอนมากเกินไปที่เรียกว่า "overfits the data" ได้ ซึ่งปัญหานี้สามารถทำการแก้ไขได้โดยทำการ Pruning การ Pruning จะทำให้แต่ละ Leaf node ของ Tree ที่ได้ไม่จำเป็นที่จะต้องประกอบด้วยข้อมูลที่อยู่ใน class เดียวกันทั้งหมด โดยแต่ละ Leaf node จะมีการระบุการกระจายของข้อมูลแต่ละ class ไว้ ซึ่งจะบอกถึงความน่าจะเป็นที่ข้อมูลจะอยู่ใน class นั้นๆ 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 จะทำโดยสมมุติว่าจะทำการแบ่งกลุ่ม set ของข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Training set โดยการคำนวณจะใช้ function ทางสถิติ ซึ่งอยู่บนพื้นฐานของการกระจายแบบ Binomial

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

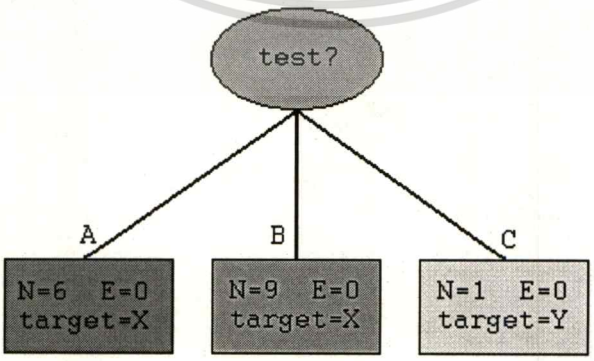
โดย N แทน ขนาดของข้อมูลที่ Leaf node ใดๆ

E แทน จำนวนของ error ที่เกิดขึ้นใน set ของข้อมูลที่ Leaf node ใดๆ

$U_{CF}(E,N)$ แทนความน่าจะเป็นสูงสุดที่จะเกิด error

และ C4.5 จะใช้ค่าระดับความเชื่อมั่น(Confidence level) ที่ 25 %

ต่อไปจะอธิบายการ Pruning โดยพิจารณา subtree ดังรูป



ภาพที่ 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

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

ดังนั้นถ้าเราใช้ Leaf node นี้ใน การแบ่งกลุ่มข้อมูลจำนวน 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

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

และจะพบว่าถ้าทำการแทนที่ 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 โดย แทนที่ด้วย Leaf node

บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งในการจัดกลุ่มลูกค้า เพื่ออนุมัติสินเชื่อเบื้องต้น

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงต้องมีการกำหนดวัตถุประสงค์และกระบวนการต่าง ๆ สำหรับจัดเตรียมข้อมูลให้อยู่ในรูปที่เหมาะสมก่อนที่จะนำมาใช้งาน ซึ่งมีขั้นตอนดังนี้

4.1 กำหนดวัตถุประสงค์

ในการตัดสินใจอนุมัติสินเชื่อในธุรกิจธนาคารนั้น ทางธนาคารจำเป็นต้องพิจารณาความเสี่ยงที่จะเกิดขึ้นจากการอนุมัติสินเชื่อนั้นด้วย เพราะถ้าลูกค้าไม่สามารถชำระเงินคืนได้ จะทำให้ธนาคารสูญเสียกำไร และความสามารถในการให้บริการแก่ลูกค้ารายใหม่ อาจส่งผลให้ธนาคารไม่สามารถคงอยู่ในธุรกิจต่อไปได้

จึงมีความคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้ในธุรกิจธนาคาร โดยมีวัตถุประสงค์เพื่อวิเคราะห์ลักษณะของลูกค้าทั้งกลุ่มที่มีเครดิตที่ดีและไม่ดี เพื่อนำมาช่วยในการตัดสินใจพิจารณาอนุมัติสินเชื่อต่อไป และยังสามารถนำผลลัพธ์ที่ได้ ไปใช้ประกอบในการพิจารณาปัจจัยอื่นได้อีกด้วย เช่น การนำไปพิจารณาอัตราดอกเบี้ยเงินกู้ เป็นต้น

4.2 การคัดเลือกข้อมูล

ทำการคัดเลือกว่าจะใช้ข้อมูลอะไร จากส่วนไหนเพื่อที่จะทำให้บรรลุวัตถุประสงค์ที่ต้องการ ซึ่งข้อมูลที่ได้อาจได้มาจากฐานข้อมูลหลายๆ แห่ง โดยจะต้องทำให้ข้อมูลเหล่านั้นอยู่ในรูปแบบเดียวกันเสียก่อน

สำหรับข้อมูลที่จะนำมาใช้เป็นข้อมูลลูกค้าที่ยื่นใบสมัครขอสินเชื่อจากธนาคารแห่งหนึ่งในประเทศเยอรมัน โดยมีรายละเอียดของข้อมูลดังนี้

ชื่อข้อมูล	ประเภทข้อมูล
จำนวนเงินในบัญชีเงินฝาก	Text
ระยะเวลาในการกู้ยืม(เดือน)	Number
ประวัติการชำระเงินคืนในอดีต	Text
วัตถุประสงค์ในการขอสินเชื่อ	Text
จำนวนเงินที่ขอสินเชื่อ	Number
จำนวนเงินในบัญชีออมทรัพย์/หุ้นกู้	Text
อายุงานในสถานที่ทำงานปัจจุบัน	Text
จำนวนเงินที่ขอกู้ต่อรายได้(%)	Number
สถานะภาพสมรสและเพศ	Text
ผู้กู้ร่วม/ผู้ค้ำประกัน	Text
ระยะเวลาที่อาศัยในที่อยู่ปัจจุบัน	Number
ทรัพย์สินที่ใช้ประกอบการกู้ยืม	Text
อายุ	Number
หน่วยงานอื่นที่ผู้กู้ทำการกู้ยืม	Text
ลักษณะที่อยู่อาศัย	Text
จำนวนสินเชื่อที่มีอยู่กับธนาคาร	Number
ลักษณะการจ้างงาน	Text
จำนวนบุคคลที่อยู่ในความดูแล	Number
ข้อมูลเบอร์โทรศัพท์	Text
พลเมืองของประเทศเยอรมัน	Text
เครดิตของลูกค้า(ดีหรือไม่ดี)	Text

ตารางที่ 4.1 ตารางข้อมูลลูกค้า

รหัส	ความหมาย
A11	ต่ำกว่า 0 DM
A12	ระหว่าง 0 ถึง 200 DM
A13	200 DM ขึ้นไป
A14	ไม่มี

ตารางที่ 4.2 รายการจำนวนเงินในบัญชีเงินฝาก

รหัส	ความหมาย
A30	ไม่เคยได้รับสินเชื่อ/จ่ายคืนตรงเวลาทุกสินเชื่อ
A31	จ่ายคืนตรงเวลาทุกสินเชื่อที่มีอยู่กับธนาคาร
A32	จ่ายคืนตรงเวลาจนถึงปัจจุบัน
A33	จ่ายคืนล่าช้า(ในอดีต)
A34	Critical account/สินเชื่อที่มีอยู่กับธนาคารอื่น

ตารางที่ 4.3 รายการประวัติการชำระเงิน

รหัส	ความหมาย
A40	รถยนต์(ใหม่)
A41	รถยนต์(ผ่านการใช้งาน)
A42	เฟอร์นิเจอร์
A43	วิทยุ/โทรทัศน์
A44	เครื่องใช้เกี่ยวกับบ้าน
A45	การซ่อมบำรุง
A46	การศึกษา

ตารางที่ 4.4 รายการวัตถุประสงค์ในการขอสินเชื่อ

รหัส	ความหมาย
A47	การพักผ่อน
A48	การฝึกอบรม
A49	ธุรกิจ
A410	อื่นๆ

ตารางที่ 4.4 (ต่อ) รายการวัตถุประสงค์ในการขอสินเชื่อ

รหัส	ความหมาย
A61	ต่ำกว่า 100 DM
A62	ระหว่าง 100 ถึง 500 DM
A63	ระหว่าง 500 ถึง 1000 DM
A64	1000 DM ขึ้นไป
A65	ไม่มี

ตารางที่ 4.5 รายการจำนวนเงินในบัญชีออมทรัพย์/หุ้นกู้

รหัส	ความหมาย
A71	ไม่มี
A72	ต่ำกว่า 1 ปี
A73	ระหว่าง 1 ถึง 4 ปี
A74	ระหว่าง 4 ถึง 7 ปี
A75	7 ปีขึ้นไป

ตารางที่ 4.6 รายการอายุงานในสถานที่ทำงานปัจจุบัน

รหัส	ความหมาย
A91	ชาย :หย่า/แยกกันอยู่
A92	หญิง :หย่า/แยกกันอยู่/สมรส
A93	ชาย : โสด
A94	ชาย :สมรส/ม่าย
A95	หญิง : โสด

ตารางที่ 4.7 รายการสถานะภาพสมรสและเพศ

รหัส	ความหมาย
A101	ไม่มี
A102	ผู้กู้ร่วม
A103	ผู้ค้ำประกัน

ตารางที่ 4.8 รายการผู้กู้ร่วม/ผู้ค้ำประกัน

รหัส	ความหมาย
A121	อสังหาริมทรัพย์
A122	ประกันชีวิต
A123	รถยนต์หรืออื่นๆ
A124	ไม่มี

ตารางที่ 4.9 รายการทรัพย์สินที่ใช้ประกอบการกู้ยืม

รหัส	ความหมาย
A141	ธนาคาร
A142	ห้างร้าน
A143	ไม่มี

ตารางที่ 4.10 รายการหน่วยงานอื่นที่ผู้กู้ทำการกู้ยืม

รหัส	ความหมาย
A151	เช่า
A152	เป็นของตนเอง
A153	อยู่อาศัย

ตารางที่ 4.11 รายการลักษณะที่อยู่อาศัย

รหัส	ความหมาย
A171	ว่างงาน/พนักงานชั่วคราวที่ไม่ชำนาญงาน
A172	พนักงานประจำที่ไม่ชำนาญงาน
A173	พนักงานที่ชำนาญงาน/ข้าราชการ
A174	ผู้บริหาร/เจ้าของธุรกิจ/พนักงานที่มีใบแสดงคุณวุฒิระดับสูง

ตารางที่ 4.12 รายการลักษณะการจ้างงาน

สำหรับข้อมูลเบอร์โทรศัพท์ มีค่าที่เป็นไปได้ดังนี้

A191 = ไม่มีเบอร์โทรศัพท์

A192 = มีเบอร์โทรศัพท์

ข้อมูลที่ระบุว่าเป็นพลเมืองของประเทศเยอรมันหรือไม่ มีค่าที่เป็นไปได้ดังนี้

A201 = เป็น

A202 = ไม่เป็น

และ Target Attribute ที่จะใช้ในการสร้างแบบจำลองพยากรณ์คือ เครดิตของลูกค้า ซึ่งอยู่ในตารางข้อมูลลูกค้า โดย 1 แทน เครดิตดี และ 2 แทน เครดิตไม่ดี

จากนั้นต้องนำข้อมูลที่คัดเลือกมาทำความสะอาด เพื่อจัดการกับ Noisy data และ Missing values แต่เนื่องจากข้อมูลที่เลือกมามีความสมบูรณ์อยู่แล้ว ไม่พบ ทั้ง Noisy data และ Missing values ดังนั้นจึงไม่มีการกระทำใดๆ กับข้อมูล

จากนั้นทำการแบ่งข้อมูลออกเป็น 2 ชุด โดยชุดแรกใช้สำหรับทำการฝึกสอนเพื่อสร้างแบบจำลองพยากรณ์จำนวน 800 รายการ และชุดที่ 2 ใช้สำหรับทดสอบความถูกต้องของแบบจำลองที่ได้จำนวน 200 รายการ

เมื่อได้ทำการจัดเตรียมข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดไปเป็นการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น ได้เลือกใช้โปรแกรมเดลไฟ เวอร์ชัน 6 (Delphi Version 6) ในการเขียนโปรแกรม

4.3 การจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น

ในส่วนของการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้นประกอบด้วย 3 ขั้นตอน ดังนี้

4.3.1 การฝึกสอนระบบเพื่อสร้างแบบจำลองพยากรณ์

ในส่วนนี้จะประกอบด้วยขั้นตอนย่อยอีก 4 ขั้นตอน คือ

4.3.1.1 การนำข้อมูลเข้าระบบ

4.3.1.2 การตรวจสอบคุณภาพของข้อมูลและการจัดกลุ่ม

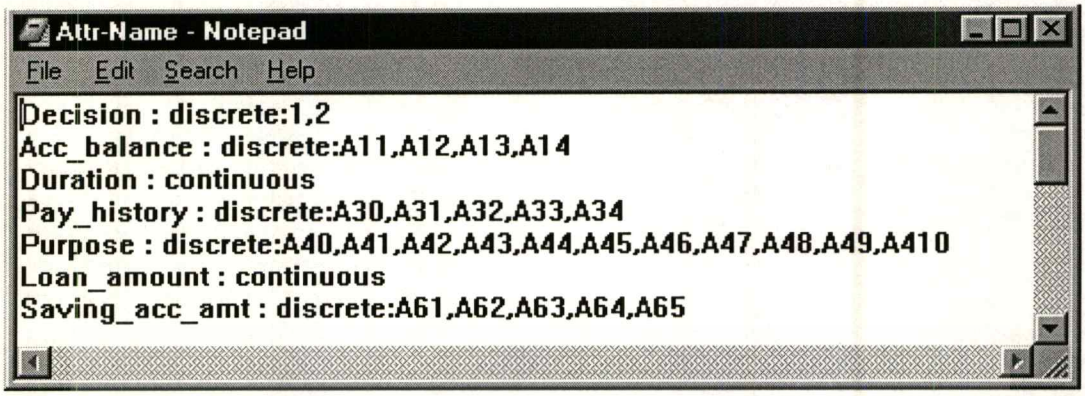
4.3.1.3 การกำหนดเงื่อนไขให้กับโปรแกรม

4.3.1.4 การแสดงผล

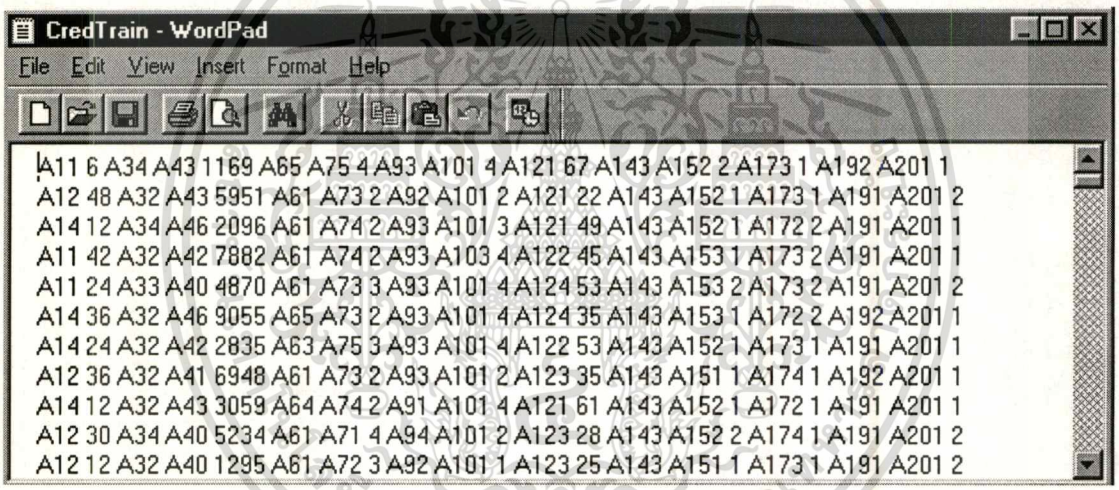
4.3.1.1 การนำข้อมูลเข้าระบบ

การนำข้อมูลเข้าระบบนี้สามารถรับข้อมูลเข้าได้ 2 แบบ คือ ข้อมูลที่อยู่ในรูปแบบฐานข้อมูลและข้อมูลรูปแบบเท็กซ์ไฟล์ (Text File) โดยในกรณีที่ข้อมูลเข้าเป็นเท็กซ์ไฟล์ จะประกอบด้วยไฟล์นามสกุล .nam, .txt และ .tst โดย

- .nam ใช้เก็บข้อมูลที่เป็นชื่อแอททริบิวของข้อมูล, ประเภทของข้อมูลว่าเป็นข้อมูลที่เป็น Continuous หรือ Dicrete และค่าที่เป็นไปได้ในแอททริบิวนั้นๆ โดยแอททริบิวแรกจะเป็นแอททริบิวเป้าหมาย แสดงตัวอย่าง ของข้อมูลดังภาพที่ 4.1
- .dat ใช้เก็บข้อมูลเรียงลำดับตามชื่อแอททริบิวที่อยู่ในไฟล์ .nam ยกเว้นแอททริบิวเป้าหมายให้อยู่ในลำดับสุดท้ายของรายการ แสดงตัวอย่างของข้อมูลดังภาพที่ 4.2



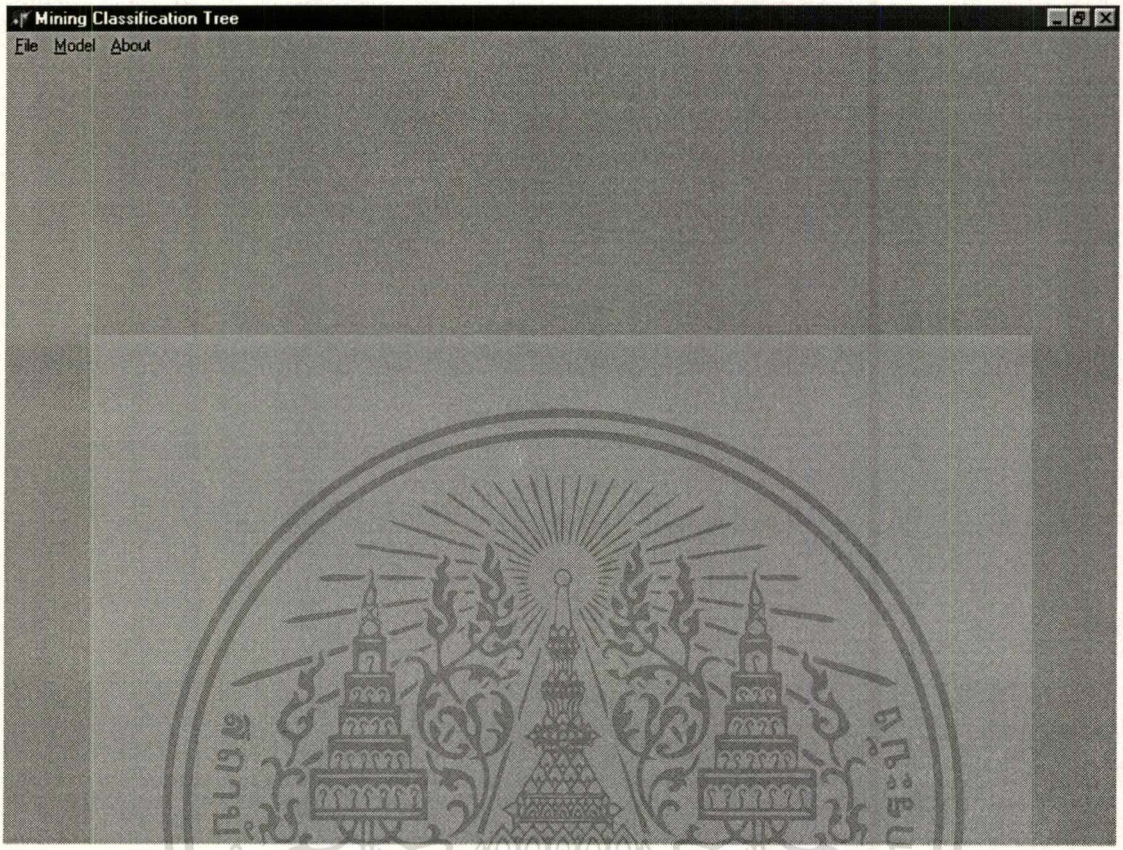
ภาพที่ 4.1 ตัวอย่างไฟล์นามสกุล .nam



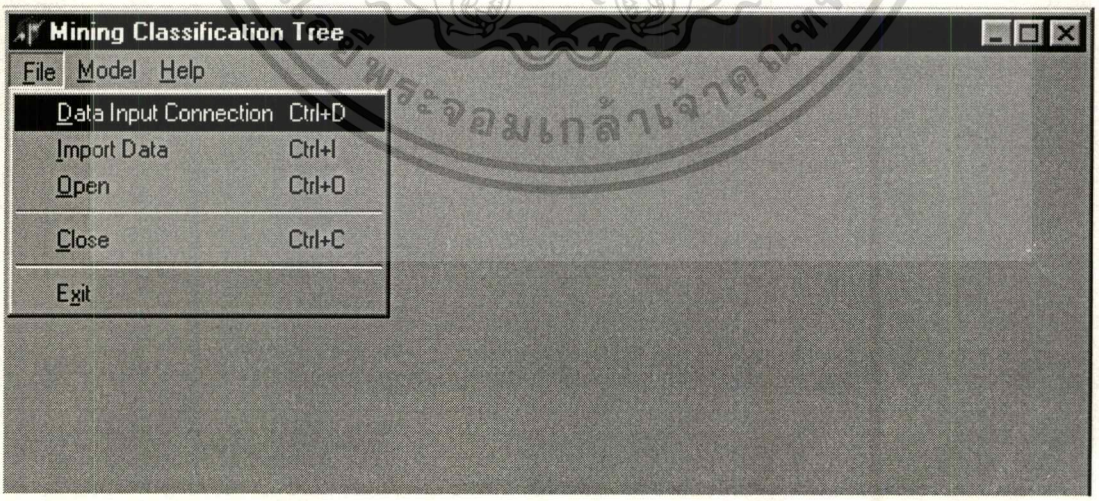
ภาพที่ 4.2 ตัวอย่างไฟล์นามสกุล .dat

ในส่วนของใช้งานโปรแกรมเพื่อนำข้อมูลเข้าระบบมีขั้นตอนการใช้งานคือ เมื่อเข้าสู่โปรแกรม จะปรากฏเมนูหลัก โดยในส่วนของเมนูหลักจะประกอบด้วย เมนู File ซึ่งใช้สำหรับนำข้อมูลเข้าระบบและนำโครงสร้างของดิสิชั่นทรี (Decision Tree) ที่ได้สร้างไว้แล้วมาแสดงผล และเมนู Model ซึ่งใช้สำหรับนำแบบจำลองพยากรณ์มาพยากรณ์กลุ่มของข้อมูล ดังภาพที่ 4.3

จากเมนูหลักเลือกเมนูย่อย จะพบเมนูย่อยในการนำข้อมูลเข้า 2 เมนูคือ Data Input Connection สำหรับติดต่อกับฐานข้อมูล และ Import Data สำหรับข้อมูลเข้าที่อยู่ในรูปของเท็กซ์ไฟล์ (Text File) ในที่นี้จะทำการวิเคราะห์ข้อมูลที่เป็นฐานข้อมูล จึงเลือกเมนูย่อย Data Input Connection ดังหน้าจอตัวอย่างภาพที่ 4.4

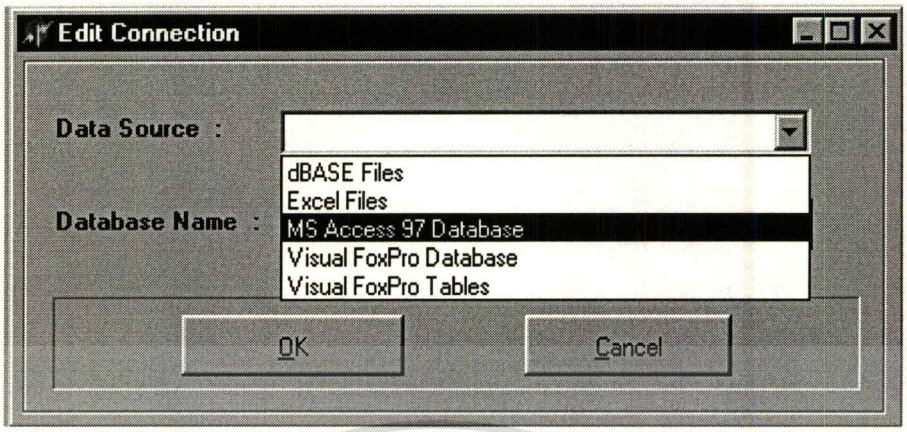


ภาพที่ 4.3 หน้าจอหลักของระบบ



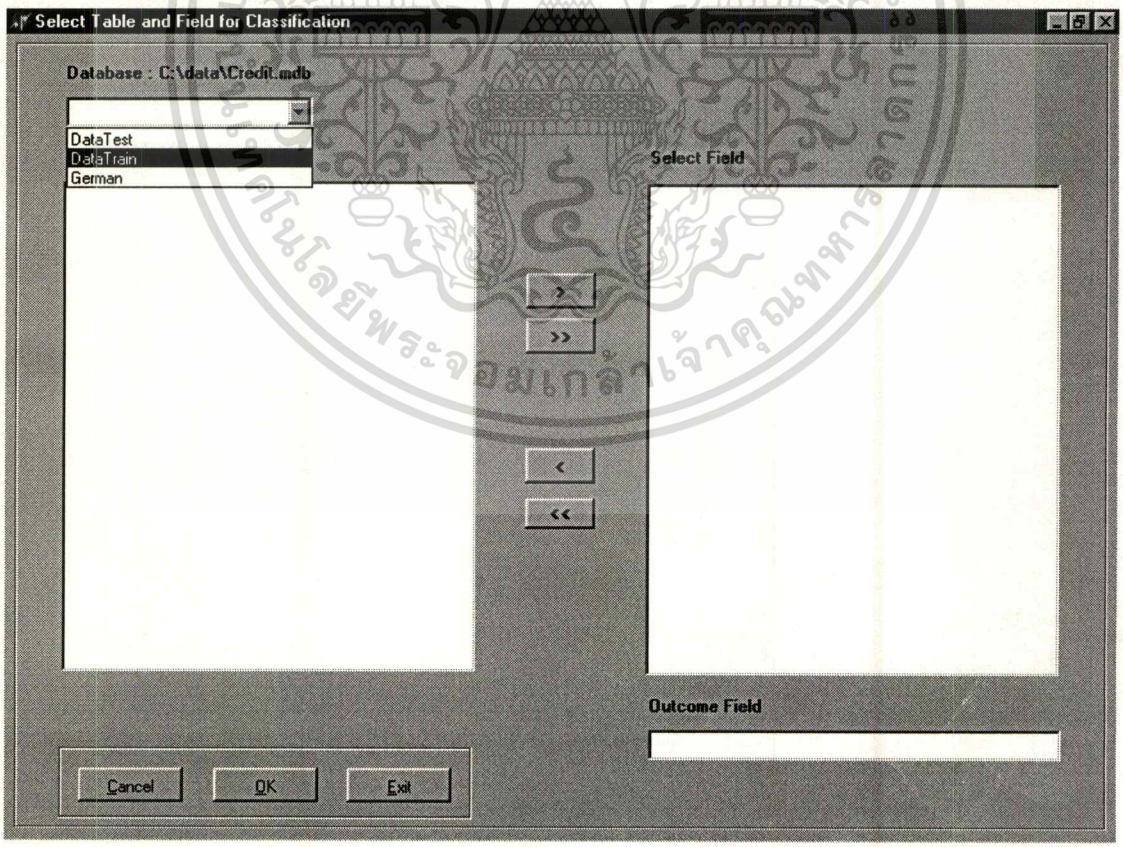
ภาพที่ 4.4 หน้าจอแสดงการเลือกประเภทข้อมูลที่ใช้ในการวิเคราะห์

จากนั้นจะปรากฏหน้าจอให้เลือกวิธีติดต่อกับฐานข้อมูลดังภาพที่ 4.5



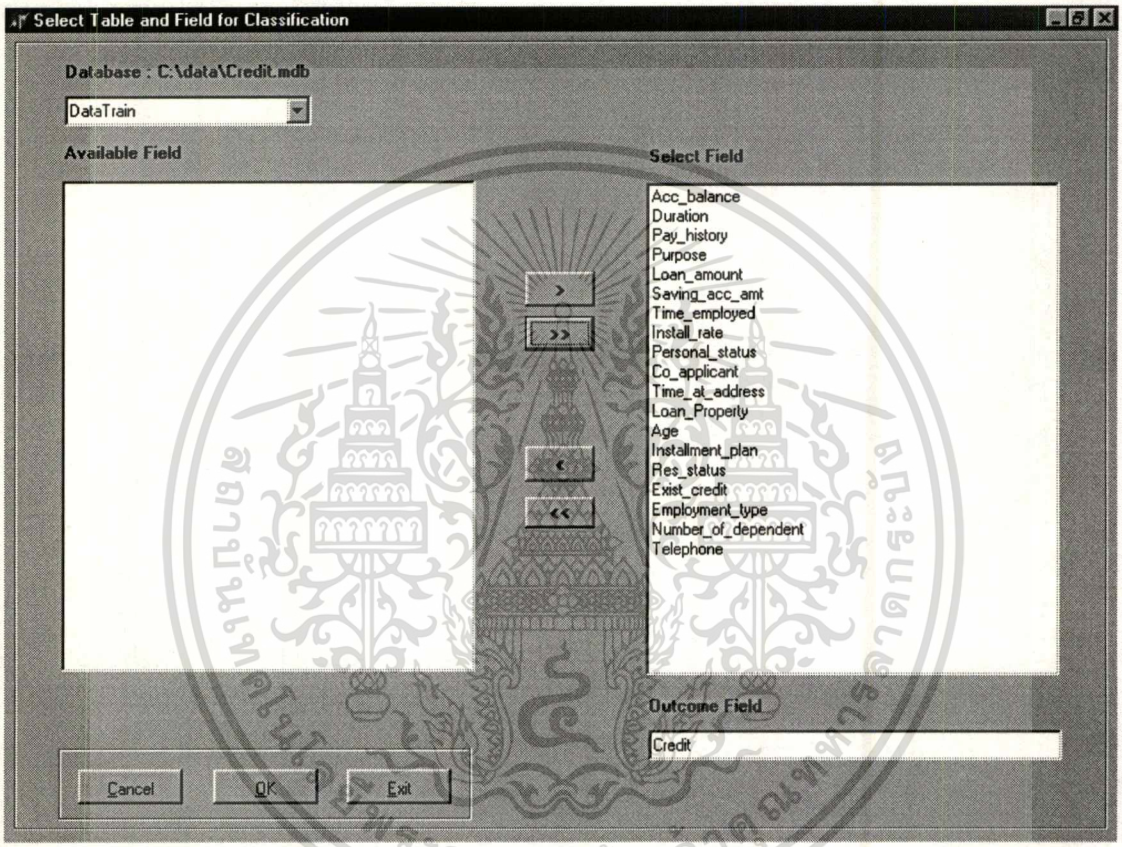
ภาพที่ 4.5 หน้าจอแสดงการเลือกประเภทฐานข้อมูลที่ใช้ในการวิเคราะห์

หลังจากเลือกฐานข้อมูลที่จะใช้วิเคราะห์แล้ว จะปรากฏหน้าจอแสดงรายละเอียดของตารางมาให้เลือก ว่าต้องการใช้ตารางใดในการวิเคราะห์ดังภาพที่ 4.6



ภาพที่ 4.6 หน้าจอแสดงการเลือกตารางมาวิเคราะห์

หลังจากเลือกตารางที่จะใช้วิเคราะห์เสร็จแล้ว จะเป็นการเลือกแอททริบิวที่จะนำมาใช้ในการวิเคราะห์ โดยระบบสามารถให้เลือกได้ว่าต้องการใช้แอททริบิวใดบ้าง โดยต้องระบุทั้งแอททริบิวที่เลือกใช้งานในส่วนของ Select Field และแอททริบิวเป้าหมายในส่วนของ Outcome Field แสดงตัวอย่างหน้าจอได้ดังภาพที่ 4.7



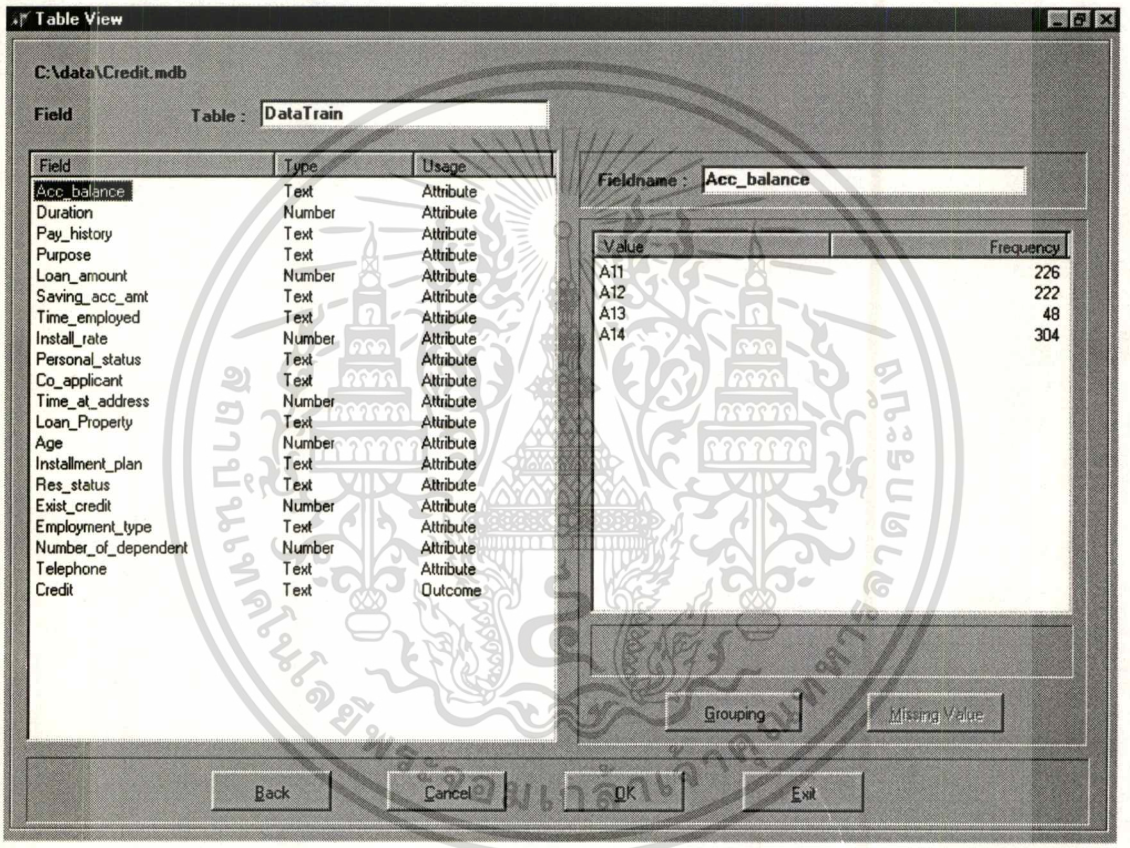
ภาพที่ 4.7 หน้าจอแสดงการเลือกแอททริบิวที่นำมาวิเคราะห์

4.3.1.2 การตรวจสอบคุณภาพของข้อมูลและการจัดกลุ่ม

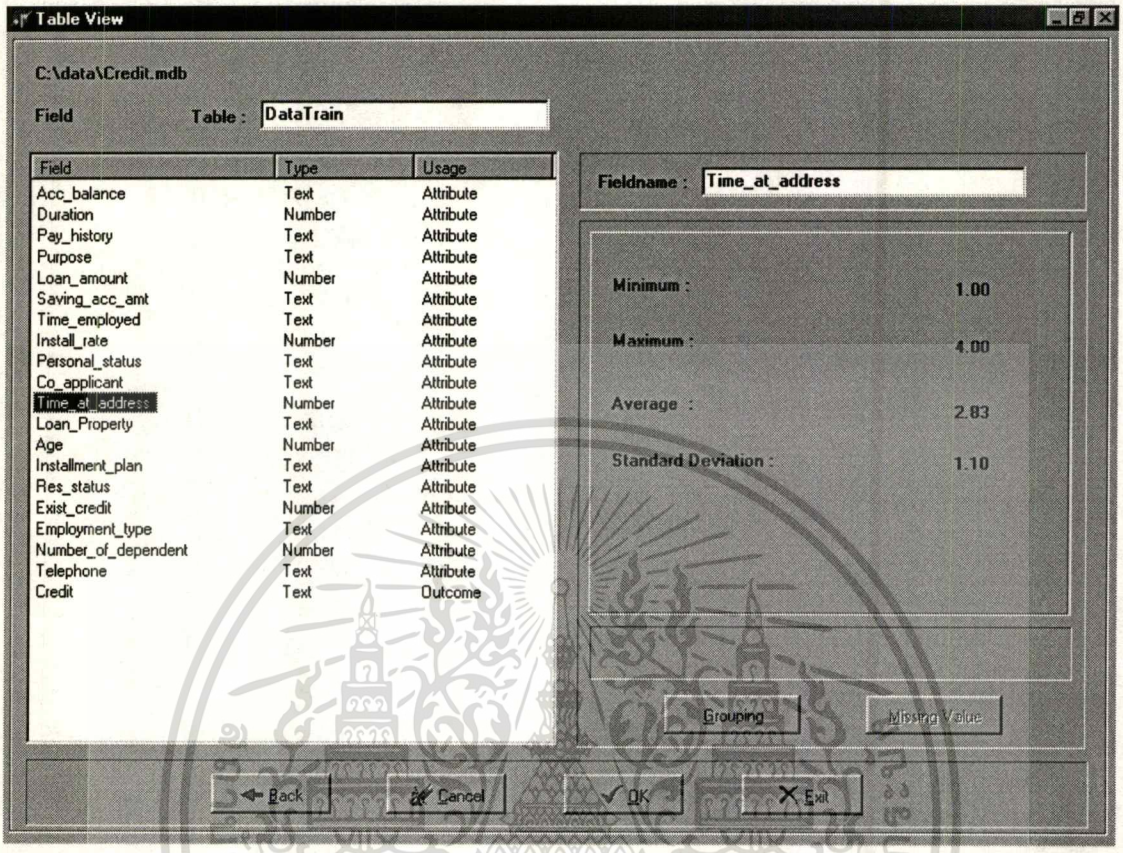
หลังจากนำข้อมูลเข้าระบบเสร็จแล้ว จะเข้าสู่กระบวนการของการตรวจสอบคุณภาพของข้อมูล การจัดการกับค่าที่หายไป (Missing Value) และการจัดกลุ่มข้อมูล โดยระบบจะแสดงรายละเอียดของข้อมูลในแต่ละแอททริบิว โดยถ้าเป็นแอททริบิวที่มีชนิดของข้อมูลเป็นข้อความ (Text) จะแสดงค่าของข้อมูลและค่าความถี่ที่เกิดขึ้นของข้อมูลในแอททริบิวนั้นๆ แสดงตัวอย่างหน้าจอดังภาพที่ 4.8 ส่วนแอททริบิวที่มีชนิดข้อมูลเป็นตัวเลข(Number) จะแสดงค่าของสูงสุด, ต่ำสุด, ค่าค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของข้อมูล และแสดงตัวอย่างหน้าจอดังภาพที่ 4.9

ในกรณีที่แอททริบิวใดมีค่าของข้อมูลที่ขาดหายไป ระบบจะแสดงข้อความเตือน ซึ่งสามารถที่จะลบเรคคอร์ดที่ประกอบด้วยค่าว่างนี้ หรือจะจัดรวมกับค่าอื่น หรือจะแทนด้วยค่าใหม่ แสดงตัวอย่างหน้าจอดังภาพที่ 4.10

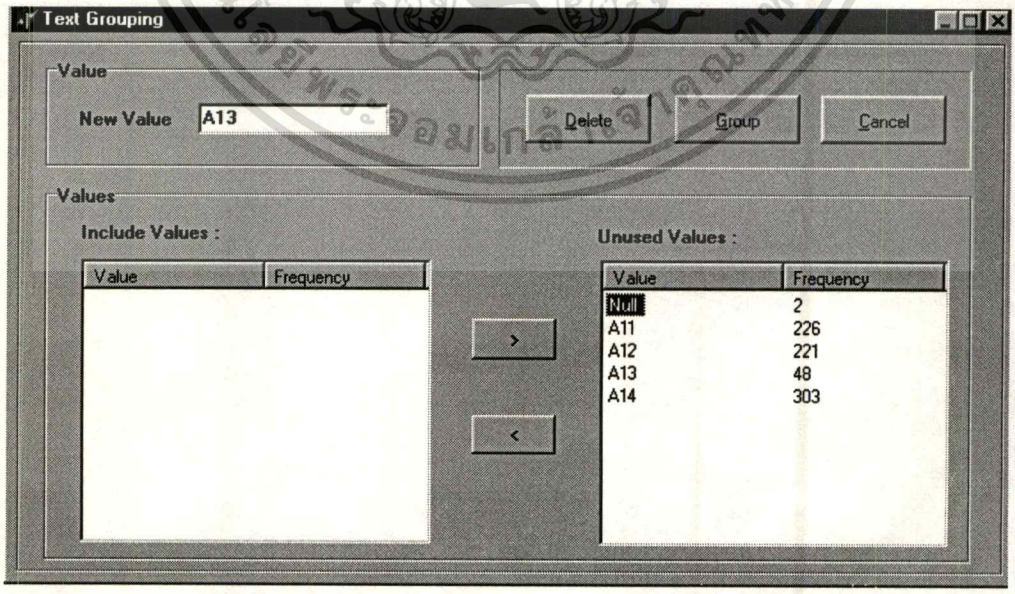
สำหรับข้อมูลที่มีค่าความถี่ที่เกิดขึ้นน้อย สามารถที่จะจัดกลุ่มรวมกับค่าอื่นได้ โดยเลือกปุ่ม Grouping แสดงตัวอย่างหน้าจอดังภาพที่ 4.11



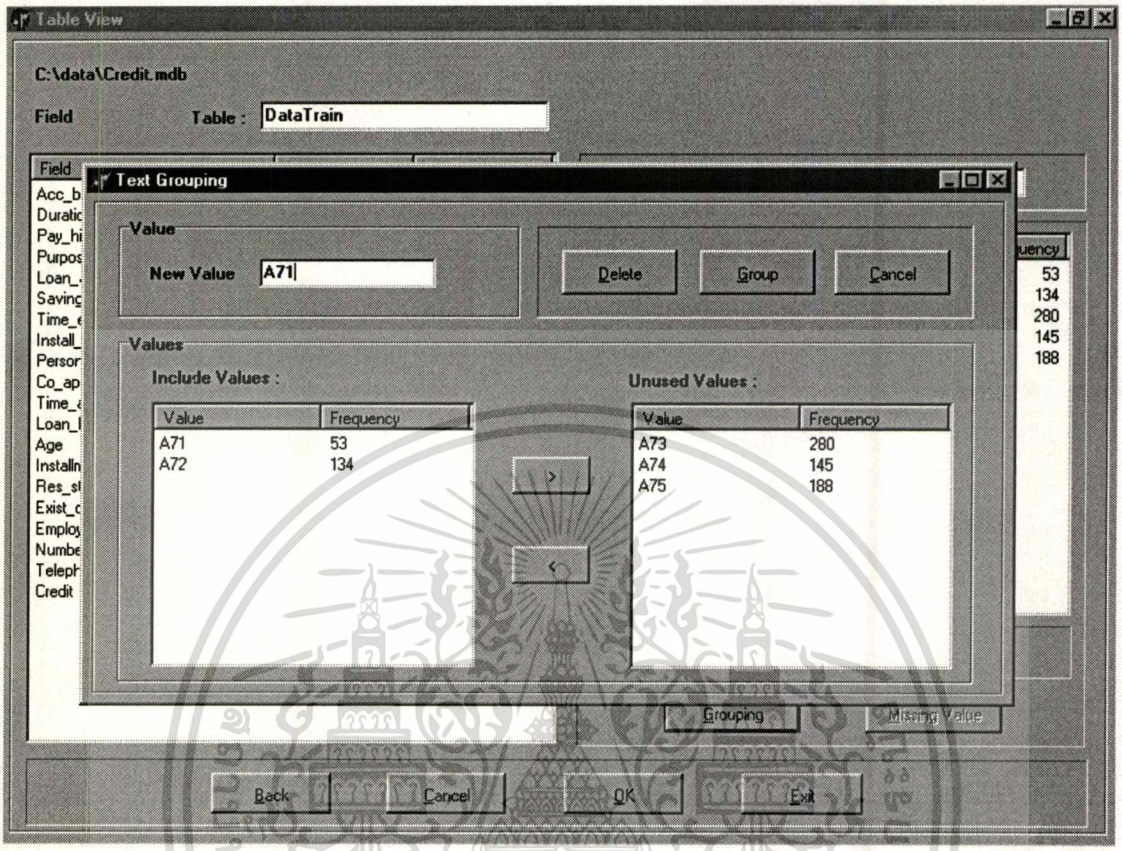
ภาพที่ 4.8 หน้าจอแสดงรายละเอียดของแอทริบิวชนิดข้อความ



ภาพที่ 4.9 หน้าจอแสดงรายละเอียดของแอทริบิวชนิดตัวเลข



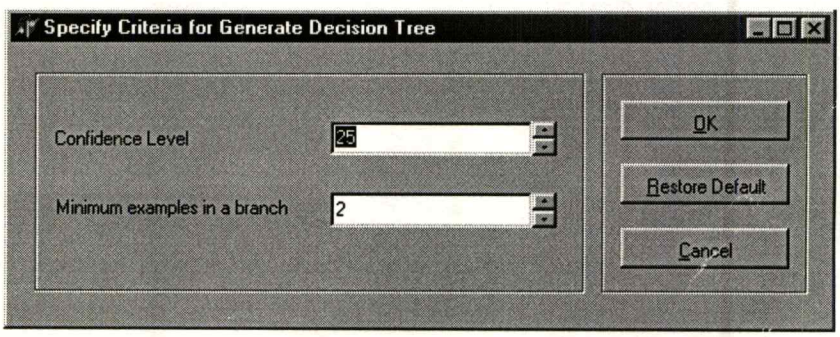
ภาพที่ 4.10 หน้าจอแสดงการจัดการกับข้อมูลที่มีค่าที่หายไป



ภาพที่ 4.11 หน้าจอแสดงการจัดกลุ่มข้อมูล

4.3.1.3 การกำหนดเงื่อนไขให้กับโปรแกรม

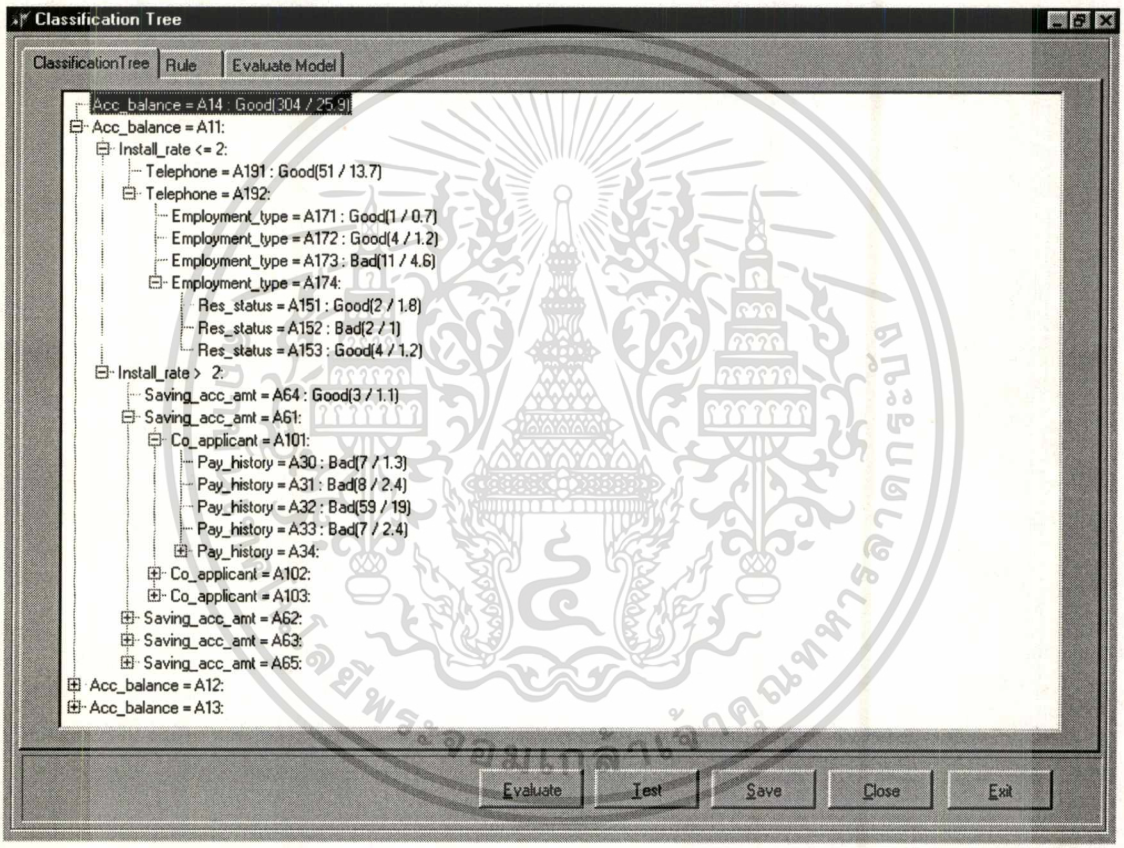
หลังจากที่ทำการตรวจสอบคุณภาพของข้อมูลเสร็จแล้ว จะเข้าสู่ขั้นตอนของการกำหนดค่า Confidence Level และ Minimum example เพื่อกำหนดเงื่อนไขในการสร้างดิสิชั่นทรีและกฎ ดังภาพที่ 4.12



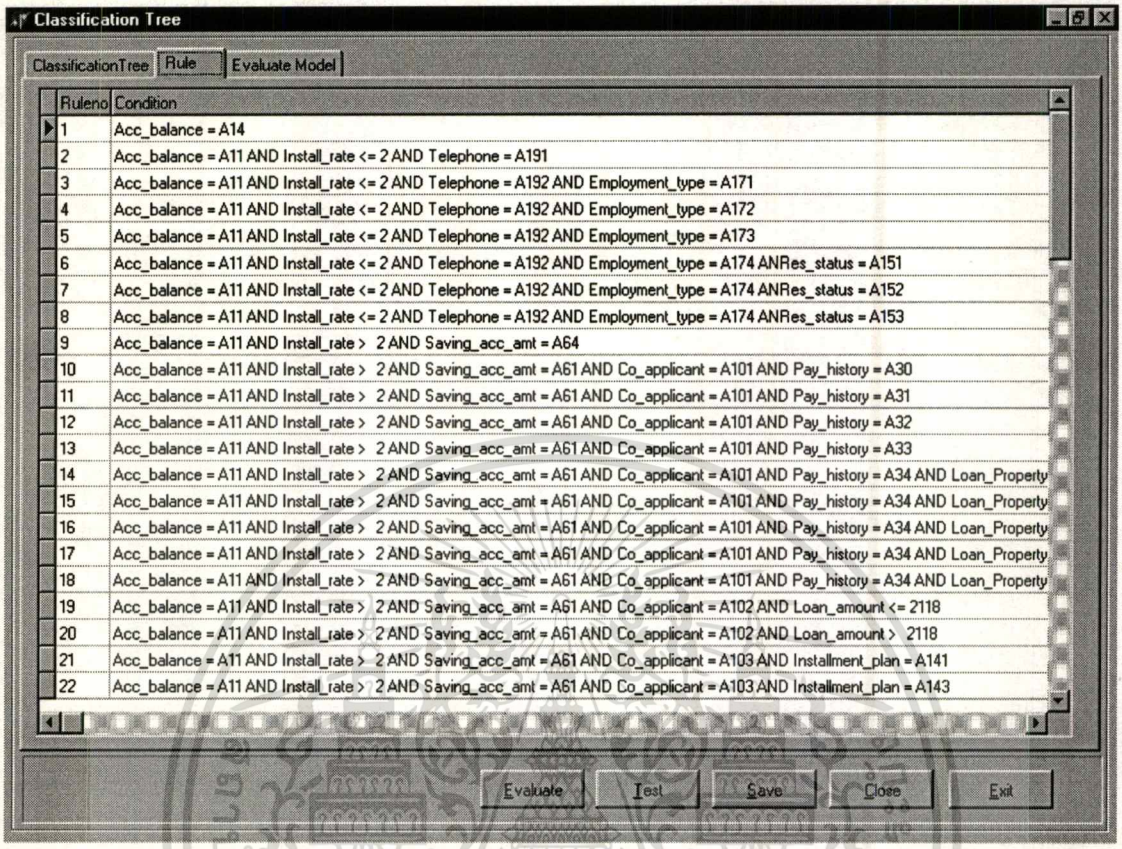
ภาพที่ 4.12 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

4.3.1.4 การแสดงผล

เมื่อโปรแกรมทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงในภาพที่ 4.13 และ 4.14 ตามลำดับ โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใดและข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภทที่ (Class) ที่ข้อมูลส่วนใหญ่ในโหนดนั้นตกอยู่ โดยสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้ โดยการเลือกปุ่ม Save



ภาพที่ 4.13 หน้าจอแสดงผลลัพธ์ในรูปดิสิชั่นทรี



ภาพที่ 4.14 หน้าจอแสดงผลลัพธ์ในรูปของกฎ

4.3.2 การทดสอบความถูกต้องของแบบจำลองพยากรณ์

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูลที่ใช้ฝึกสอนแล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากเพียงใด โดยการนำข้อมูลอีกชุดหนึ่งมาทำการทดสอบกับแบบจำลองพยากรณ์ที่ได้ โดยเลือกปุ่ม Evaluate เพื่อให้ระบบแสดงความถูกต้องของแบบจำลองโดยใช้ข้อมูลชุดฝึกสอน และเลือกปุ่ม Test เพื่อทำการทดสอบแบบจำลองโดยใช้ข้อมูลทดสอบ โดยความถูกต้องของระบบสามารถคำนวณได้จาก

$$\text{Accuracy} = \text{Sensitivity} \times (\text{pos}/(\text{pos} + \text{neg})) + \text{Specificity} \times (\text{neg}/(\text{pos} + \text{neg}))$$

โดย $\text{Sensitivity} = \text{t\_pos}/\text{pos}$

$\text{Specificity} = \text{t\_neg}/\text{neg}$

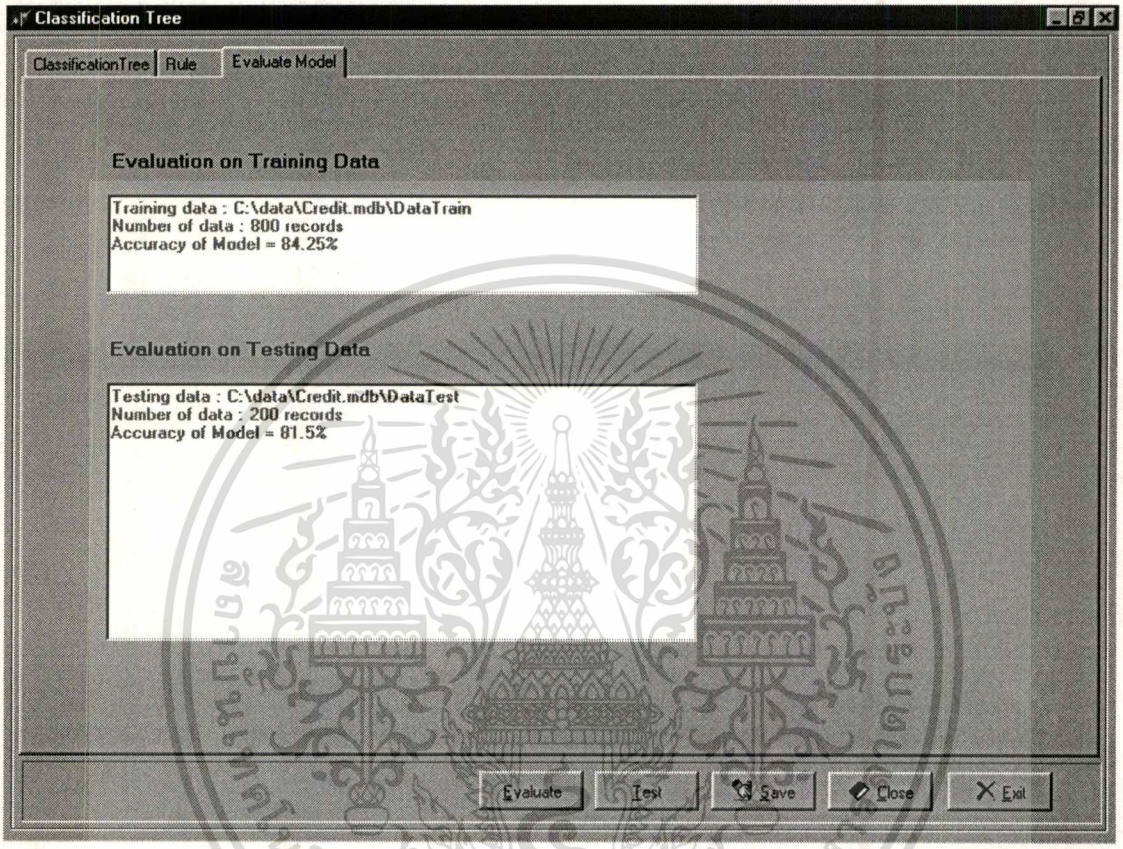
t_pos แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive ที่สามารถทำนายกลุ่มได้ถูกต้อง

pos แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive

t_neg แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative ที่สามารถทำนายกลุ่มได้ถูกต้อง

neg แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative

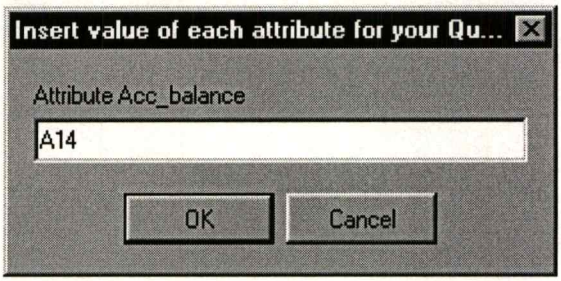
โดยหลังจากสั่งให้ระบบทำการตรวจสอบความถูกต้องของแบบจำลองพยากรณ์ ระบบจะแสดงผลการทดสอบ ดังภาพที่ 4.15



ภาพที่ 4.15 หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

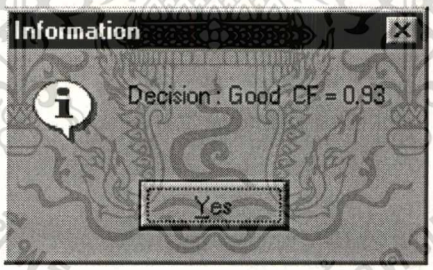
4.3.3 การนำแบบจำลองพยากรณ์มาใช้จัดกลุ่มข้อมูล

เมื่อทำการสร้างแบบจำลองและทดสอบความถูกต้องจนได้ผลอยู่ในระดับที่พึงพอใจแล้ว ผู้ใช้สามารถสอบถามเกี่ยวกับข้อมูลของผู้ใช้ว่าจัดอยู่ในกลุ่มใดได้ โดยเลือกเมนูย่อย Predict Class ในหน้าจอหลักของระบบ จากนั้นจะปรากฏหน้าต่างให้ใส่ข้อมูลในแต่ละแอททริบิวดังภาพที่ 4.16



ภาพที่ 4.16 หน้าต่างสำหรับใส่ข้อมูลเพื่อสอบถามกลุ่มของข้อมูล

โดยผู้ใช้ไม่ต้องใส่ข้อมูลทุกแอททริบิว แต่จะไล่ไปตามทรีจากรากไปยังปลาย และในกรณีที่ผู้ใช้ไม่ทราบค่าในแอททริบิวใดก็สามารถเว้นว่างไว้หรือใส่เป็น "-" ก็ได้ เมื่อถึงส่วนปลายของทรีซึ่งระบุกลุ่มที่คาดว่าข้อมูลจะจัดอยู่ ระบบก็จะแสดงผลการทำนายว่าข้อมูลควรจะจัดอยู่ในกลุ่มใดและมี Certainty Factor(CF) เป็นเท่าใด โดยค่า CF คือค่าความน่าจะเป็นที่ข้อมูลจะตกอยู่ในกลุ่มนั้นๆ ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1 ดังภาพที่ 4.17



ภาพที่ 4.17 หน้าต่างแสดงผลการทำนายกลุ่มของข้อมูล

4.4 วิเคราะห์ผลการดำเนินงาน

ผลจากการทดสอบแบบจำลองพยากรณ์ที่ได้โดยใช้ข้อมูลชุดทดสอบจำนวน 200 รายการ สามารถวัดความถูกต้องของแบบจำลองการจัดหมวดหมู่สำหรับการทำนายข้อมูลในแต่ละกลุ่ม โดยสรุปรายการที่แบบจำลองสามารถทำนายได้ถูกต้องคิดเป็น 81.5 เปอร์เซ็นต์ของข้อมูลทั้งหมด และจากการวิเคราะห์พบว่าความผิดพลาดจำนวน 18.5 เปอร์เซ็นต์นี้อาจเกิดจากข้อมูลมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ เนื่องจากข้อมูลชุดนี้เป็นข้อมูลชุดทดลองของธนาคารแห่งหนึ่งในประเทศเยอรมันที่เตรียมไว้สำหรับทำการศึกษาการทำงานของอัลกอริทึมต่างๆ ทางด้านดาต้าไมนิ่งเท่านั้น ซึ่งอาจส่งผลให้ข้อมูลไม่สามารถเป็นตัวแทนทางสถิติได้อย่างสมบูรณ์

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

โครงงานพัฒนาระบบฉบับนี้มีวัตถุประสงค์หลักเพื่อที่จะนำเสนอและประยุกต์ใช้คาต้าไมนิ่งในธุรกิจ ซึ่งดาต้าไมนิ่งนั้นเป็นกระบวนการที่ใช้เพื่อค้นหาข้อมูลที่มีประโยชน์ออกจากฐานข้อมูล เพื่อนำมาช่วยในการตัดสินใจ ซึ่งวิธีการแก้ปัญหาด้วยดาต้าไมนิ่งนั้นมีอยู่ด้วยกันหลายรูปแบบขึ้นอยู่กับวัตถุประสงค์ของการทำงาน โดยในโครงงานนี้ได้เสนอเทคนิคการสร้างแบบจำลองพยากรณ์ (Predictive Modeling) เพื่อทำนายเครดิตของลูกค้าที่ยื่นใบสมัครขอสินเชื่อจากธนาคาร โดยมีวัตถุประสงค์เพื่อเพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้ประกอบการตัดสินใจในการอนุมัติสินเชื่อ เพื่อลดความเสี่ยงในการอนุมัติสินเชื่อให้แก่ลูกค้า รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้เพื่อพิจารณาปัจจัยอื่นๆ ที่มีผลต่อการดำเนินธุรกิจ โดยใช้อัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมของ Classification Tree ที่มีการใช้งานกันอย่างกว้างขวาง อันเนื่องมาจากความมีประสิทธิภาพในการแก้ปัญหา นอกจากนี้ยังมีความยืดหยุ่นและเข้าใจได้ง่าย

ผลจากการศึกษาทำให้ได้ระบบที่ใช้สำหรับจัดกลุ่มของข้อมูล ซึ่งสามารถนำไปประยุกต์ใช้ได้กับทุกธุรกิจ และจากการนำข้อมูลเข้าไปสร้างแบบจำลองพยากรณ์และทำการทดสอบพบว่าแบบจำลองพยากรณ์ที่ได้มีความถูกต้องคิดเป็น 81.5 เปอร์เซ็นต์ของข้อมูลที่ใช้ทดสอบ และจากการวิเคราะห์พบว่าความผิดพลาดจำนวน 18.5 เปอร์เซ็นต์นี้เอาจเกิดจากข้อมูลมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ เนื่องจากข้อมูลชุดนี้เป็นข้อมูลชุดทดลองของธนาคารแห่งหนึ่งในประเทศเยอรมันที่เตรียมไว้สำหรับทำการศึกษาการทำงานของอัลกอริทึมต่างๆ ทางด้านดาต้าไมนิ่งเท่านั้น ซึ่งอาจส่งผลให้ข้อมูลไม่สามารถเป็นตัวแทนทางสถิติได้ทั้งหมด

5.2 ข้อเสนอแนะ

ระบบนี้สามารถที่จะลดขั้นตอนการทำงานลงได้ด้วยการใช้ข้อมูลจาก Data Warehouse เนื่องจากเป็นข้อมูลที่ได้รับการทำความสะอาดมาเรียบร้อยแล้ว ทำให้สามารถลดขั้นตอนในการทำความสะอาดข้อมูลลงได้ ทั้งยังให้ความมั่นใจในความถูกต้องมากขึ้นด้วย

ระบบที่พัฒนาขึ้นนี้สามารถนำไปใช้กับข้อมูลในธุรกิจอื่นได้ เนื่องจากไม่ได้จำกัดขอบเขตไว้กับธุรกิจการธนาคารเท่านั้น

ข้อมูลที่นำมาใช้ในการทดสอบโปรแกรมอาจมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ เนื่องจากข้อมูลชุดนี้เป็นข้อมูลชุดทดลองของธนาคารแห่งหนึ่งในประเทศเยอรมันที่เตรียมไว้สำหรับทำการศึกษาการทำงานของอัลกอริทึมต่างๆ ทางด้านดาต้าไมนิ่งเท่านั้น

จากการนำอัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมของ Classification Tree พบว่ากฎที่ได้สามารถเข้าใจได้ง่าย และสามารถบอกได้ว่าปัจจัยใดที่มีอิทธิพลต่อการทำนายมากที่สุด แต่อย่างไรก็ตามวิธีนี้ก็มีข้อเสียคือ ถ้าข้อมูลมีจำนวนน้อย ความผิดพลาดในการทำนายจะสูงขึ้น เนื่องจากเมื่อทำการแตก Tree ไปเรื่อยๆ จำนวนของข้อมูลจะลดลง ซึ่งข้อมูลจำนวนน้อยจะเป็นตัวแทนทางสถิติได้ยาก โดยยิ่งทำให้มี level มากขึ้นความน่าเชื่อถือจะยิ่งน้อยลง

บรรณานุกรม

Artificial Intelligence and Computer Science Laboratory. **German credit dataset.** [Online]. Available : http://www.ncc.up.pt/liacc/ML/statlog/datasets/german.

Building Classification Models: ID3 and C4.5. [Online]. Available : http://yoda.cis.temple.edu:8080/UGAIWWW/lectures95/learn/c45/.

Han Jiawei and Kamber Micheline. 2001. **Data Mining: Concepts and Techniques.** California : Morgan Kaufmanns.

Machine Learning. [Online]. Available : http://www.cs.uregina.ca/~dbd/cs831/ notes/ml/dtrees/c4.5/tutorial.html.

Pruning. [Online]. Available : http://www.cs.bgu.ac.il/~westrich/decision/pruning.html.

Quinlan, J. R. 1993. **C4.5: Programs for Machine Learning.** California : Morgan Kaufmann.

Simoudis, E. 1998. **Discovering Data Mining From Concept to implementation.** New Jersey : Prentice Hall.

ภาคผนวก ก

คู่มือการใช้ระบบอนุมัติสินเชื่อเบื้องต้นโดยใช้อัลกอริทึม C4.5

ก.1 ความต้องการของระบบ

1) ความต้องการทางด้านซอฟต์แวร์

ซอฟต์แวร์ที่ต้องการใช้ มีดังนี้

(1.1) ระบบปฏิบัติการวินโดวส์ 95 ขึ้นไป

(1.2) โปรแกรมบอร์แลนด์เดลไฟ 6.0

(1.3) โปรแกรมไมโครซอฟต์แอกเซส เวอร์ชัน 97

2) ความต้องการทางด้านฮาร์ดแวร์

ความต้องการทางด้านฮาร์ดแวร์ สามารถสรุปได้ดังนี้

(2.1) เครื่องคอมพิวเตอร์ที่มีโปรเซสเซอร์เพนเทียมทรี ขึ้นไป

(2.2) เนื้อที่ว่างบนฮาร์ดดิสก์อย่างน้อย 2 เมกะไบต์

(2.3) หน่วยความจำสำรอง อย่างน้อย 128 เมกะไบต์

(2.4) จอภาพชนิด Super VGA ความละเอียดของจอภาพอย่างน้อย 256 สี

ก.2 การทำงานของระบบอนุมัติสินเชื่อเบื้องต้นโดยใช้อัลกอริทึม C4.5

เมื่อเริ่มต้นทำงานโปรแกรมจะแสดงหน้าจอแรกดังภาพที่ ก.1 โดยประกอบด้วยเมนู (Menu) ดังนี้

1) เมนู File

ประกอบด้วยเมนูย่อย 5 เมนู คือ

1.1) เมนู Data Input Connection ใช้สำหรับเลือกติดต่อกับข้อมูลที่นำมาวิเคราะห์ที่เป็นฐานข้อมูล

1.2) เมนู Import Data ใช้สำหรับติดต่อกับข้อมูลที่นำมาวิเคราะห์ที่เป็นเท็กซ์ไฟล์ (Text File)

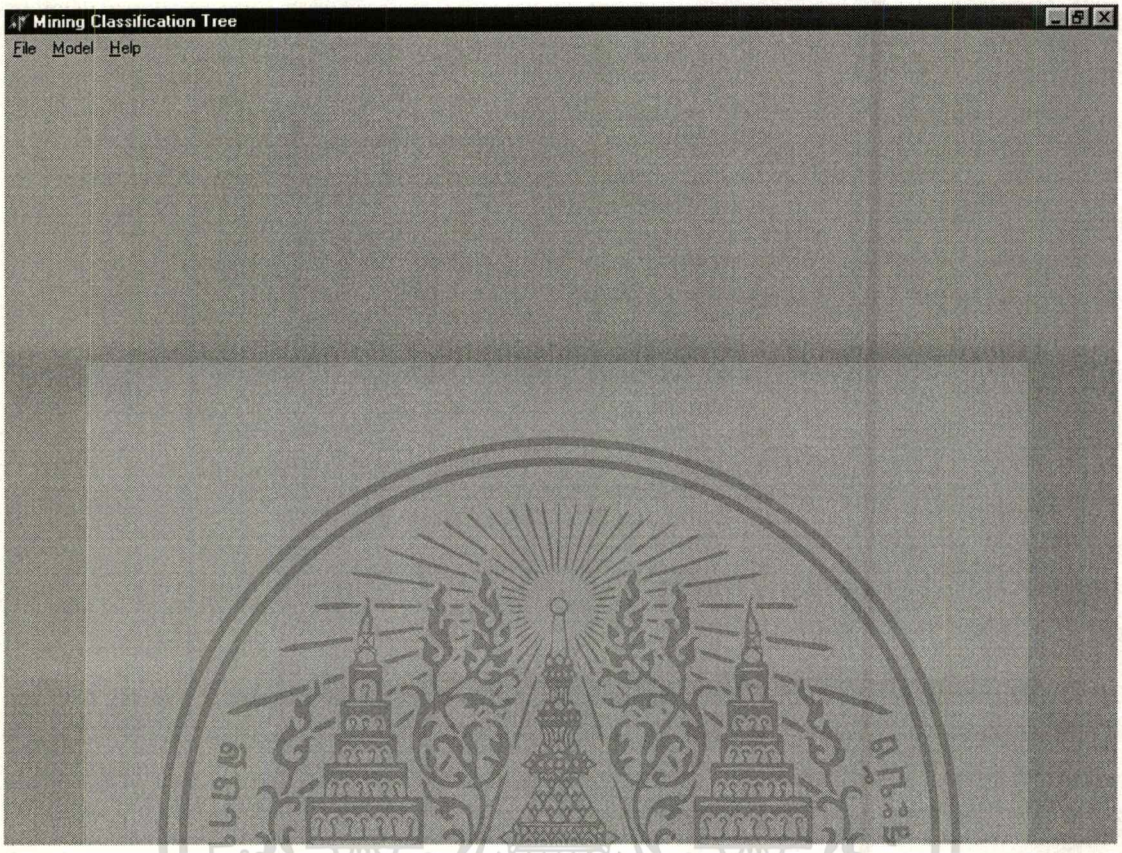
1.3) เมนู Open ใช้เพื่อเรียกดูโครงสร้างของดิสชั่นทรีที่ทำการบันทึกไว้แล้ว

1.4) เมนู Close ใช้เพื่อปิดหน้าต่างที่เปิดอยู่ขณะนั้น

1.5) เมนู Exit ใช้เพื่อออกจากระบบ

2) เมนู Model ประกอบด้วยเมนูย่อย Predict Class ใช้เพื่อทำการทำนายกลุ่มของข้อมูล

3) เมนู Help



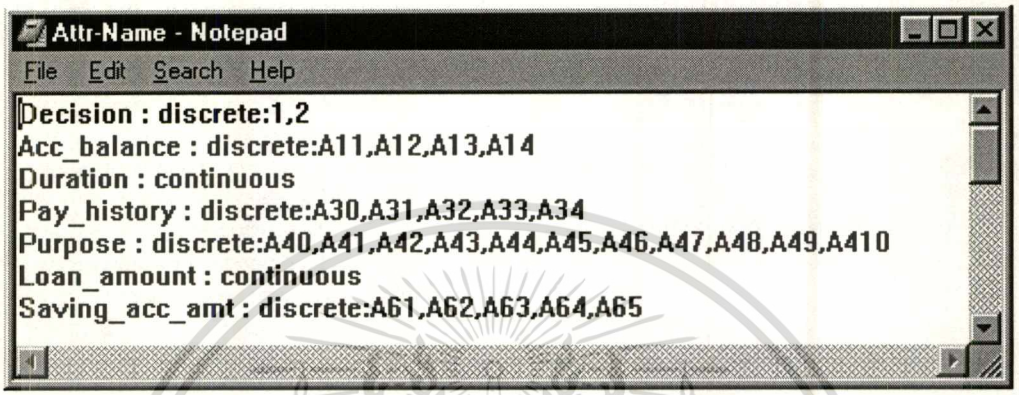
ภาพที่ ก.1 หน้าจอแรกของระบบ

สำหรับโปรแกรมที่พัฒนาขึ้น สามารถติดต่อกับข้อมูลที่นำมาวิเคราะห์ได้ 2 รูปแบบคือ ข้อมูลที่เป็นฐานข้อมูลและ ข้อมูลที่เป็นเท็กซ์ไฟล์

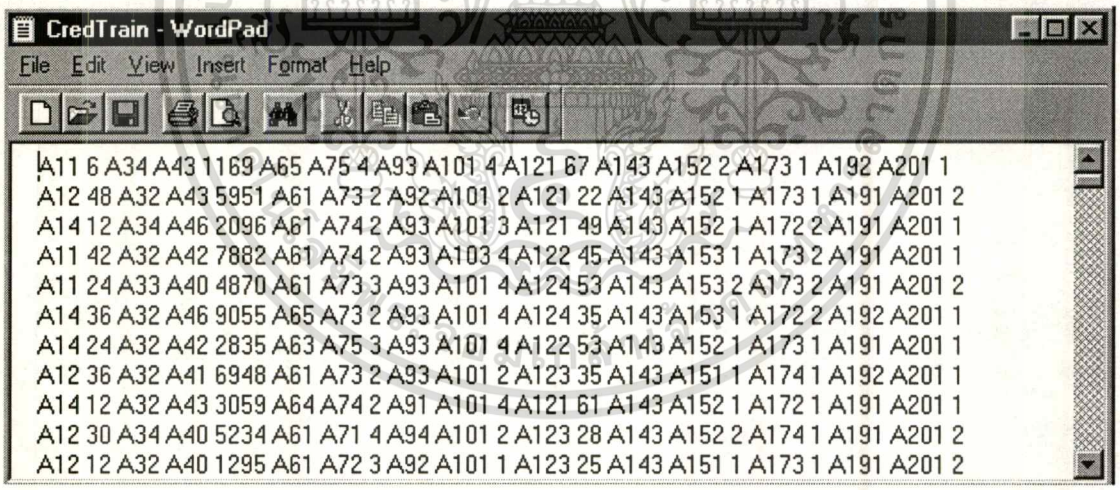
- กรณีที่เป็นการเตรียมข้อมูลด้วยตาราง (Table) จะคล้ายกับฐานข้อมูลของระบบ Relational Database ทั่ว ๆ ไป
- กรณีที่เป็นการเตรียมข้อมูลด้วยเท็กซ์ไฟล์ (Text File) จะประกอบด้วยไฟล์นามสกุล .nam และ .dat โดย

  .nam เก็บข้อมูลที่เป็นชื่อแอททริบิวของข้อมูล, ประเภทของข้อมูลว่าเป็นข้อมูลที่เป็น Continuous หรือ Dicrete และค่าที่เป็นไปได้ในแอททริบิวนั้นๆ โดยแอททริบิวแรกจะเป็นแอททริบิวเป้าหมาย (Target Attribute) แสดงตัวอย่างของข้อมูลดังภาพที่ ก.2

.dat เก็บข้อมูลเรียงลำดับตามชื่อแอททริบิวที่อยู่ในไฟล์ .nam ยกเว้นแอททริบิวเป้าหมายให้เป็นแอททริบิวสุดท้ายของแต่ละรายการ โดยสามารถคั่นด้วยเครื่องหมายอะไรก็ได้ แสดงตัวอย่างของข้อมูลดังภาพที่ ก.3



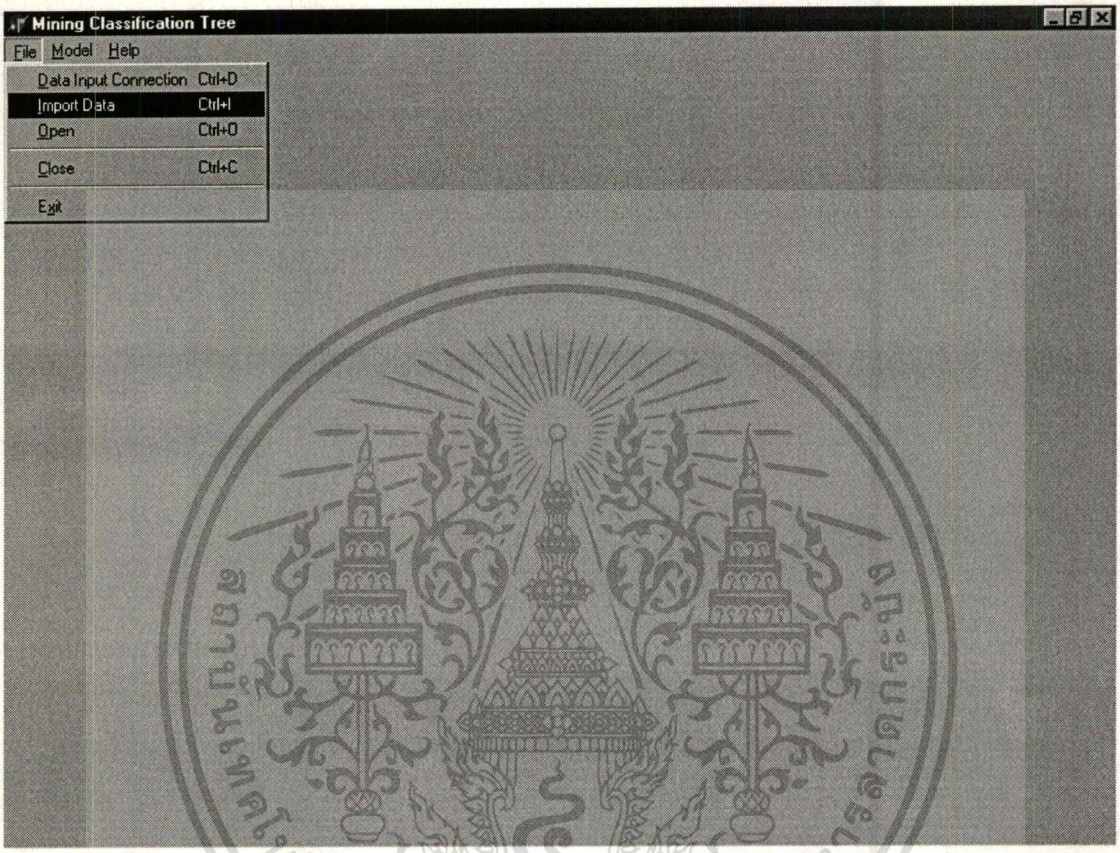
ภาพที่ ก.2 ตัวอย่าง file .nam



ภาพที่ ก.3 ตัวอย่าง file .dat

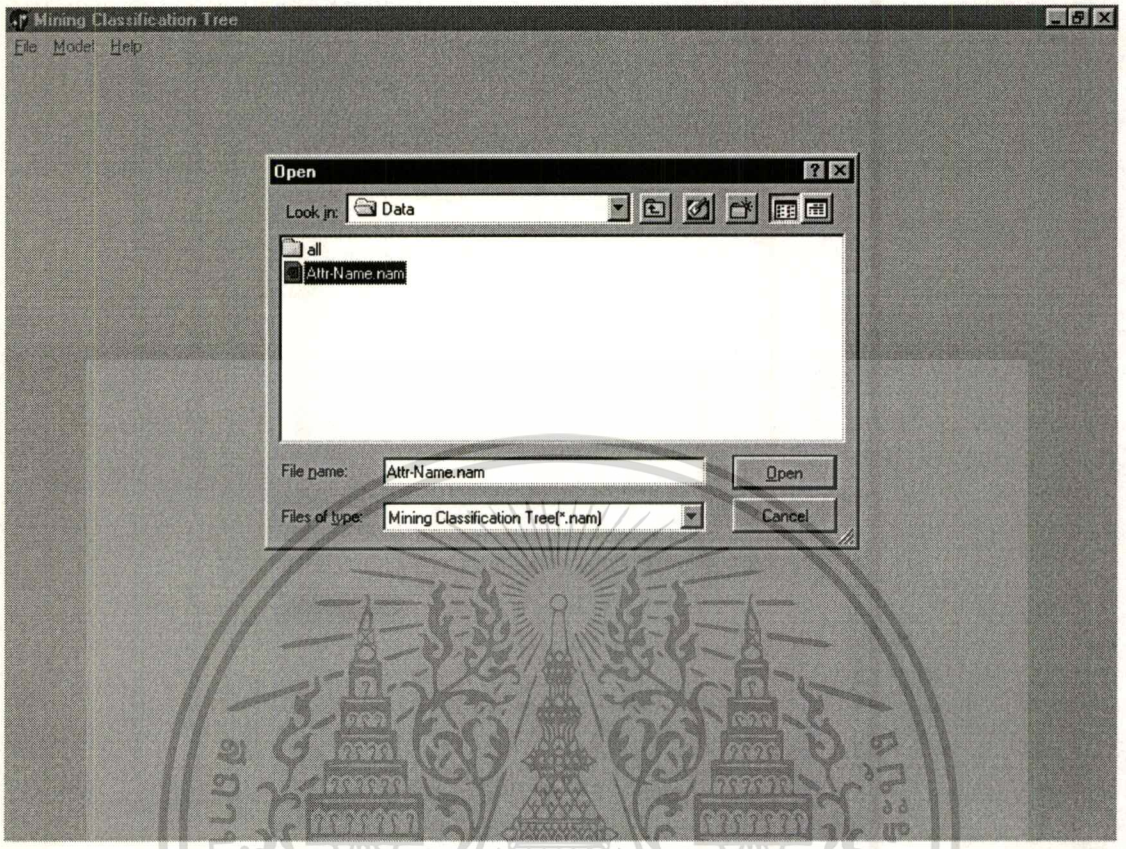
เนื่องจากขั้นตอนการทำงานในส่วนของข้อมูลเข้าที่เป็นฐานข้อมูลได้อธิบายในเอกสารบทที่ 4 แล้ว ในที่นี้จึงขออธิบายถึงขั้นตอนการทำงานในส่วนของข้อมูลเข้าที่อยู่ในรูปแบบเท็กซ์ไฟล์ (Text File) ดังนี้

การติดต่อกับข้อมูลที่เป็นเท็กซ์ไฟล์ ทำได้โดยคลิกที่เมนู Import Data จากเมนูหลัก ดังภาพที่ ก.4



ภาพที่ ก.4 เมนูการติดต่อกับข้อมูลที่เป็นเท็กซ์ไฟล์

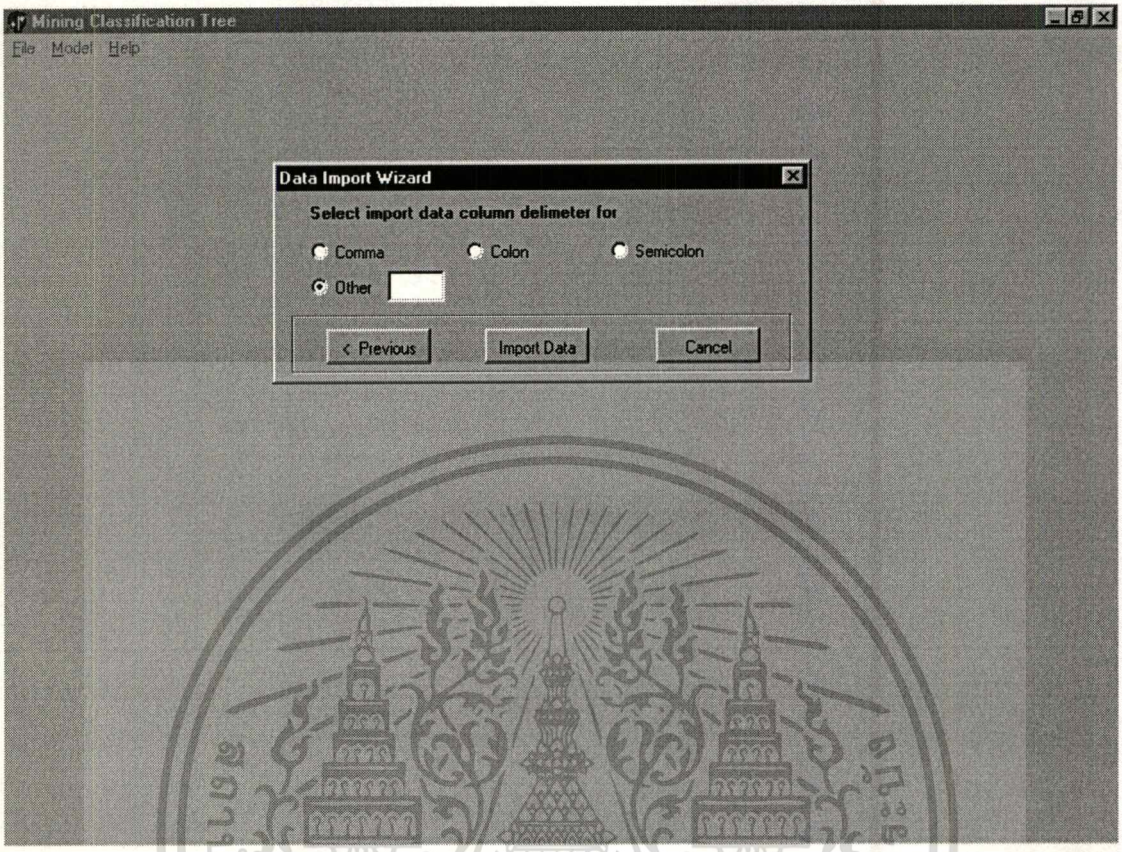
จากนั้น จะปรากฏหน้าจอดังภาพที่ ก.5 เพื่อให้เลือกชื่อไฟล์ที่ต้องการติดต่อ โดยระบบจะกำหนดให้เลือกได้เฉพาะไฟล์ที่นามสกุล .nam



ภาพที่ ก.5 หน้าจอการเลือกไฟล์ .nam

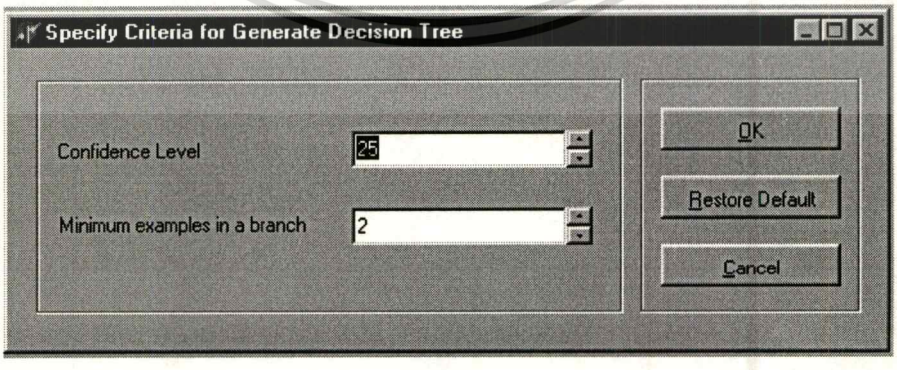
ภาพที่ ก.6 หน้าจอการเลือกไฟล์ .dat

จากภาพที่ ก.6 ผู้ใช้สามารถเลือกปุ่ม Select เพื่อทำการเลือกไฟล์ที่ต้องการ เมื่อผู้ใช้เลือกไฟล์เสร็จแล้ว ให้เลือกปุ่ม Next เพื่อเลือกเงื่อนไขในการดึงข้อมูลเท็กซ์ไฟล์เข้าว่าข้อมูลที่พิมพ์เข้ามาแต่ละค่า คั่นด้วยเครื่องหมายอะไร ดังภาพที่ ก.7



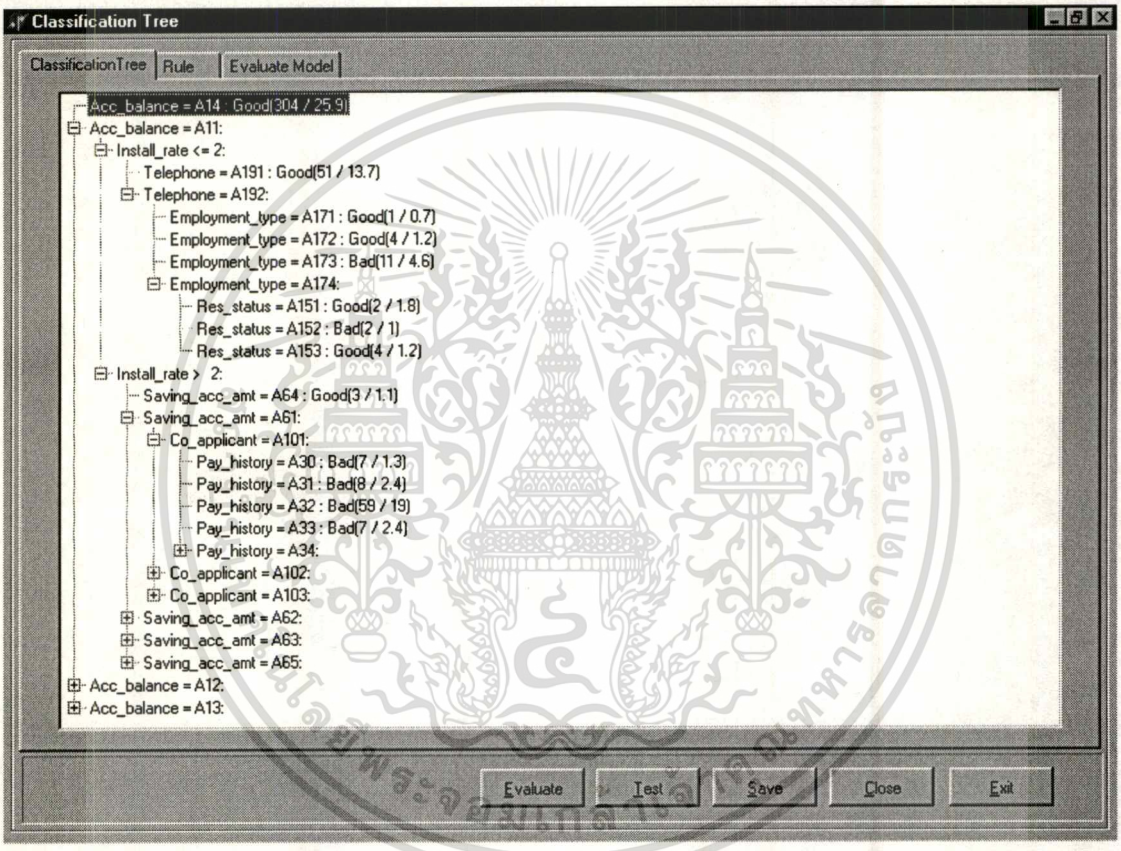
ภาพที่ ก.7 หน้าจอการเลือกเงื่อนไขการดึงข้อมูลเข้า

จากนั้นจะปรากฏหน้าจอดังภาพที่ ก.8 เพื่อให้ระบุค่า Confidence Level และ Maximum examples และให้เลือกปุ่ม OK เพื่อให้ระบบสร้างดิสิชั่นทรีและกฎ

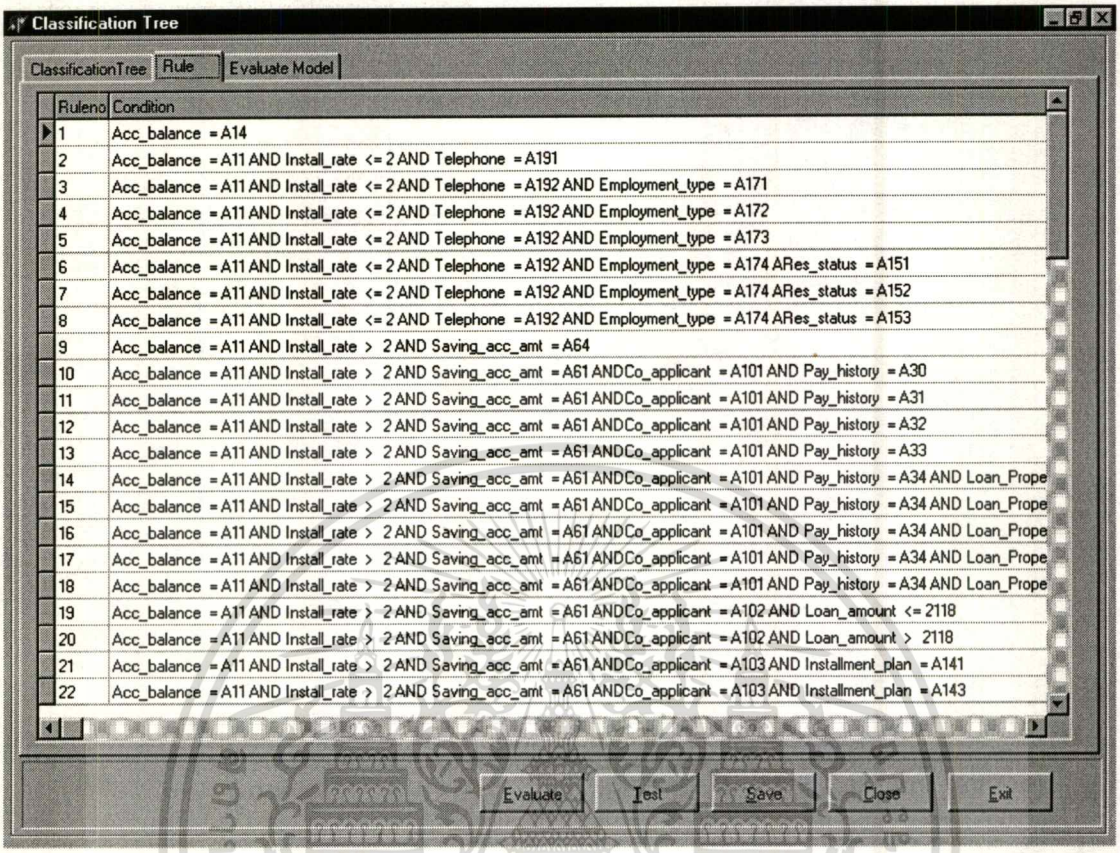


ภาพที่ ก.8 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

เมื่อโปรแกรมทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงในภาพที่ ก.9 และ ก.10 ตามลำดับ โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใดและข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภทที่ (Class) ที่ข้อมูลส่วนใหญ่ในโหนดนั้นตกอยู่ โดยสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้โดยการเลือกปุ่ม Save

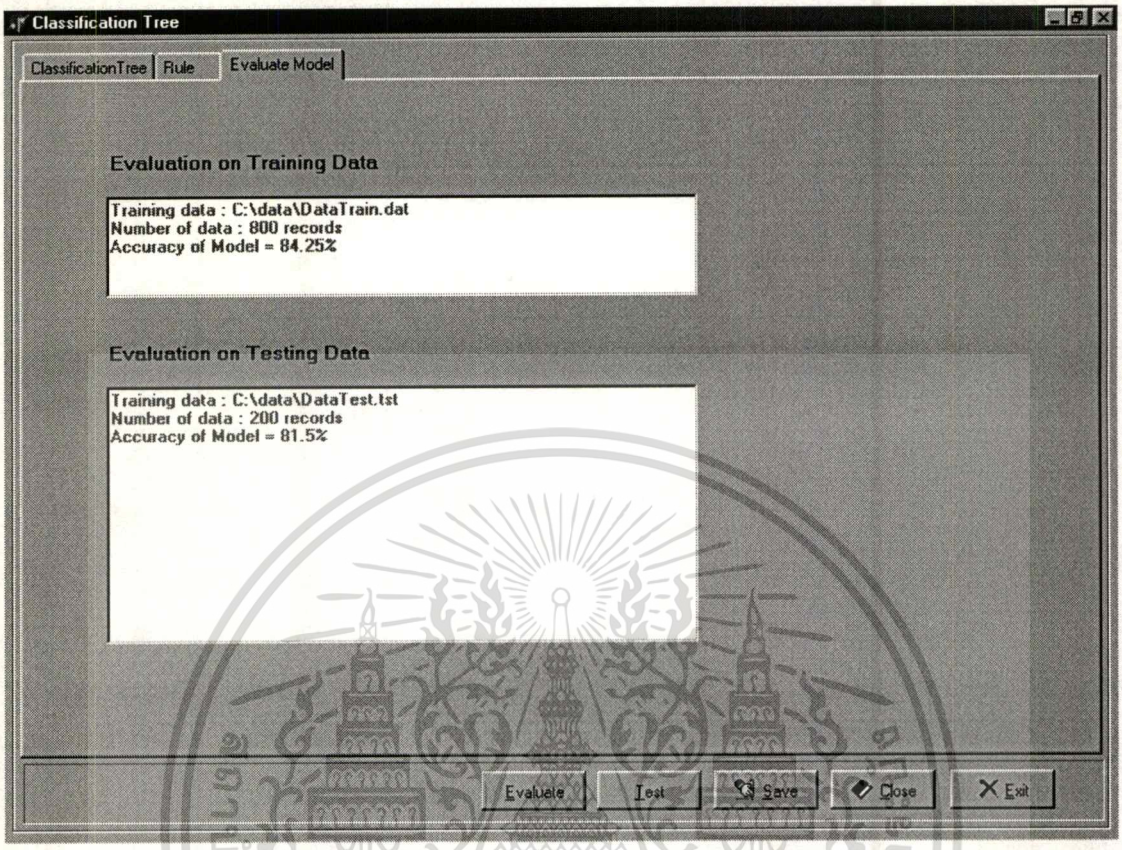


ภาพที่ ก.9 หน้าจอแสดงผลลัพธ์ในรูปดิสิชั่นทรี



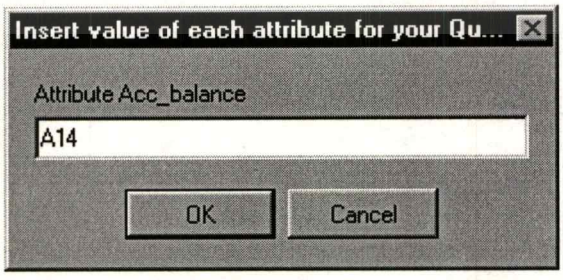
ภาพที่ ก.10 หน้าจอแสดงผลลัพธ์ในรูปของกฎ

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูลที่ใช้ฝึกสอนแล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากเพียงใด โดยการนำข้อมูลอีกชุดหนึ่งมาทำการทดสอบกับแบบจำลองพยากรณ์ที่ได้ โดยเลือกปุ่ม Evaluate เพื่อให้ระบบแสดงความถูกต้องของแบบจำลองโดยใช้ข้อมูลชุดฝึกสอน และเลือกปุ่ม Test เพื่อทำการทดสอบแบบจำลองโดยใช้ข้อมูลทดสอบ โดยระบบจะให้ระบุชื่อไฟล์และชนิดของตัวอักษรที่ใช้คั่นระหว่างข้อมูลแต่ละแอททริบิว โดยหลังจากสั่งให้ระบบทำการตรวจสอบความถูกต้องของแบบจำลองพยากรณ์ ระบบจะแสดงผลการทดสอบ ดังภาพที่ ก.11



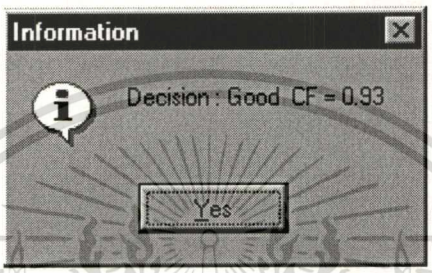
ภาพที่ ก.11 หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

เมื่อทำการสร้างแบบจำลองและทดสอบความถูกต้องจนได้ผลอยู่ในระดับที่พึงพอใจแล้ว ผู้ใช้สามารถสอบถามเกี่ยวกับข้อมูลของผู้ใช้ว่าจัดอยู่ในกลุ่มใดได้ โดยเลือกเมนูย่อย Predict Class ในหน้าจอหลักของระบบ จากนั้นจะปรากฏหน้าต่างให้ใส่ข้อมูลในแต่ละแอททริบิวดังภาพที่ ก.12



ภาพที่ ก.12 หน้าต่างสำหรับใส่ข้อมูลเพื่อสอบถามกลุ่มของข้อมูล

โดยผู้ใช้ไม่ต้องใส่ข้อมูลทุกแอททริบิว แต่จะไล่ไปตามทรีจากรากไปยังปลาย และในกรณีที่ผู้ใช้ไม่ทราบค่าในแอททริบิวใดก็สามารถเว้นว่างไว้หรือใส่เป็น "-" ก็ได้ เมื่อถึงส่วนปลายของทรีซึ่งระบุกลุ่มที่คาดว่าข้อมูลจะจัดอยู่ ระบบก็จะแสดงผลการทำนายว่าข้อมูลควรจะจัดอยู่ในกลุ่มใดและมี Certainty Factor(CF) เป็นเท่าใด โดยค่า CF คือค่าความน่าจะเป็นที่ข้อมูลจะตกอยู่ในกลุ่มนั้นๆ ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1 ดังภาพที่ ก.13



ภาพที่ ก.13 หน้าต่างแสดงผลการทำนายกลุ่มของข้อมูล

ประวัติผู้เขียน

นางสาวนฤมล สมบูรณ์เงิน เกิดวันที่ 12 สิงหาคม พ.ศ. 2516 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต จากภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ในปีการศึกษา 2539 และศึกษาต่อในระดับปริญญามหาบัณฑิต สาขาวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในปีการศึกษา 2543 ปัจจุบันทำงานในตำแหน่ง Technical/Business Analyst ให้กับบริษัท ไอทีวัน จำกัด