การวิเคราะห์ความเสี่ยงสินเชื่อธนาคาร

Banking Credit Risk Analysis

โดย

นางวาสนา เทียมประดิษฐ์

รหัส 44067262

อาจารย์ที่ปรึกษา

ผศ.ดร. อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการศึกษากรณีพิเศษ
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2545
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การวิเคราะห์ความเสี่ยงสินเชื่อธนาคาร
นักศึกษา	นางวาสนา เทียมประดิษฐ์
อาจารย์ที่ปรึกษา	ผศ.ดร. อาริต ธรรมโน
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	การจัดการเทคโนโลยีสารสนเทศ
ปีการศึกษา	2545

บทคัดย่อ

จากภาวะทางเศรษฐกิจในปัจจุบันที่มีอัตราการแข่งขันค่อนข้างสูง ทำให้แนวทางในการดำเนินธุรกิจของธนาคารจำเป็นต้องมีการวางนโยบายและกลยุทธ์ให้รัดกุม และหาทางป้องกันความเสี่ยงที่อาจเกิดขึ้น ในที่นี้จะเน้นความเสี่ยงในด้านสินเชื่อ โดยทำการศึกษาข้อมูลลูกค้าและแบ่งกลุ่มลูกค้าในด้านสินเชื่อ เพื่อทำการวิเคราะห์พฤติกรรมของลูกค้าด้วยกระบวนการ Data Mining ในการทำนายความเสี่ยงที่อาจจะเกิดขึ้นในอนาคต ซึ่งจะเป็นแนวทางในการวางแผนนโยบาย และกลยุทธ์รวมทั้งแผนการตลาด ทำให้ฝ่ายสินเชื่อของธนาคารสามารถตัดสินใจปล่อยเงินกู้ให้แก่ลูกค้าได้อย่างเหมาะสม เพื่อป้องกัน NPL ที่อาจจะเกิดขึ้นได้

Title	Banking Credit Risk Analysis
Student	Mrs. Vasana Thaimpradist
Advisor	Asst.Prof.Dr. Arit Thammano
Level of Study	Master of Science in Information Technology
Major	Information Technology Management
Academic Year	2002

ABSTRACT

Under recent circumstances, economical competition is really high. Banks have to carefully plan good policies and find good procedures to prevent the risk, especially credit risk, by studying information of customers and classifying customers for credit purpose in order to analyse customer behaviors with data mining process. Data mining is able to forebode the uncertain risk used for layout policies and tactics including marketing schemes. With these advantages, credit department of the bank can track and recall debts properly so as to inhibit NPL

กิตติกรรมประกาศ

โครงการกรณีพิเศษฉบับนี้เป็นการศึกษา และทำการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารด้วยกระบวนการคาต้าไมนิ่ง ผู้เขียนขอขอบพระคุณ ผศ.ดร. อาริต ธรรมโน อาจารย์ที่ปรึกษา ซึ่งได้กรุณาให้คำแนะนำและคำปรึกษา และขอขอบคุณผู้ที่มีส่วนเกี่ยวข้องทุกท่านที่ได้กรุณาให้คำปรึกษา และคำแนะนำวิธีการต่างๆ ทำให้โครงการกรณีศึกษาฉบับนี้สำเร็จลุล่วง หากมีข้อบกพร่องส่วนใดผู้เขียนขอรับไว้เพียงผู้เดียว

วาสนา เทียมประดิษฐ์

20 กุมภาพันธ์ 2546

สารบัญ

หน้า

สารบัญ

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญภาพ

หน้า

สารบัญภาพ

หน้า

ภาพที่

บทที่ 1
บทนำ

ในปัจจุบันการให้บริการสินเชื่อของสถาบันการเงินมีความเสี่ยงหลาย ๆ ด้าน เนื่องจากความไม่แน่นอนของสภาวะทางเศรษฐกิจและปัจจัยอื่น ๆ ซึ่งก่อให้การดำเนินธุรกิจการเงินการธนาคารมีอัตราเสี่ยงในการเกิดหนี้สูญเป็นจำนวนมาก ทำให้เกิดผลเสียหายต่อธนาคารหลายอย่าง ดังนั้นโครงการศึกษากรณีพิเศษฉบับนี้จึงขอศึกษาและวิเคราะห์ Credit Risk โดยอาศัยหลักการหรือกระบวนการดาต้าไมนิ่งมาใช้ในการจัดประเภท และวิเคราะห์ข้อมูลช่วยในการติดตามหนี้สินเชื่อกรณีที่ผลการวิเคราะห์มีความเป็นไปได้สูง เพื่อช่วยลดความเสี่ยงในการให้บริการสินเชื่อของธนาคาร

1.1 วัตถุประสงค์

เพื่อให้องค์กรสามารถนำสารสนเทศที่มีอยู่ไปใช้ประกอบการตัดสินใจในการติดตามหนี้สินเชื่อที่มีแนวโน้มเป็นหนี้เสีย (NPL) เพื่อลดความเสี่ยงที่อาจเกิดขึ้นได้ โดยอาศัยกระบวนการดาต้าไมนิ่งช่วยในการวิเคราะห์ลักษณะพฤติกรรมการทำธุรกรรมทางด้านสินเชื่อของลูกค้าธนาคาร

1.2 ขอบเขตการศึกษา

โครงการศึกษากรณีพิเศษนี้เป็นการศึกษาและวิเคราะห์ความเสี่ยงของลูกค้าสินเชื่อธนาคารโดยอาศัยกระบวนการดาต้าไมนิ่ง ที่เสนอผลการวิเคราะห์เป็นรูปแบบ Decision Tree และใช้หลักการอัลกอริทึม ID3 ซึ่งเป็นอัลกอริทึมประเภทหนึ่งใน Classification

1.3 ขั้นตอนและวิธีการดำเนินงาน

สำหรับการวิเคราะห์ความเสี่ยงสินเชื่อของลูกค้าธนาคารจะนำกระบวนการดาต้าไมนิ่งเข้ามาเป็นส่วนหนึ่งในการวิเคราะห์ความเสี่ยงของการชำระสินเชื่อที่ให้บริการแก่ลูกค้า เพื่อพิจารณาความเป็นไปได้ในการชำระหนี้ทำให้การติดตามหนี้ดำเนินไปอย่างมีประสิทธิภาพ การทำให้บรรลุวัตถุประสงค์และขอบเขตการศึกษาโครงการพิเศษที่ได้กล่าวมาในข้างต้น สามารถกำหนดขั้นตอนและวิธีการดำเนินงานได้ ดังนี้

1) ศึกษาทฤษฎีและหลักการของกระบวนการดาต้าไมนิ่ง เพื่อนำมาประยุกต์ใช้

2) กำหนดวัตถุประสงค์และขอบเขตของงานที่จะทำการวิเคราะห์
3) รวบรวมข้อมูลที่คาดว่าจะเกี่ยวข้อง เพื่อทำการวิเคราะห์ และทำการ Cleaning ข้อมูลให้พร้อมที่จะนำมาศึกษา
4) สร้างแบบจำลองและทำการทดสอบรวมทั้งประเมินความผิดพลาดที่เกิดจากการวิเคราะห์ เพื่อนำไปปรับปรุงแบบจำลองให้มีประสิทธิภาพดียิ่งขึ้น

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากโครงการพิเศษฉบับนี้เมื่อได้ทำการศึกษาแล้วพอจะสรุปสังเขปได้ดังนี้

1) เข้าใจหลักการและขั้นตอนรวบถึงอัลกอริทึมพื้นฐานของกระบวนการดาต้าไมนิ่ง
2) สามารถนำทฤษฎีและหลักการของกระบวนการดาต้าไม่นิ่งที่ได้ศึกษามา ไปประยุกต์ใช้ในงานธุรกิจที่ดำเนินอยู่จริง
3) ทำให้สามารถใช้ข้อมูลความสัมพันธ์ของข้อมูลในอดีตมาก่อให้เกิดประโยชน์ในการติดตามหนี้ เพื่อลดอัตราเสี่ยงสินเชื่อธนาคาร โดยทำให้สามารถติดตามหนี้ได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

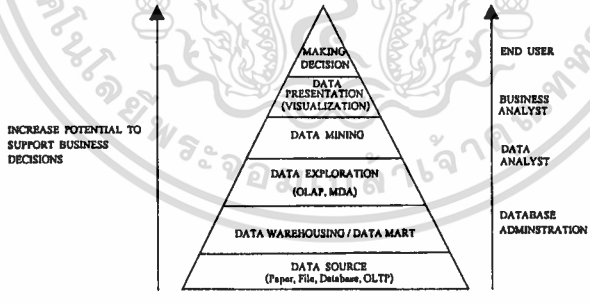
บทที่ 2
หลักการทั่วไปของดาต้าไมนิ่ง

2.1 คำจำกัดความของดาต้าไมนิ่ง

ดาต้าไมนิ่ง เป็นกระบวนการวิเคราะห์ข้อมูลจากฐานข้อมูลซึ่งเป็นข้อมูลในอดีตหรือข้อมูลในปัจจุบัน โดยอาศัยความก้าวหน้าทางการวิเคราะห์เชิงสถิติ และเทคนิคแบบจำลองในการหารูปแบบและความสัมพันธ์ของข้อมูลที่มีอยู่ ซึ่งมีความรู้บางอย่างที่ถูกซ่อนในข้อมูลที่มีอยู่ การใช้วิธีการวิเคราะห์ธรรมดาอาจจะไม่สามารถทราบได้ ดังนั้นจึงใช้กระบวนการดาต้าไมนิ่งเป็นพื้นฐานในการประกอบการตัดสินใจในเชิงธุรกิจ เพื่อทำให้เข้าใจแนวโน้มและรูปแบบของตลาด

ดาต้าไมนิ่งไม่ได้เข้ามาทดแทนเทคนิคหรือวิธีการวิเคราะห์ทางสถิติ แต่เป็นส่วนขยายวิธีการทางสถิติ หรือกล่าวได้อีกอย่างว่า เป็นการประยุกต์การวิเคราะห์ทางสถิติ เพื่อให้สามารถรองรับข้อมูลขนาดใหญ่ได้

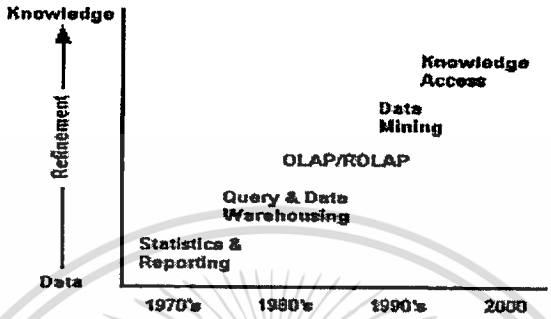
เครื่องมือหรือกระบวนการวิเคราะห์ข้อมูลที่ใช้ในการสนับสนุนการตัดสินใจในเชิงธุรกิจมีหลายประเภท ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 เครื่องมือและวิธีการวิเคราะห์ต่าง ๆ ทางธุรกิจ

จากรูปจะเห็นว่า คุณค่าของข้อมูลที่ใช้สนับสนุนการตัดสินใจจะเพิ่มขึ้นจากล่างไปบนสุดของรูปปิรามิด ระดับของผู้ตัดสินใจจะแตกต่างกัน เช่น Database Administrator จะตัดสินใจบนระดับของ Data Warehousing และแหล่งข้อมูลเท่านั้น ส่วน Data Analyst และ Business Analyst จะตัดสินใจช่วงบนของปิรามิด

เครื่องมือหรือวิธีการวิเคราะห์ข้อมูล ถูกพัฒนาให้ดีขึ้นเรื่อย ๆ จากอดีตจนถึงปัจจุบันเพื่อให้ได้ Information ที่ถูกต้องและใช้ประโยชน์จาก Information เหล่านั้นให้มากขึ้น เพิ่มความรู้ที่ได้จากข้อมูลดิบที่มีอยู่ ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 กราฟแสดงการพัฒนาเครื่องมือและวิธีการวิเคราะห์กับแนวโน้มความรู้ที่ได้รับจากผลการวิเคราะห์ (จากปี ค.ศ. 1970 ถึงปี ค.ศ. 2000)

เพราะฉะนั้นจึงอาจจะถือได้ว่ากระบวนการดาต้าไมนิ่ง เป็นวิธีที่จะนำเอาข้อมูลมาใช้เพื่อการวิเคราะห์ให้ก่อเกิดประโยชน์สูงสุด โดยเฉพาะอย่างยิ่งการตัดสินใจของฝ่ายบริหารขององค์กร ดาต้าไมนิ่งช่วยธุรกิจในการวิเคราะห์หาแนวทางและความสัมพันธ์ของข้อมูล ซึ่งจะไม่บอกว่ามูลค่าของแนวทางแก่องค์กร โดยแนวทางที่ถูกเปิดเผยด้วยกระบวนการดาต้าไมนิ่งจำเป็นต้องถูกตรวจสอบในโลกของความจริงอีกครั้ง

ดาต้าไมนิ่งถูกนำมาใช้จัดการด้านการตลาดเป็นส่วนใหญ่ บางครั้งเราเรียกแทนด้วยคำว่า "Database Marketing" จุดประสงค์เพื่อต้องการทราบลูกค้ากลุ่มเป้าหมายที่แท้จริง และสามารถสร้างโปรแกรมส่งเสริมการขาย และโปรแกรมทางการตลาดให้มีประสิทธิภาพ โดยวิธีการทำงานของกระบวนการดาต้าไมนิ่งจะมองผ่านเข้าไปในข้อมูลอย่างละเอียด เพื่อมองหาแบบจำลอง (Model) พฤติกรรมของผู้บริโภคในกลุ่มที่มีลักษณะคล้ายกัน เช่น ระดับรายได้, พฤติกรรมการใช้เงิน เป็นต้น ทำให้องค์กรธุรกิจประหยัดค่าใช้จ่ายในการโฆษณาหรือประชาสัมพันธ์เพื่อหาลูกค้ากลุ่มเป้าหมาย เพราะองค์กรธุรกิจสามารถมุ่งไปสู่ลูกค้ากลุ่มเป้าหมายได้โดยตรง

จากที่กล่าวมาจะเห็นว่ากระบวนการวิเคราะห์ ดาต้าไมนิ่งจัดเป็นเทคโนโลยีที่เข้ามามีส่วนเกี่ยวข้องสัมพันธ์กับระบบการทำงานในองค์กร เป็นเครื่องมือเพื่อเพิ่มประสิทธิภาพ แต่ไม่ใช่องค์ประกอบหลักในการทำงาน ทำให้เห็นความเกี่ยวข้องของ ดาต้าไมนิ่งกับระบบ Operation System เป็นดังรูปที่ 2.3

2.2 องค์ประกอบที่ทำให้ประสบความสำเร็จในการวิเคราะห์ด้วยดาต้าไมนิ่ง

องค์ประกอบสำคัญที่จะทำให้ประสบความสำเร็จในการวิเคราะห์ด้วยดาต้าไมนิ่ง ได้แก่

1. การใช้ข้อมูลที่มีความถูกต้องและมีความเหมาะสมกับปริมาณของข้อมูล เพราะมีส่วนทำให้การเรียนรู้แบบจำลอง (Model) มีประสิทธิภาพและความถูกต้องแม่นยำมากขึ้น นอกจากนั้นยังทำให้ประหยัดเวลาในการทำดาต้าไมนิ่ง

2. เลือกใช้เทคนิคหรือ Algorithm ที่เหมาะสมกับชนิดของข้อมูลที่นำมาวิเคราะห์ และเป้าหมายหรือวัตถุประสงค์ของงานที่กำลังจะทำการวิเคราะห์

3. เลือกใช้ Tool ในการวิเคราะห์ดาต้าไมนิ่งที่มีประสิทธิภาพและเหมาะสมกับงานที่จะทำการวิเคราะห์

2.3 ขั้นตอนในการทำดาต้าไมนิ่ง

กระบวนการของการทำดาต้าไมนิ่งมีการทำหลายขั้นตอน และสามารถย้อนกลับไปกลับมาได้ตลอดเวลา เพื่อให้ได้ผลการวิเคราะห์ที่มีประสิทธิภาพ และความน่าเชื่อถือมากที่สุดเท่าที่จะทำการวิเคราะห์ได้ โดยขึ้นอยู่กับปัจจัยแวดล้อมด้วย เช่น จำนวนข้อมูลที่นำมาวิเคราะห์, ลักษณะของข้อมูลที่นำมาวิเคราะห์ เป็นต้น สำหรับโครงการพิเศษที่ได้ศึกษามาฉบับนี้จะแบ่งขั้นตอนการดาต้าไมนิ่ง ออกเป็น 6 ขั้นตอนหลัก ได้แก่

ขั้นตอนที่ 1 กำหนดวัตถุประสงค์และปัญหาที่ต้องการวิเคราะห์ (Define the objective and problem)

ในขั้นตอนแรกนี้เราต้องรู้และเข้าใจปัญหาอย่างถ่องแท้และชัดเจน กำหนดขอบเขตและวัตถุประสงค์ กำหนดปัญหาและสิ่งที่ต้องการจากผลการวิเคราะห์ ซึ่งในส่วนการวิเคราะห์นี้จะประกอบด้วยการวิเคราะห์แนวทางธุรกิจตลอดจนถึงข้อมูลเบื้องต้นว่าเกี่ยวข้องกับข้อมูลอะไรบ้าง เพื่อเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่งให้บรรลุวัตถุประสงค์ที่กำหนดไว้

ขั้นตอนที่ 2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนที่ 2 จัดเป็นขั้นตอนที่สำคัญและใช้เวลาในการทำมากที่สุด โดยปกติจะใช้เวลาประมาณ 60 % ของเวลาทั้งหมดในการทำดาต้าไมนิ่ง ซึ่งขั้นตอนนี้สามารถจำแนกออกเป็น 3 ขั้นตอนย่อย ๆ ดังนี้

- **การรวบรวมและเลือกข้อมูล (Data Collection and Selection)** เมื่อกำหนดปัญหาและวัตถุประสงค์แล้วขั้นตอนต่อไปคือ การรวบรวมและเลือกข้อมูลที่อยู่ในช่วงหรือลักษณะที่ต้องการ

เพื่อให้สามารถแก้ปัญหาหรือทำให้บรรลุวัตถุประสงค์ที่กำหนดไว้ โดยข้อมูลที่จะใช้วิเคราะห์จะได้มาจากฐานข้อมูลซึ่งมีอยู่แล้วในระดับปฏิบัติการต่าง ๆ หรืออาจได้มาจากคลังข้อมูล โดยปกติอัลกอริทึมของดาต้าไมนิ่งไม่สามารถทำงานได้โดยตรงกับฐานข้อมูลหลายๆ รูปแบบพร้อมกันได้ จะต้องมีการรวบรวมข้อมูลเหล่านั้นให้เป็นรูปแบบเดียวกันก่อนที่จะนำไปวิเคราะห์ด้วยดาต้าไมนิ่ง ต่อไป

- **การกลั่นกรองข้อมูล (Data Preprocessing and Clearing)** เมื่อทำการเก็บรวบรวมและเลือกข้อมูลที่จะใช้วิเคราะห์เรียบร้อยแล้ว ขั้นตอนต่อไปคือการกรองข้อมูลเป็นการทำให้ข้อมูลมีความถูกต้องครบถ้วน โดยนำข้อมูลที่ไม่ถูกต้องออกไปด้วยเทคนิคต่างๆ เช่น Clustering , Binning และ Regression เป็นต้น ส่วนข้อมูลบางส่วนที่ขาดหายไป สามารถแก้ไขด้วยการตัดข้อมูลนั้นทิ้งทั้งรายการ ในกรณีที่ข้อมูลนั้นมีปริมาณน้อยมากเมื่อเทียบกับข้อมูลทั้งหมด แต่ถ้ากรณีที่ข้อมูลที่ไม่สมบูรณ์มีปริมาณค่อนข้างมากเมื่อเทียบกับข้อมูลทั้งหมดควรจะใช้วิธีการหาค่าเฉลี่ย (Mean) เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ได้ทำการเลือกไว้นั้นเหมาะสมและมีความถูกต้องครบถ้วน ปัจจัยที่ต้องมีการกลั่นกรองข้อมูลมี 2 ประการ ได้แก่

  1) นักวิเคราะห์จำเป็นต้องมีความคุ้นเคยและรู้ซึ้งถึงข้อมูลที่ใช้วิเคราะห์ ไม่ใช่รู้เพียงชื่อและความหมายของ Attribute แต่ต้องรู้ถึงเนื้อหาหรือความมุ่งหมายที่แท้จริงของข้อมูลด้วย

  2) ในขณะที่ทำการเก็บรวบรวมข้อมูลจากแหล่งข้อมูลหลายแหล่งอาจก่อให้เกิดความผิดพลาดขึ้น จึงจำเป็นต้องกลั่นกรองข้อมูลอีกครั้ง

- **การปรับแต่งหรือแปลงข้อมูล (Data Transformation)** หลังจากขั้นตอนกลั่นกรองข้อมูลเรียบร้อยแล้ว ขั้นตอนย่อยนี้จะทำการแปลงข้อมูลให้อยู่ในรูปแบบของข้อมูลที่มีความสอดคล้องและพร้อมที่จะนำไปทำการวิเคราะห์ โดยไม่เกิดความรู้สึกขัดแย้งซึ่งจะถูกจัดให้ข้อมูลเป็นระเบียบเรียบร้อยและเหมาะสม

ขั้นตอนที่ 3 การเลือกใช้อัลกอริทึม (Algorithm Engineering)

เป็นขั้นตอนในการเลือกใช้อัลกอริทึมที่เหมาะสมกับปัญหาที่ต้องการวิเคราะห์ เพื่อใช้ในการวิเคราะห์ค้นหา Pattern รวมถึงการกำหนดกฎเกณฑ์และความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ผ่านวิธีทางด้านสถิติ การเลือกอัลกอริทึมนั้นอาจสามารถเลือกใช้มากกว่า 1 อัลกอริทึมโดยใช้ในการเปรียบเทียบผลลัพธ์ของการวิเคราะห์ด้วยอัลกอริทึมแต่ละอัลกอริทึมว่าผลของการวิเคราะห์มีความน่าเชื่อถือมากกว่า

ขั้นตอนที่ 4 การทำดาต้าไมนิ่ง (Running the data mining algorithm)

หลังจากเลือกอัลกอริทึมที่เหมาะสมกับลักษณะของปัญหาและลักษณะของข้อมูลแล้ว เราจะนำอัลกอริทึมนั้นมาทำการวิเคราะห์ข้อมูลที่เตรียมไว้ ซึ่งบางครั้งขั้นตอนนี้จะถูกเรียกว่า "data mining" ในขณะที่จะเรียกกระบวนทั้งหมดว่า "knowledge discovery in databases" ผลลัพธ์ที่ได้จากขั้นตอนนี้จะเป็นรูปแบบของความสัมพันธ์ของข้อมูลหรือแบบจำลอง (model) ที่จะนำมาใช้ในการพยากรณ์ (prediction) หรือทำการวิเคราะห์ต่อไป

ขั้นตอนที่ 5 การวิเคราะห์ผลลัพธ์ (Analysis of Result)

เป็นขั้นตอนการทำการวิเคราะห์และตรวจสอบผลลัพธ์ที่ได้ และทดลองนำแบบจำลอง (Model) ที่ได้ไปใช้งานจริง เพื่อจะนำเอาผลลัพธ์ที่ได้มาเปรียบเทียบกับผลตามแบบจำลอง (Model) ว่ามีความแม่นยำหรือไม่ และยอมรับได้หรือไม่หากยอมรับไม่ได้ ก็อาจทำการแก้ไข โดยการเพิ่มจำนวนข้อมูลให้มากขึ้น หรือเปลี่ยนไปใช้อัลกอริทึมอื่น เป็นต้น ซึ่งหากแบบจำลองที่ได้ให้ผลลัพธ์ที่เที่ยงตรงในระดับที่ยอมรับได้ก็จะนำแบบจำลองนี้ไปใช้งานจริงต่อไป

ขั้นตอนที่ 6 การนำไปใช้งานจริง (Impletement)

ในการใช้งานจริงนั้น Model ที่ได้อาจไม่ถูกต้องตลอดไป ต้องมีการตรวจสอบ Model ตลอดเวลา ทั้งนี้เพราะว่าสิ่งแวดล้อมของระบบมักมีการเปลี่ยนแปลงอยู่เสมอ ดังนั้นกระบวนการการทำดาต้าไมนิ่งจึงเป็นกระบวนการที่ต้องมีการตรวจสอบและทำซ้ำอยู่ตลอดเวลา

จากขั้นตอนหลักทั้ง 6 ขั้นตอนจะเห็นได้ว่าการ ซึ่งขั้นตอนการเตรียมข้อมูลจะใช้ทรัพยากรมากที่สุด เพราะฉะนั้นควรมีข้อมูลมากพอสมควร ไม่เช่นนั้นอาจจะไม่สามารถวิเคราะห์ เพื่อดึงความรู้บางอย่างที่แฝงอยู่ในข้อมูลนั้นออกมาใช้ประโยชน์ได้หรือผลลัพธ์ของการวิเคราะห์มีโอกาสผิดพลาดสูง

2.4 เทคนิคของดาต้าไมนิ่ง

ในการทำดาต้าไมนิ่งขั้นตอนที่สำคัญอย่างหนึ่ง คือ การเลือกเทคนิคหรืออัลกอริทึมที่เหมาะสมในการวิเคราะห์ ปัจจัยที่นำมาพิจารณาในการเลือกเทคนิคต่าง ๆ ได้แก่

- เป้าหมายที่แตกต่างกัน ย่อมต้องการเทคนิคการวิเคราะห์ซึ่งแตกต่างกันด้วย
- ชนิดของข้อมูลที่นำมาวิเคราะห์ที่แตกต่างกัน จำเป็นต้องอาศัยเทคนิคการวิเคราะห์ที่แตกต่างกันด้วย

เทคนิคในการวิเคราะห์ด้วยกระบวนการดาต้าไมนิ่งมีหลายเทคนิค ซึ่งขึ้นอยู่กับปัจจัยในการพิจารณาที่กล่าวมาข้างต้น ในโครงการพิเศษฉบับนี้จะขอนำเสนอเทคนิคที่เลือกใช้ในการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารคือ เทคนิค Decision Tree

Decision Tree เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากเหมาะสำหรับข้อมูลที่สามารถแบ่งเป็นประเภทต่างๆ ได้อย่างชัดเจน หรือสามารถคาดการณ์ผลลัพธ์ที่อาจจะเกิดขึ้นได้ในอนาคต นอกจากนี้เทคนิค Decision Tree ยังเป็นการกำหนดกฎเกณฑ์ที่มีความง่ายในการทำความเข้าใจและง่ายต่อการอธิบายให้เข้าใจ นอกจากนั้นยังง่ายในการแปลงอยู่ในรูป SQL ที่เราคุ้นเคย มีประสิทธิภาพสูงในการวิเคราะห์ เพราะฉะนั้นจึงเลือกใช้ในการวิเคราะห์การตัดสินใจในเรื่องความเสี่ยงสินเชื่อธนาคาร ในโครงการพิเศษนี้จะขออธิบายรายละเอียดเทคนิคและอัลกอริทึมที่ได้ศึกษาดังจะกล่าวต่อไปนี้

2.5 ประเภทลักษณะของดาต้าไมนิ่ง

ดาต้าไมนิ่งสามารถแบ่งออกได้เป็นหลายลักษณะ ได้แก่

- **Classification** เป็นการจัดกลุ่มของข้อมูลที่เป็นประเภทเดียวกัน เพื่อวิเคราะห์คุณสมบัติเป็นกลุ่ม ๆ อย่างชัดเจนแน่นอน เช่น “ใช่”, “ไม่ใช่” หรือ “สูง”, “กลาง”, “ต่ำ” เป็นต้น ตัวอย่างเช่น กลุ่มลูกค้าที่ขอทำบัตรเครดิต สามารถแบ่งเป็น 3 กลุ่ม คือ กลุ่มที่มีอัตราเสี่ยงสูง, กลุ่มที่มีอัตราเสี่ยงปานกลาง และกลุ่มที่มีอัตราเสี่ยงต่ำ เทคนิคที่ใช้ในลักษณะนี้อาจจะเป็น Decision Tree หรือ Neual Network

- **Estimation** เป็นการประเมินหรือการคาดคะเน ซึ่งไม่สามารถแบ่งได้อย่างชัดเจนเหมือนแบบ Classification เป็นการกะประมาณในการวัดที่ต่อเนื่อง ตัวอย่างเช่น การประเมินราคาที่ดินในอีก 2 ปีข้างหน้า หรือ การประเมินรายได้รวมของแต่ละครอบครัว

- **Prediction** เป็นการทำนายจะมีลักษณะคล้ายกับแบบ Classification และแบบ Estimation แต่ต่างกันที่ข้อมูลที่นำมาแบ่งกลุ่มจะเกิดขึ้นจากค่าการทำนายในอดีตและถูกสร้างเป็นแบบจำลอง (Model) เพื่อทำนายหรืออธิบายสิ่งที่จะเกิดขึ้นในอนาคต ตัวอย่างเช่น Credit Risk Model สำหรับทำนายความเสี่ยงของลูกค้าที่กู้เงินจะไม่สามารถชำระได้ในอนาคต เกิดเป็นหนี้สูญ

- **Affinity Grouping or Association Rule** เป็นการตัดสินใจและวิเคราะห์จัดกลุ่มสิ่งที่ไปด้วยกันได้ หรือกล่าวอีกนัยหนึ่งว่า เป็นการกำหนดว่าสิ่งใดควรจะเกิดขึ้นเมื่อเกิดเหตุการณ์หนึ่งแล้ว เช่น คนที่มาซื้ออาหารสุนัขในร้านมักจะซื้อกระดูกปลอม จะมีโอกาสเกิดขึ้นกี่เปอร์เซ็นต์

2.6 ประโยชน์ของการนำกระบวนการดาต้าไมนิ่งไปประยุกต์ใช้

กำไรจากการพิจารณาถึงองค์ประกอบในด้านต่างๆของการนำเอากระบวนการดาต้าไมนิ่งมาใช้งานเราจะพบว่ารูปแบบการใช้งานดาต้าไมนิ่งจะมีลักษณะการใช้งานดังต่อไปนี้

- นำเอาคุณค่าที่ถูกซ่อนอยู่ภายในตัวข้อมูลออกมาใช้ให้เกิดประโยชน์สูงสุดทั้งจากข้อมูลภายในและภายนอกองค์กร
- นำเอาข้อมูลต่างๆจากอดีตมาทำความเข้าใจเพื่อนำมาปรับปรุงใช้งานในอนาคตได้อย่างเหมาะสม
- สามารถคาดการณ์เหตุการณ์ที่อาจเกิดในอนาคตได้อย่างใกล้เคียงความเป็นจริงที่สุด
- สามารถแบ่งแยกลูกค้ากลุ่มต่างๆได้อย่างถูกต้องและเหมาะสม
- สามารถนำเสนอผลิตภัณฑ์ขององค์กรได้อย่างเหมาะสมกับกลุ่มลูกค้าในเวลาที่ถูกต้อง
- สามารถตอบสนองความต้องการของกลุ่มลูกค้าที่มีความสำคัญสูงได้อย่างรวดเร็วเพื่อเป็นแรงจูงใจให้ลูกค้าใช้บริการขององค์กรต่อไป
- สามารถทำการวิเคราะห์ความเสี่ยงของลูกค้ากลุ่มต่างๆได้อย่างถูกต้องเพื่อลดความเสียหายที่อาจจะเกิดขึ้น และหาแนวทางป้องกันแก้ไขต่อไป
- สามารถคาดการณ์แนวโน้มความต้องการของตลาดได้อย่างมีประสิทธิภาพ
- สามารถสร้างผลกำไรได้มากยิ่งขึ้นจากการนำเอาดาต้าไมนิ่ง มาใช้ในการดำเนินธุรกิจ และลดความเสียหายในด้านต่างๆได้อย่างมีประสิทธิภาพ

2.7 ข้อจำกัดของกระบวนการดาต้าไมนิ่ง

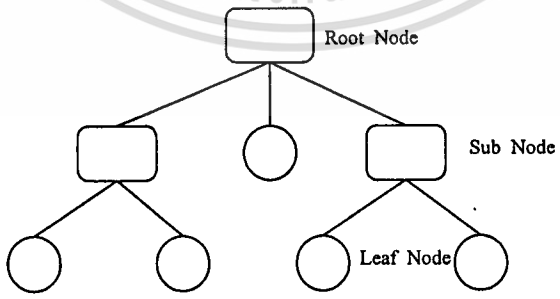
ข้อจำกัดบางประการของดาต้าไมนิ่ง คือ ข้อมูลที่เลือกมาใช้เป็นข้อมูลที่ไม่มีความสัมพันธ์กัน หรือความสัมพันธ์อยู่ในรูปแบบที่ไม่ถูกต้อง จะทำให้ผลลัพธ์ของการวิเคราะห์ผิดพลาด นอกจากนั้นยังมีเรื่องของปริมาณข้อมูลที่ไม่เพียงพอในการทำดาต้าไมนิ่งซึ่งจะทำให้ไม่สามารถมองเห็นผลลัพธ์ที่แท้จริงได้จัดเชน หรือขาดความน่าเชื่อถือของผลการวิเคราะห์ ซึ่งอาจก่อให้เกิดผลเสียหายต่อองค์กรได้

บทที่ 3
ประเภทและเทคนิคที่ใช้ในกระบวนการดาต้าไมนิ่ง

3.1 เทคนิค Decision Tree

Decision Tree เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากเหมาะสำหรับข้อมูลที่สามารถแบ่งเป็นประเภทต่าง ๆ ได้อย่างชัดเจน หรือสามารถคาดการณ์ผลลัพธ์ที่อาจจะเกิดขึ้นได้ในอนาคต นอกจากนี้เทคนิค Decision Tree ยังเป็นการกำหนดกฎเกณฑ์ที่มีความง่ายในการทำความเข้าใจ และง่ายต่อการอธิบายให้เข้าใจ นอกจากนั้นยังง่ายในการแปลงอยู่ในรูป SQL ที่เราคุ้นเคยมีประสิทธิภาพสูงในการวิเคราะห์ เพราะฉะนั้นจึงเลือกใช้ในการวิเคราะห์การตัดสินใจในเรื่องความเสี่ยงสินเชื่อธนาคาร ในโครงการพิเศษฉบับนี้จะขออธิบายรายละเอียดเทคนิคและอัลกอริทึมที่ได้ศึกษามา ดังต่อไปนี้

Decision Tree มีลักษณะคล้ายต้นไม้ที่แตกกิ่งก้านสาขาออกไป จะประกอบด้วย Node แรกสุด เรียกว่า Root Node จาก Root Node จะค่อย ๆ แตกย่อยลงไปเป็น Sub Node ซึ่งแต่ละ Sub Node ก็ยังสามารถแตกเป็น Sub Node ย่อยลงไปอีก ซึ่งการแตก Sub Node อาจแตกได้มากกว่า 2 Nodes ก็ได้ Sub Node ที่อยู่ท้ายสุดเรียกว่า Leaf Node ลักษณะโครงสร้างของ Decision Tree จะแสดงดังรูป 3.1



รูปที่ 3.1 รูปแบบโครงสร้างพื้นฐานของ Decision Tree

โดยปกติเทคนิค Decision Tree จะอยู่ในรูปแบบของกฎ "IF-THEN" เช่น

IF age "<=30" AND student = "yes"

THEN buy_computer = "yes"

IF age "<=30" AND student = "no"

THEN buy_computer = "no"

ในการวิเคราะห์ด้วยเทคนิคแบบ Decision Tree จะมีการแบ่งข้อมูลออกเป็น 2 กลุ่มหลัก ๆ คือ

1. กลุ่มข้อมูลเพื่อการเรียนรู้ (Training Data Set) หมายถึง การนำเอากลุ่มข้อมูลซึ่งเป็นข้อมูลในอดีตมาทำการวิเคราะห์ โดยใช้อัลกอริทึมที่เลือกทำการหารูปแบบ Model ออกมา กลุ่มข้อมูลนี้จำเป็นต้องใช้ข้อมูลขนาดค่อนข้างใหญ่ เพื่อทำให้ Model มีความถูกต้องมากที่สุดเท่าที่จะทำได้

2. กลุ่มข้อมูลเพื่อการทดสอบ (Testing Data Set) หมายถึง กลุ่มของข้อมูลที่นำมาใช้เพื่อทดสอบการทำงานของ Model ว่ามีความถูกต้องมากน้อยเพียงใด ซึ่งจะต้องไม่ใช่กลุ่มข้อมูลชุดเดียวกับที่สร้าง Model เพื่อพิจารณาว่า Model ที่ถูกสร้างขึ้น สามารถนำไปใช้กับกลุ่มข้อมูลจริงหรือกลุ่มข้อมูลใหม่ ๆ ได้หรือไม่

อัลกอริทึมที่ใช้ในการสร้าง Decision Tree มีอยู่หลายอัลกอริทึม เช่น

- CHAID (Chi-squared Automatic Interaction Detection)
- CHAT (Classification And Regression Tree)
- ID3 (Iterative Dichotomiser)
- C4.5 พัฒนาเพิ่มเติมจาก ID3

สำหรับในโครงการพิเศษฉบับนี้จะเป็นการศึกษารายละเอียดของอัลกอริทึม ID3 เนื่องจาก ID3 เป็นอัลกอริทึมพื้นฐานแบบ Greedy Algorithm

หลักการของอัลกอริทึม ID3 คือต้องคำนวณหาค่า Gain ของทุก Attribute แล้วทำการเลือก Attribute ที่มีค่า Gain สูงสุด สูตรการหา Gain เริ่มต้นจากสูตรดังนี้

$$P_i = \frac{S_i}{|S|} \qquad \text{..... (1)}$$

โดย P_i หมายถึง Probability ของ Class i

S_i หมายถึง จำนวนข้อมูลใน Class i

$|S|$ หมายถึง จำนวนข้อมูลทั้งหมดใน Set ข้อมูล

จากสูตรที่ 1 เป็นการหาค่าความน่าจะเป็นของข้อมูล Class ที่ i โดยนำค่าความน่าจะเป็น (P_i) มาใช้ในสูตร เพื่อวัดค่าของ Information ต่อไปด้วยสูตรดังนี้

$$I(S_1,S_2,\ldots,S_m) = -\sum_{i=1}^{m} P_i \log_2 (P_i) \qquad \ldots\ldots\ldots\ldots (2)$$

โดย $I(S_1,S_2,\ldots,S_m)$ หมายถึง entropy ของ Set S เป็น การวัดค่าของ Information ของ Set S ซึ่ง $S = S_1,S_2,\ldots,S_m$

m หมายถึง จำนวน Class ที่แจกแจงซึ่งแบ่งเป็น m Class

เมื่อนำสูตรที่ 2 ไปใช้ในการคำนวณหาค่าที่วัดค่าของ Information เพื่อแบ่ง Traning Data Set ที่สนใจโดยใช้ค่าที่เป็นไปได้ของ Attribute A ดังสูตรที่ 3

$$E(A) = \sum_{j=1}^{v} \frac{Sij+\ldots+Smj}{S} I(S_{1j},\ldots,S_{mj}) \qquad \ldots\ldots\ldots\ldots (3)$$

จากสูตรที่ 2 จะได้ $I(S_{1j},\ldots,S_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2 (P_{ij})$

และจากสูตรที่ 1 จะได้ $P_{ij} = \frac{S_{ij}}{|S_j|}$

จากสูตรที่ 2 และ สูตรที่ 3 จะได้สูตรการคำนวณค่า Gain ของ Attribute A ดังนี้

$$Gain(A) = I(S_1,S_2,\ldots,S_m) - E(A) \qquad \ldots\ldots\ldots\ldots (4)$$

เพื่อความเข้าใจยิ่งขึ้นจะขอยกตัวอย่างการวิเคราะห์ด้วยเทคนิค Decision Tree ซึ่งเป็นข้อมูลของลูกค้าที่ซื้อเครื่องใช้ไฟฟ้าที่เก็บอยู่ในฐานข้อมูล เพื่อพิจารณาว่าลูกค้ารายใดจะซื้อเครื่องคอมพิวเตอร์ และลูกค้ารายใดไม่ซื้อเครื่องคอมพิวเตอร์ โดยใช้ Algorithm ID3 ซึ่งจะคำนวณจากสูตรดังกล่าวข้างต้น ดังตารางที่ 3.1

RID	AGE	INCOME	STUDENT	CREDIT_RATING	CLASS
1	<=30	High	no	Fair	not buy
2	<=30	High	no	Excellent	not buy
3	31…40	High	no	Fair	buy
4	>40	medium	no	Fair	buy
5	>40	low	yes	Fair	buy
6	>40	low	yes	Excellent	not buy
7	31…40	low	yes	Excellent	buy
8	<=30	medium	no	Fair	not buy
9	<=30	low	yes	Fair	buy
10	>40	medium	yes	Fair	buy
11	<=30	medium	yes	excellent	buy
12	31…40	medium	no	excellent	buy
13	31…40	high	yes	fair	buy
14	>40	medium	no	excellent	not buy

ตารางที่ 3.1 Training data tuples from the AllElectronics Customer database

จากข้อมูลตัวอย่างในตารางที่ 3.1 มีจำนวนข้อมูลทั้งหมด 14 รายการ (S = 14) โดย Attribute "CLASS" ถือว่าเป็น "Target Attribute" เป็น Attribute เป้าหมายที่เราสนใจสามารถแยก CLASS ออกเป็น 2 CLASS คือ

- CLASS ที่ลูกค้าซื้อเครื่องคอมพิวเตอร์ (buy) ซึ่งมีข้อมูลใน CLASS นี้อยู่ 9 รายการ ($S_1 = 9$)
- CLASS ที่ลูกค้าไม่ซื้อเครื่องคอมพิวเตอร์ (not buy) ซึ่งมีข้อมูลใน CLASS นี้อยู่ 5 รายการ ($S_2 = 5$)

หาค่า $I(S_1,S_2)$ จากสูตรที่ 2 โดยที่ค่า $P_1 = 9/14$, $P_2 = 5/14$ จะได้ว่า

$$I(S_1,S_2) = -(9/14) \times \log_2(9/14) - (5/14) \times \log_2(5/14)$$

$$= 0.940$$

คำนวณค่า Gain Attribute "AGE" โดยพิจารณา Attribute "AGE" สามารถแบ่งเป็น 3 กลุ่มได้แก่

- AGE ที่น้อยกว่าหรือเท่ากับ 30 (AGE <= 30) มีอยู่จำนวน 5 รายการ ($S_1 = 5$) แบ่งเป็น CLASS ที่ buy 2 รายการ ($S_{11} = 2$) และ CLASS ที่ not buy 3 รายการ ($S_{21} = 3$) คำนวณค่า $I(S_{11},S_{21})$ โดย $P_{11} = 2/5$, $P_{21} = 3/5$ จะได้ว่า

$$I(S_{11},S_{21}) = -(2/5) \times \log_2(2/5) - (3/5) \times \log_2(3/5)$$
$$= 0.971$$

- AGE ที่น้อยกว่าหรือเท่ากับ 30 (30<AGE<=40) มีอยู่จำนวน 4 รายการ ($S_2 = 4$) แบ่งเป็น CLASS ที่ buy 4 รายการ ($S_{12} = 4$) และ CLASS ที่ not buy 0 รายการ ($S_{22} = 0$) คำนวณค่า $I(S_{12},S_{22})$ โดย $P_{12} = 4/4$, $P_{22} = 0/4$ จะได้ว่า

$$I(S_{12},S_{22}) = -(4/4) \times \log_2(4/4) - (0/4) \times \log_2(0/4)$$
$$= 0$$

- AGE ที่น้อยกว่าหรือเท่ากับ 30 (AGE > 40) มีอยู่จำนวน 5 รายการ ($S_3 = 5$) แบ่งเป็น CLASS ที่ buy 3 รายการ ($S_{13} = 3$) และ CLASS ที่ not buy 2 รายการ ($S_{23} = 2$) คำนวณค่า $I(S_{13},S_{23})$ โดย $P_{13} = 3/5$, $P_{23} = 2/5$ จะได้ว่า

$$I(S_{13},S_{23}) = -(3/5) \times \log_2(3/5) - (2/5) \times \log_2(2/5)$$
$$= 0.971$$

จากสูตรที่ 3 Attribute "AGE" 3 กลุ่ม นำมาคำนวณค่า E(AGE) ได้ดังนี้

$$E(AGE) = ((5/14) \times I(S_{11},S_{21})) + ((4/14) \times I(S_{12},S_{22})) + ((5/14) \times I(S_{13},S_{23}))$$
$$= ((5/14) \times 0.971)) + ((4/14) \times 0)) + ((5/14) \times 0.971))$$
$$= 0.694$$

ดังนั้นสามารถคำนวณค่า Gain ของ Attribute "AGE" ได้จากสูตรที่ 4 ดังนี้

$$Gain(AGE) = I(S_1,S_2) - E(AGE)$$
$$= 0.940 - 0.694 = 0.246$$

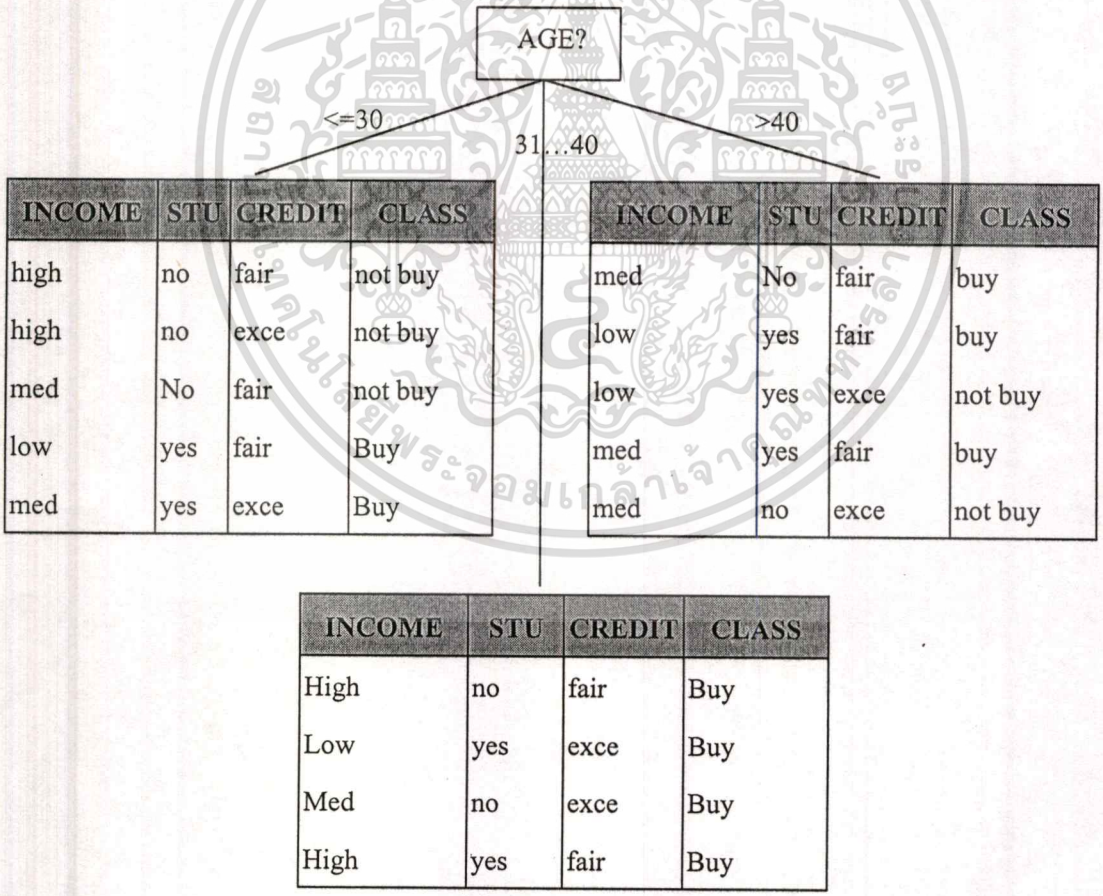
ผลจากการคำนวณค่า Gain ของ Attribute "AGE" เท่ากับ 0.246 ในทำนองเดียวกันเมื่อคำนวณค่า Gain ของ Attribute อื่น ๆ จะได้ค่าต่าง ๆ ดังนี้

- Gain(INCOME) เท่ากับ 0.029
- Gain(STAUDENT) เท่ากับ 0.151
- Gain(CREDIT_RATING) เท่ากับ 0.048

เมื่อพิจารณาค่า Gain ของ Attribute ทุกตัวที่คำนวณได้ จะพบว่า Gain(AGE) มีค่าสูงสุด เพราะฉะนั้นจึงเลือก Attribute "AGE" ให้เป็น Root node เนื่องจากผลการคำนวณค่า Gain แสดงให้เห็นว่า "AGE" มีความสำคัญที่สุด โดยแยก Tree ออกเป็น 3 Sub Node คือ

- กลุ่มที่อายุน้อยกว่า 30
- กลุ่มที่อายุระหว่าง 31 ถึง 40
- กลุ่มที่อายุมากกว่า 40

เมื่อพิจารณา CLASS ที่อายุระหว่าง 30 ถึง 40 ทั้ง CLASS จะเห็นว่าเป็น buy ทั้งหมด ดังนั้นจึงถือว่าเป็น Leaf Node ดังรูปที่ 3.2



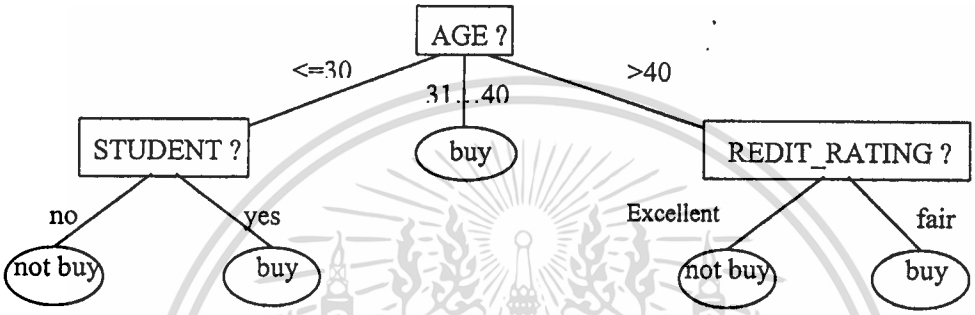
INCOME	STU	CREDIT	CLASS
high	no	fair	not buy
high	no	exce	not buy
med	No	fair	not buy
low	yes	fair	Buy
med	yes	exce	Buy

INCOME	STU	CREDIT	CLASS
High	no	fair	Buy
Low	yes	exce	Buy
Med	no	exce	Buy
High	yes	fair	Buy

INCOME	STU	CREDIT	CLASS
med	No	fair	buy
low	yes	fair	buy
low	yes	exce	not buy
med	yes	fair	buy
med	no	exce	not buy

รูปที่ 3.2 แสดงผลการเลือก Attribute "AGE" ในขั้นแรกเพื่อทำการวิเคราะห์แยก Sub Node ย่อยต่อไป

จากรูปที่ 3.2 เมื่อเลือก Attribute "AGE" เป็น Root Node แล้ว หลังจากนั้นคำนวณค่า Gain ของ Attribute ที่เหลือด้วยวิธีการเดิมจนกระทั่งแต่ละ CLASS มีค่าเป็นค่าเดียวกันหมด ซึ่งถือว่าเป็น Leaf Node ผลขั้นสุดท้ายของการแจกแจง Decision Tree จากข้อมูลตัวอย่างจะเป็นลักษณะดังรูปที่ 3.3



รูปที่ 3.3 Decision Tree ที่แจกแจงได้จากตัวอย่างข้อมูลลูกค้าที่จะซื้อหรือไม่ซื้อเครื่องคอมพิวเตอร์

จากการศึกษาอัลกอริทึม ID3 พบว่ามีข้อจำกัดในเรื่องของข้อมูลที่นำมาสร้าง Decision Tree กรณีเป็นข้อมูลต่อเนื่อง หรือมีความแตกต่างหลายระดับในการแบ่งกลุ่ม สามารถแก้ปัญหาดังกล่าว โดยการ Grouping ข้อมูลเป็นกลุ่มใหญ่ๆ ให้มีจำนวนกลุ่มไม่มาก ซึ่งในแต่ละกลุ่มจะมีการกระจายข้อมูลที่ใกล้เคียงกัน ทำให้ข้อมูลที่มีลักษณะต่อเนื่องถูกแปลงเป็นข้อมูลไม่ต่อเนื่อง

บทที่ 4
การวิเคราะห์ความเสี่ยงสินเชื่อด้วยกระบวนการดาต้าไมนิ่ง

ในการดำเนินธุรกรรมของธนาคารปัจจุบันมีความเสี่ยงในหลายๆ ด้าน แต่ในโครงการพิเศษฉบับนี้จะขอทำการวิเคราะห์ความเสี่ยงทางด้านสินเชื่อ เพื่อให้สามารถติดตามหนี้ได้อย่างมีประสิทธิภาพ โดยอาศัยกระบวนการดาต้าไมนิ่ง ดังต่อไปนี้

4.1 กำหนดวัตถุประสงค์และปัญหาที่ต้องการวิเคราะห์

เพื่อวางแผนกลยุทธ์ในด้านการติดตามหนี้โดยความระมัดระวัง และเอาใจใส่ในการติดตามลูกหนี้ที่มีความเสี่ยง ได้อย่างเหมาะสมทั้งนี้เพื่อลด NPL ของธนาคาร

4.2 การเตรียมข้อมูล

ข้อมูลที่ใช้ในการศึกษาความเสี่ยงสินเชื่อธนาคารสำหรับโครงการพิเศษฉบับนี้ได้มาจากฝ่ายบริหารความเสี่ยงของธนาคารแห่งหนึ่ง ซึ่งเป็นหน่วยงานที่ดูแลเกี่ยวกับการวิเคราะห์ความเสี่ยงต่างๆ โดยข้อมูลจะถูกเลือกมาจาก Data Warehouse ของธนาคาร เป็นข้อมูลที่เกี่ยวข้องกับข้อมูลทางด้านสินเชื่อบุคคล โดยมีรายละเอียด ดังนี้

- ข้อมูลพฤติกรรมการชำระหนี้ (Transactions) ของลูกค้าสินเชื่อจากระบบ Credit Rating and Collection System (CLS) ได้แก่ จำนวนครั้งที่ผ่อนใน 1 ปี, ช่องทางวิธีการชำระ และจำนวนวันที่ชำระช้ากว่ากำหนด เป็นต้น
- ข้อมูล Profile ที่มาจากระบบ Personal Processing System (PLPS) ได้แก่ ประเภทผู้กู้, เงินเดือนเฉลี่ยของผู้กู้, อายุเฉลี่ยของผู้กู้, ประเภทหลักประกัน, เพศ เป็นต้น

ข้อมูลที่ใช้ในการศึกษาวิเคราะห์เหล่านี้เป็นข้อมูลที่คาดว่าจะเกี่ยวข้องกับการวิเคราะห์ความเสี่ยงสินเชื่อ ซึ่งเป็นข้อมูลย้อนหลังที่ศึกษาพฤติกรรมในช่วงเดือนมกราคม พ.ศ. 2545 ถึงเดือนพฤศจิกายน พ.ศ. 2545 จำนวน 1,000 รายการ โดยมีรายละเอียดของ Attribute ต่าง ๆ ดังตารางที่ 4.1 และค่าที่เป็นไปได้ของ Attribute ดังตารางที่ 4.2 ถึงตารางที่ 4.9 และจะใช้พยากรณ์เหตุการณ์อนาคตในอีก 2 เดือนล่วงหน้า คือ เดือนธันวาคม พ.ศ. 2545 ถึง เดือนมกราคม พ.ศ. 2546 ว่าลูกค้ารายใดมีโอกาสจะเป็นหนี้เสียที่ก่อให้เกิด NPL ต่อไป เพื่อจะได้ติดตามหนี้ได้อย่างเหมาะสม

ATTRIBUTE (ต่อ)	รายละเอียด
เพศของผู้กู้	"1" = หญิง, "2" = ชาย
ระดับการศึกษาของผู้กู้	(ตารางที่ 4.2)
อาชีพของผู้กู้	(ตารางที่ 4.3)
ระดับตำแหน่งงานของผู้กู้	(ตารางที่ 4.4)
สถานะภาพสมรสของผู้กู้	(ตารางที่ 4.5)
สถานะการอยู่อาศัยของผู้กู้	(ตารางที่ 4.6)
อายุสัญญาเงินกู้	วันสิ้นสุดสัญญา – วันเริ่มทำสัญญา (ตารางที่ 4.7)
อายุบัญชีเงินกู้	วันที่ปัจจุบัน - วันที่เปิดบัญชี (ตารางที่ 4.8)
อัตราดอกเบี้ยเฉลี่ยต่อปี	(ตารางที่ 4.9)
วงเงินกู้	เงินต้นรับ + ยอดหนี้คงเหลือ (ตารางที่ 4.10)
จำนวนครั้งที่ผ่อนต่อปี	(ตารางที่ 4.11)
ตัดชำระเงินกู้ผ่านระบบ ATS	"Y"= มี ATS, "N"=ไม่มี ATS
จำนวนครั้งที่ชำระเงินกู้ช้ากว่ากำหนด	(ตารางที่ 4.12)
จำนวนวันเฉลี่ยที่ชำระเงินกู้ช้ากว่ากำหนด	(ตารางที่ 4.13)
ช่องทางการชำระเงินกู้	ATS, ATM, COUNTER (ตารางที่ 4.14)
วิธีการชำระเงินกู้	(ตารางที่ 4.15)
ประเภทหลักประกัน	(ตารางที่ 4.16)
พฤติกรรมการชำระเงินกู้ในรอบ 1 ปี	(ตารางที่ 4.17)
ประเภทผู้กู้	"1" = กู้เดี่ยว, "2" = กู้ร่วม
เงินเดือนของผู้กู้ในบัญชี	กรณีกู้ร่วมคิดค่าเฉลี่ยเงินเดือน (ตารางที่ 4.18)
อายุของผู้กู้ในบัญชี	กรณีกู้ร่วมคิดค่าเฉลี่ยอายุ (ตารางที่ 4.19)
อายุงานของผู้กู้ในบัญชี	กรณีกู้ร่วมคิดค่าเฉลี่ยอายุงาน (ตารางที่ 4.20)
ประเภทลูกหนี้	'1' = Good-ลูกค้าที่มีความเสี่ยงต่ำ, '2' = Bad- ลูกค้าที่มีความเสี่ยงสูง

ตารางที่ 4.1 แสดง Attribute ของข้อมูลที่ใช้ในการวิเคราะห์ ความเสี่ยงสินเชื่อธนาคาร

VALUE	รายละเอียด
1	สูงกว่าระดับปริญญาตรี
2	ระดับปริญญาตรี
3	ต่ำกว่าระดับปริญญาตรี

ตารางที่ 4.2 แสดงค่า Attribute ระดับการศึกษาของผู้กู้

VALUE	รายละเอียด
10	ข้าราชการ, รัฐวิสาหกิจ
20	พนักงานบริษัทเอกชน
30	ผู้ประกอบอาชีพอิสระส่วนตัว
40	อื่นๆ

ตารางที่ 4.3 แสดงค่า Attribute อาชีพของผู้กู้

VALUE	รายละเอียด
1	พนักงานปฏิบัติการ
2	ผู้บริหารระดับล่าง
3	ผู้บริหารระดับกลาง
4	ผู้บริหารระดับบน

ตารางที่ 4.4 แสดงค่า Attribute ระดับตำแหน่งงานของผู้กู้

VALUE	รายละเอียด
1	โสด
2	แต่งงาน
3	ม่าย

ตารางที่ 4.5 แสดงค่า Attribute สถานะภาพสมรสของผู้กู้

VALUE	รายละเอียด
R1	อาศัยอยู่บ้านบิดามารดา
R2	อาศัยอยู่บ้านเช่า
R3	อาศัยอยู่บ้านตนเอง

ตารางที่ 4.6 แสดงค่าของ Attribute สถานะอยู่อาศัยของผู้กู้

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 10 ปี
2	มากกว่า 10 ปี แต่น้อยกว่าหรือเท่ากับ 20 ปี
3	มากกว่า 20 ปี

ตารางที่ 4.7 แสดงค่าของ Attribute อายุสัญญาเงินกู้

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 4 ปี
2	มากกว่า 4 ปี แต่น้อยกว่าหรือเท่ากับ 8 ปี
3	มากกว่า 8 ปี

ตารางที่ 4.8 แสดงค่าของ Attribute อายุบัญชีเงินกู้

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 9 % ต่อปี
2	มากกว่า 9 % ต่อปี แต่น้อยกว่าหรือเท่ากับ 12 % ต่อปี
3	มากกว่า 12 ปี

ตารางที่ 4.9 แสดงค่าของ Attribute อัตราดอกเบี้ยเฉลี่ยต่อปี

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 100,000.00 บาท
2	มากกว่า 100,000.00 บาท แต่น้อยกว่าหรือเท่ากับ 500,000.00 บาท
3	มากกว่า 500,000.00 บาท แต่น้อยกว่าหรือเท่ากับ 1,000,000.00 บาท
4	มากกว่า 1,000,000.00 บาท

ตารางที่ 4.10 แสดงค่าของ Attribute วงเงินกู้

VALUE	รายละเอียด
1	น้อยกว่า 12 ครั้งต่อปี
2	เท่ากับ 12 ครั้งต่อปี
3	มากกว่า 12 ครั้งต่อปี

ตารางที่ 4.11 แสดงค่าของ Attribute จำนวนครั้งที่ผ่อนต่อปี

VALUE	รายละเอียด
0	ไม่มีการผ่อนชำระช้า
1	ผ่อนชำระช้าน้อยกว่าหรือเท่ากับ 5 ครั้ง
2	ผ่อนชำระช้ามากกว่า 5 ครั้ง แต่น้อยกว่าหรือเท่ากับ 7 ครั้ง
3	ผ่อนชำระช้ามากกว่า 7 ครั้ง แต่น้อยกว่าหรือเท่ากับ 10 ครั้ง
4	มากกว่า 10 ครั้ง

ตารางที่ 4.12 แสดงค่าของ Attribute จำนวนครั้งที่ผ่อนชำระช้ากว่ากำหนด

VALUE	รายละเอียด
0	ไม่มีการผ่อนชำระช้า
1	ผ่อนชำระช้าน้อยกว่าหรือเท่ากับ 10 วัน
2	ผ่อนชำระช้ามากกว่า 10 วัน แต่น้อยกว่าหรือเท่ากับ 20 วัน
3	ผ่อนชำระช้ามากกว่า 20 วัน

VALUE	รายละเอียด
1	ผ่อนชำระผ่าน COUNTER
2	ผ่อนชำระผ่าน ATS
3	ผ่อนชำระผ่าน ATM

ตารางที่ 4.14 แสดงค่าของ Attribute ช่องทางการชำระเงินกู้

VALUE	รายละเอียด
1	CASH
2	CHEQUE IC
3	CHEQUE CL
4	TRANSFER

ตารางที่ 4.15 แสดงค่าของ Attribute วิธีการชำระเงินกู้

VALUE	รายละเอียด
1	สังหาริมทรัพย์
2	อสังหาริมทรัพย์
3	อื่นๆ

ตารางที่ 4.16 แสดงค่าของ Attribute ประเภทหลักประกัน

VALUE	รายละเอียด
1	ค้างชำระ, จ่ายล่ำซำ
2	ค้างชำระ, ไม่จ่ายล่ำซำ
3	ชำระสม่ำเสมอ, จ่ายล่ำซำ
4	ชำระสม่ำเสมอ, ไม่จ่ายล่ำซำ

ตารางที่ 4.17 แสดงค่าของ Attribute พฤติกรรมชำระเงินกู้ในรอบ 1 ปี

จากตารางที่ 4.17 เป็นการแสดงค่าของ Attribute พฤติกรรมชำระเงินในรอบ 1 ปี โดย

พฤติกรรมที่ 1 : ค้างชำระ, จ่ายล่ำซำ หมายถึง ลูกค้าที่ชำระค่างวดไม่ครบ 12 เดือนต่อปี แต่ยอดรวมค่างวดในรอบ 1 ปีมากกว่า 12 เท่าของอัตราผ่อนต่อเดือน

พฤติกรรมที่ 2 : ค้างชำระ, ไม่จ่ายล่ำซำ หมายถึง ลูกค้าที่ชำระค่างวดไม่ครบ 12 เดือนต่อปี และยอดรวมค่างวดในรอบ 1 ปีไม่มากกว่า 12 เท่าของอัตราผ่อนต่อเดือน

พฤติกรรมที่ 3 : ชำระสม่ำเสมอ, จ่ายล่ำซำ หมายถึง ลูกค้าที่ชำระค่างวดครบ 12 เดือนต่อปี และยอดรวมค่างวดในรอบ 1 ปีมากกว่า 12 เท่าของอัตราผ่อนต่อเดือน

พฤติกรรมที่ 4 : ชำระสม่ำเสมอ, ไม่จ่ายล่ำซำ หมายถึง ลูกค้าที่ชำระค่างวดครบ 12 เดือนต่อปี แต่ยอดรวมค่างวดในรอบ 1 ปีไม่มากกว่า 12 เท่าของอัตราผ่อนต่อเดือน

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 15,000.00 บาทต่อเดือน
2	มากกว่า 15,000.00 บาทต่อเดือน แต่น้อยกว่าหรือเท่ากับ 30,000.00 บาทต่อเดือน
3	มากกว่า 30,000 บาทต่อเดือน

ตารางที่ 4.18 แสดงค่าของ Attribute เงินเดือนของผู้กู้ในบัญชี

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 35 ปี
2	มากกว่า 35 ปี แต่น้อยกว่าหรือเท่ากับ 45 ปี
3	มากกว่า 45 ปี แต่น้อยกว่าหรือเท่ากับ 50 ปี
4	มากกว่า 50 ปี

ตารางที่ 4.19 แสดงค่าของ Attribute อายุของผู้กู้ในบัญชี

VALUE	รายละเอียด
1	น้อยกว่าหรือเท่ากับ 8 ปี
2	มากกว่า 8 ปี แต่น้อยกว่าหรือเท่ากับ 15 ปี
3	มากกว่า 15 ปี แต่น้อยกว่าหรือเท่ากับ 20 ปี
4	มากกว่า 20 ปี

ตารางที่ 4.20 แสดงค่าของ Attribute อายุงานของผู้กู้ในบัญชี

ในการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารนี้จะพิจารณา Attribute "ประเภทลูกหนี้" เป็น Target Attribute ในการสร้างแบบจำลองพยากรณ์ โดย Attribute "ประเภทลูกหนี้" มีค่า 2 ค่า คือ

'1' = Good หมายถึง ลูกค้าที่ดีมีความเสี่ยงในการเกิด NPL ต่ำ

'2' = Bad หมายถึง ลูกค้าที่ไม่ดีมีความเสี่ยงในการเกิด NPL สูง

ข้อมูลที่เลือกมาศึกษาวิเคราะห์ความเสี่ยงสินเชื่อเหล่านี้เป็นข้อมูลที่มีความสมบูรณ์ เนื่องจากได้รับการกลั่นกรองเรียบร้อยแล้วจากฝ่ายบริหารความเสี่ยงของธนาคาร เพราะฉะนั้นจึงไม่ Missing Value และ Noisy data โดยฝ่ายบริหารความเสี่ยงจะทำการกลั่นกรองข้อมูลด้วยระบบกึ่ง Manual โดยการอาศัย Feature ของ Software Mineset ช่วยในบางส่วนที่เกี่ยวกับการหาค่าทางสถิติของข้อมูล เช่น การหาค่าเฉลี่ย, ค่า Maximum, ค่า Minimum เป็นต้น เพื่อนำไปแทนที่ข้อมูลบางส่วนที่สูญหาย แต่ถ้าข้อมูลที่สูญหายนั้นมีปริมาณน้อยมากเมื่อเทียบกับข้อมูลทั้งหมด ทำให้ไม่มีผลกระทบในการวิเคราะห์ ฝ่ายบริหารความเสี่ยงก็จะตัดข้อมูลสูญหายนั้นทิ้ง ส่วนข้อมูลที่ผิดปกติเช่น เพศหญิงแทนด้วย 'F', ส่วนเพศชายแทนด้วย 'M' แต่ข้อมูลเพศกับมีค่าเป็น 'X' ก็จะทำการแก้ไขโดยดูจากค่า Attribute อื่น เช่น คำนำหน้าชื่อว่าเป็น นาย, นางสาว, นาง แล้วทำการแก้ไขให้ถูกต้อง นอกจากนั้นก็มีการแปลงค่าของข้อมูลบางค่าให้เหมาะสมด้วย เช่น กรณีแปลงข้อมูลต่อเนื่องให้เป็นข้อมูลไม่ต่อเนื่อง ทางฝ่ายบริหารความเสี่ยงจะมีวิธีการแบ่งช่วงข้อมูล 2 วิธี คือ

- แบ่งช่วงข้อมูลเป็นช่วงเท่าๆ กัน โดยไม่คำนึงถึงปริมาณข้อมูลที่กระจายในแต่ละช่วงว่าจะต้องเท่ากันหรือไม่
- แบ่งช่วงข้อมูลไม่เท่ากัน แต่จะพิจารณาปริมาณข้อมูลที่กระจายในแต่ละช่วงควรจะมีกระจายข้อมูลที่ค่อนข้างใกล้เคียงกัน

ซึ่งจากการศึกษาลักษณะการทำงานของฝ่ายบริหารความเสี่ยงมักจะนิยมใช้วิธีแรกคือแบ่งช่วงข้อมูลเท่าๆ กัน มากกว่าแบบแบ่งช่วงข้อมูลไม่เท่ากัน

ดังนั้นในการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารสำหรับโครงการพิเศษฉบับนี้ ผู้เขียนจึงได้ทำการแบ่งข้อมูลต่อเนื่องเป็นช่วงๆ โดยพยายามจัดกลุ่ม Attribute ที่มีลักษณะเป็นข้อมูลต่อเนื่อง เช่น รายได้ผู้กู้, วงเงินกู้ เป็นต้น โดยแบ่งเป็นช่วงๆ ให้การกระจายข้อมูลแต่ละช่วงค่อนข้างมีจำนวนเท่าๆ กัน

ในการพัฒนาโปรแกรมสำหรับใช้ในการวิเคราะห์ นั้นจะพัฒนาด้วย Delphi Version 6 และใช้ Microsoft Access 97 เป็นฐานข้อมูลในการเก็บข้อมูลชั่วคราวระหว่างที่โปรแกรมทำงาน ซึ่งข้อมูลที่จะเป็นข้อมูลนำเข้าสู่ระบบจะอยู่ในรูป Text File โดยโปรแกรมที่พัฒนาขึ้นเพื่อใช้วิเคราะห์นี้จะมีหน้าที่การทำงานหลัก ๆ 3 ฟังก์ชั่น ได้แก่

4.2.1 นำข้อมูลในรูป Text File มาทำการสร้างแบบจำลองด้วยอัลกอริทึม ID3 เพื่อใช้ในการพยากรณ์

4.2.2 ทดสอบและประเมินแบบจำลองที่สร้างขึ้น

4.2.3 นำแบบจำลองที่ได้มาพยากรณ์เหตุการณ์ในอนาคต

โปรแกรมที่พัฒนาขึ้นนี้จะไม่มีฟังก์ชันส่วนของการตรวจสอบหาค่าเฉลี่ยและค่าสถิติต่างๆ รวมทั้งไม่มีฟังก์ชันการจัดกลุ่มข้อมูล ดังนั้นการจัดกลุ่มข้อมูลในการวิเคราะห์ความเสี่ยงสินเชื่อในโครงการฉบับนี้จะจัดด้วยระบบ Manual ซึ่งอาศัย Software Microsoft Office ช่วย เนื่องจากปริมาณข้อมูลมีจำนวนไม่มาก และถือว่าข้อมูลมีความสมบูรณ์ค่อนข้างมาก ข้อมูลที่ใช้ในการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารจะแบ่งข้อมูลเป็น 2 ชุด ดังนี้

ข้อมูลชุดที่ 1 เป็นข้อมูลที่ใช้ในการสร้างแบบจำลอง (Training Data Set) มีปริมาณข้อมูล 800 รายการ เมื่อพิจารณา Attribute "ประเภทลูกหนี้" (Traget Attribute) จะแบ่งเป็น

- ลูกค้าที่ดีมีความเสี่ยงในการเกิด NPL ต่ำ จำนวน 570 รายการ คิดเป็น 71.25 %
- ลูกค้าที่ไม่ดีมีความเสี่ยงในการเกิด NPL สูง จำนวน 230 รายการ คิดเป็น 28.75 %

ข้อมูลชุดที่ 2 เป็นข้อมูลที่ใช้ในการสร้างแบบจำลอง (Test Data Set) มีปริมาณข้อมูล 200 รายการ เมื่อพิจารณา Attribute "ประเภทลูกหนี้" (Traget Attribute) จะแบ่งเป็น

- ลูกค้าที่ดีมีความเสี่ยงในการเกิด NPL ต่ำ จำนวน 154 รายการ คิดเป็น 77 %
- ลูกค้าที่ไม่ดีมีความเสี่ยงในการเกิด NPL สูง จำนวน 46 รายการ คิดเป็น 23 %

4.3 การเลือกใช้อัลกอริทึม

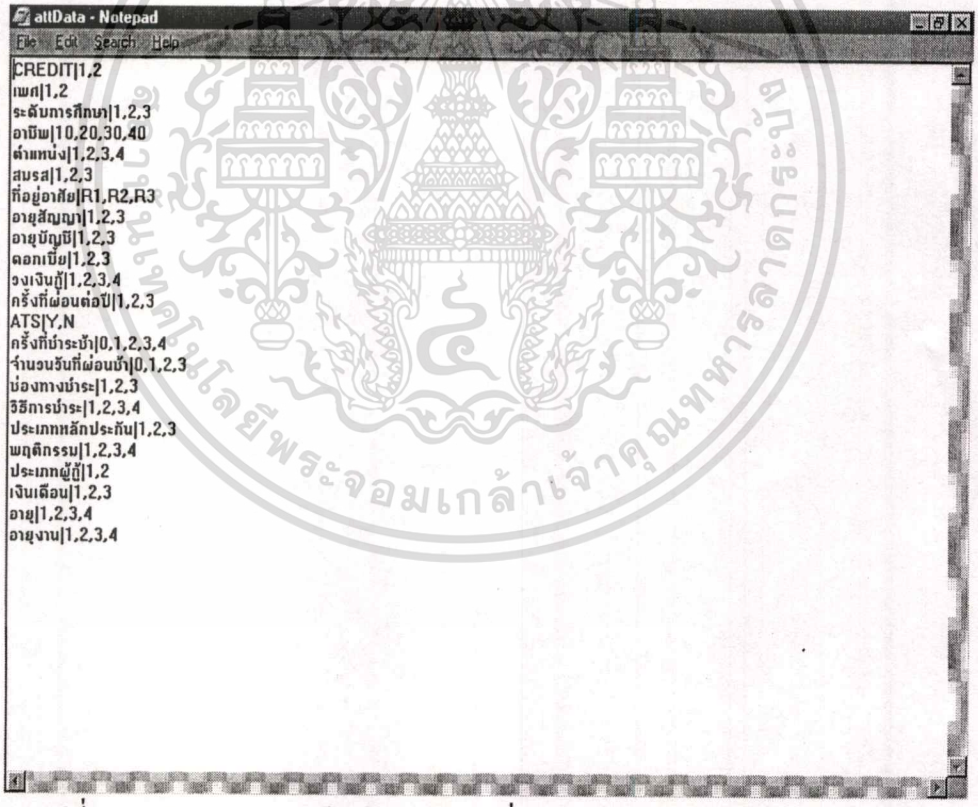
สำหรับในโครงการพิเศษฉบับนี้ได้ทำการศึกษาและใช้หลักการของอัลกอริทึม ID3 ซึ่งถือว่าเป็นอัลกอริทึมพื้นฐาน โดยอัลกอริทึม ID3 นี้มีข้อจำกัดในเรื่องข้อมูลที่ใช้ในการวิเคราะห์ไม่

ควรเป็นข้อมูลต่อเนื่อง ดังนั้นในขั้นตอนการเตรียมข้อมูลจึงได้ทำการแบ่งข้อมูลต่อเนื่องเป็นช่วงๆ เรียบร้อยแล้ว

4.4 การทำดาต้าไมนิ่ง

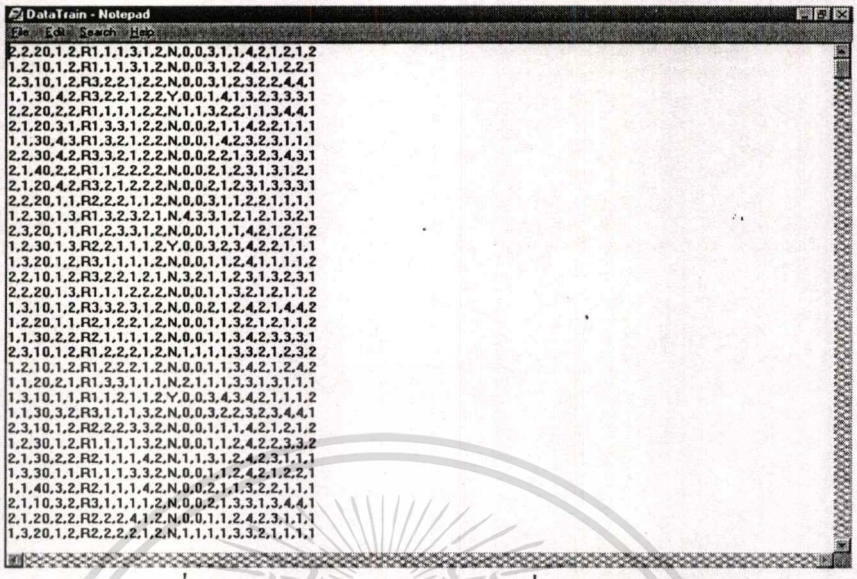
ขั้นตอนนี้จะนำโปรแกรมที่พัฒนามาทำการสร้างแบบจำลอง ซึ่งจำเป็นต้องอาศัยข้อมูล Text File 2 File ได้แก่

4.4.1 แฟ้มข้อมูลที่ระบุชื่อของ Attribute และระบุค่าที่เป็นไปได้ของ Attribute นั้นๆ โดย Traget Attribute จะถูกระบุไว้บรรทัดแรก แฟ้มข้อมูลประเภทนี้เป็นแฟ้มข้อมูลนามสกุล .att ลักษณะของแฟ้มข้อมูลจะเป็นลักษณะดังรูปที่ 4.1



รูปที่ 4.1 ลักษณะของแฟ้มข้อมูลอธิบายชื่อและค่าของ Attribute (*.att)

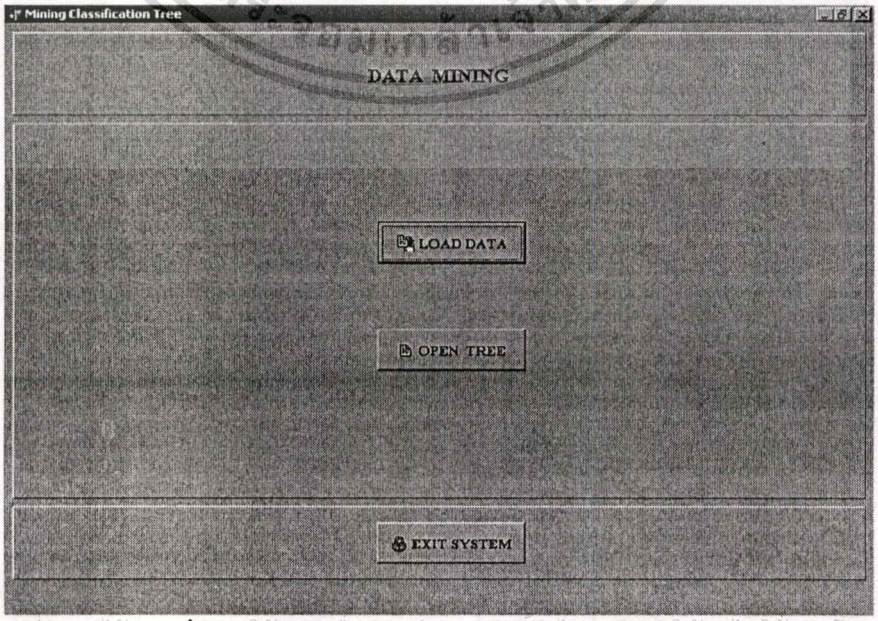
4.4.2 แฟ้มข้อมูลที่เป็นข้อมูลจริงที่ใช้วิเคราะห์ตาม Format ที่กำหนด เป็นแฟ้มข้อมูลนามสกุล .txt ลักษณะของแฟ้มข้อมูลจะมีลักษณะดังรูปที่ 4.2



รูปที่ 4.2 ลักษณะของแฟ้มข้อมูลที่ใช้ในการวิเคราะห์ (*.txt)

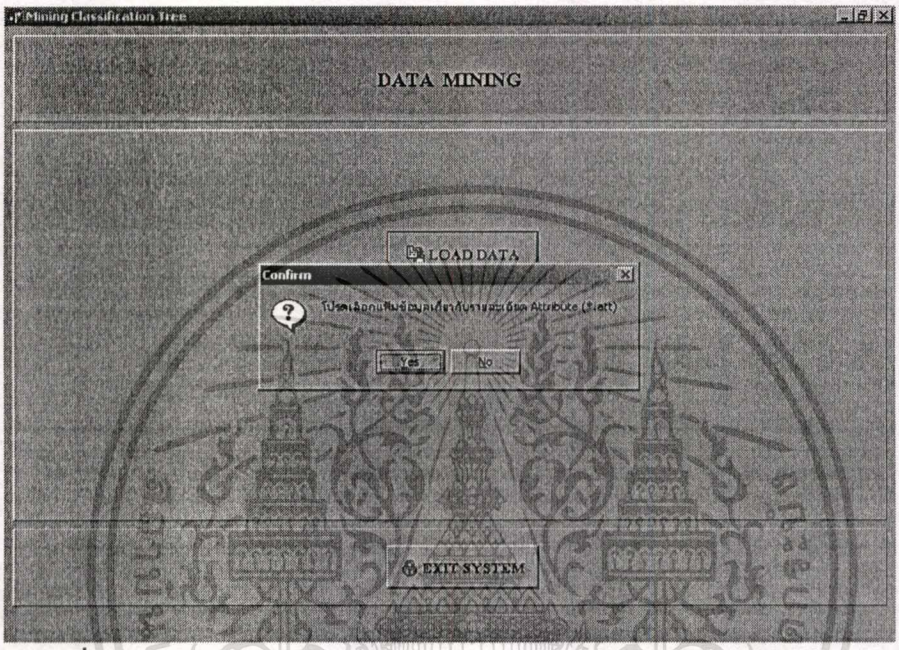
โปรแกรมจะมีเมนูหลัก 2 เมนู คือ LOAD DATA และ OPEN TREE (ดังรูปที่ 4.3) โดยมีรายละเอียด ดังนี้

- LOAD DATA ทำหน้าที่ LOAD ข้อมูลเพื่อนำมาสร้างแบบจำลองพร้อมทั้งประเมิน และทดสอบ เพื่อให้แบบจำลองมีความน่าเชื่อถือในระดับที่พอใจ และทำการ Save แบบจำลอง
- OPEN TREE ทำหน้าที่พยากรณ์ข้อมูลโดยอาศัยแบบจำลองในรูปของ Decision Tree ที่สร้าง

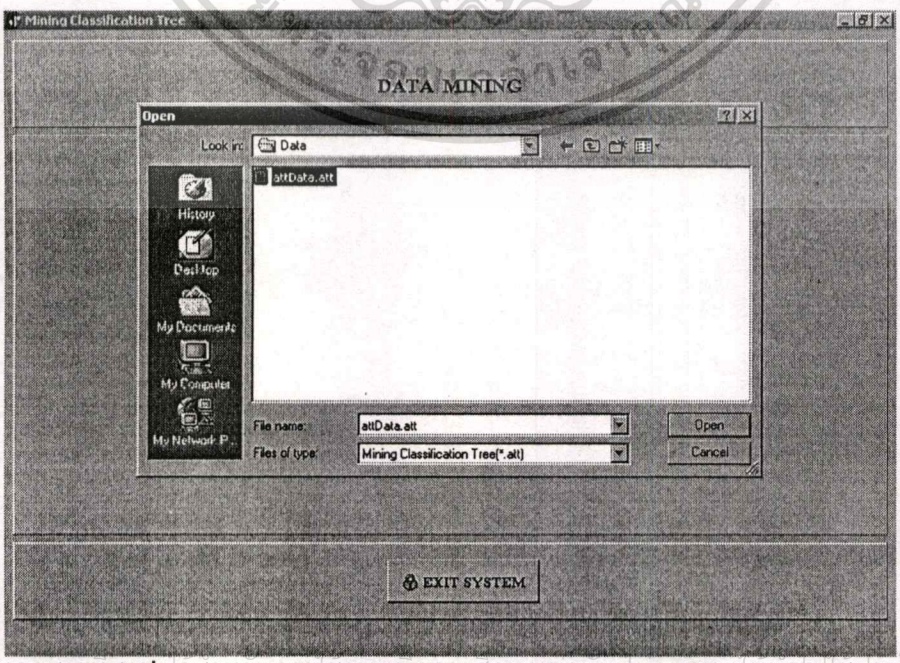


รูปที่ 4.3 เมนูหลักของโปรแกรมที่ใช้ในการวิเคราะห์

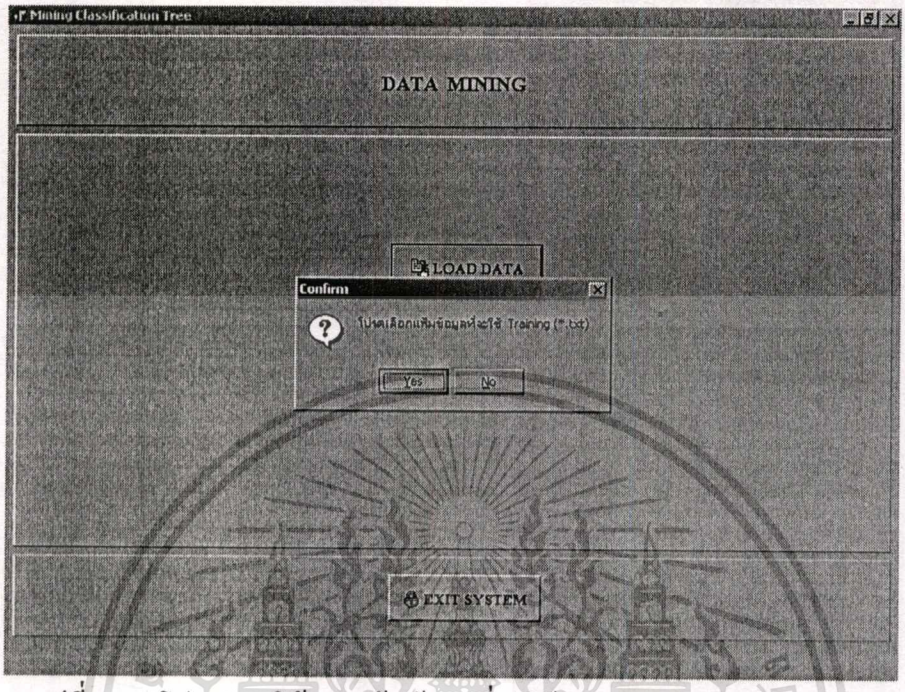
เมื่อเลือกเมนู LOAD DATA โปรแกรมจะให้ระบุแฟ้มข้อมูลอธิบาย Attribute (*.att) ดังรูปที่ 4.4 และรูปที่ 4.5 ต่อจากนั้นโปรแกรมก็จะให้ระบุแฟ้มข้อมูลที่เก็บข้อมูลที่จะนำมา Training เพื่อสร้าง Decision Tree ซึ่งอยู่ในรูป Text File (*.txt) ดังรูปที่ 4.6 และรูปที่ 4.7



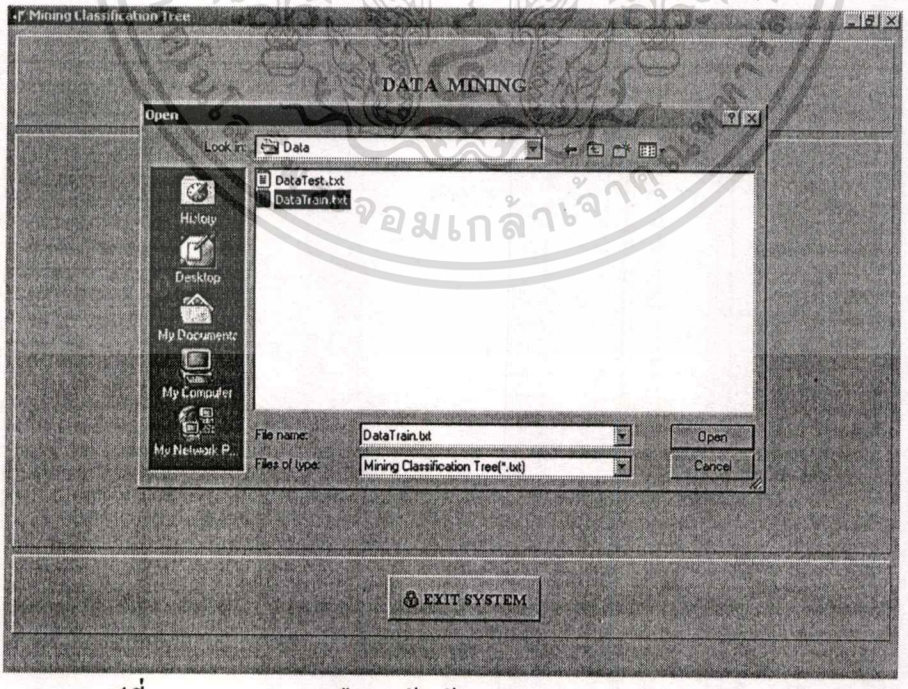
รูปที่ 4.4 โปรแกรมให้ระบุแฟ้มข้อมูลอธิบายลักษณะของ Attribute (*.att)



รูปที่ 4.5 แสดงการเลือกแฟ้มข้อมูลอธิบาย Attribute (attData.att)



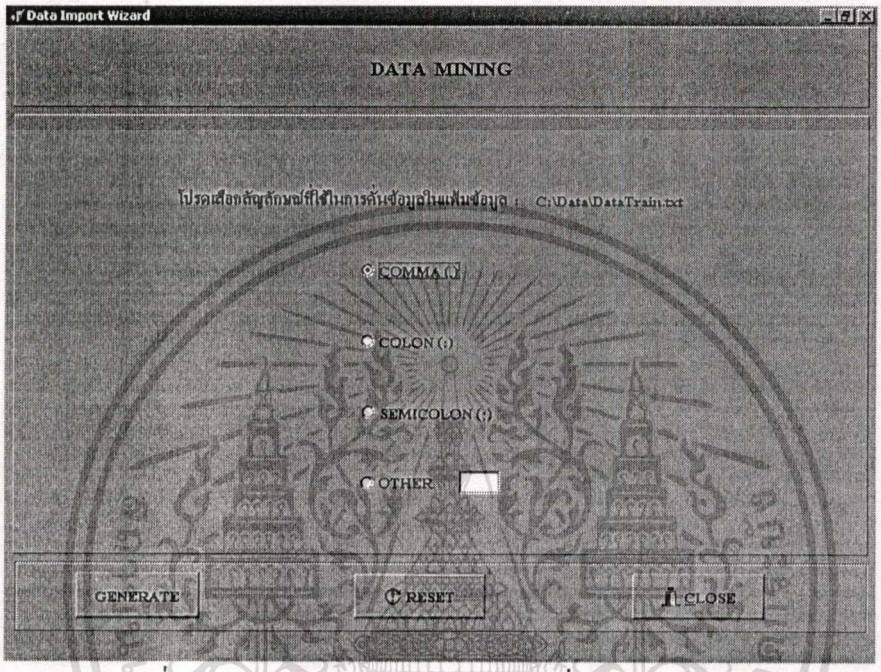
รูปที่ 4.6 โปรแกรมให้ระบุแฟ้มข้อมูลที่จะใช้ Training (*.txt)



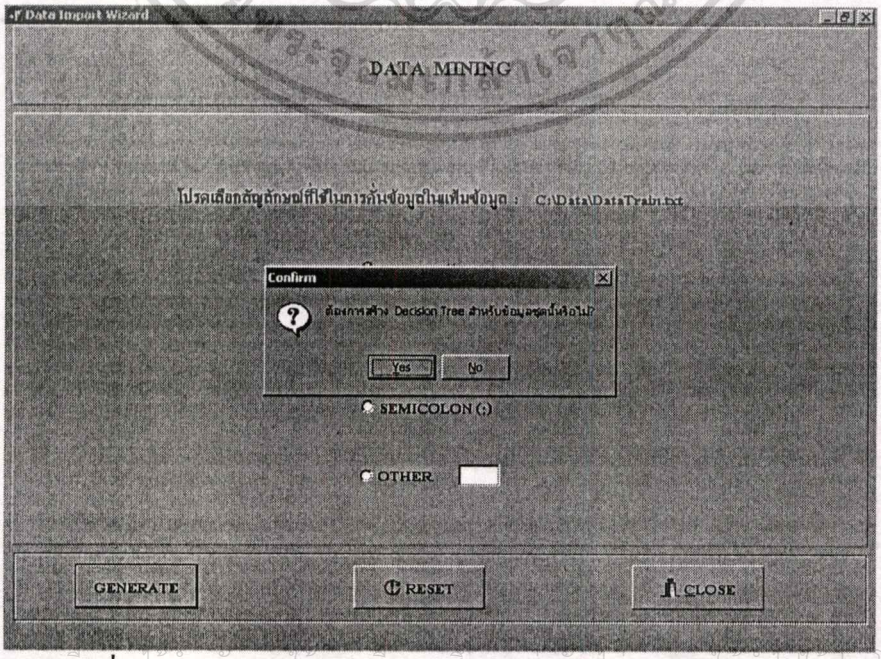
รูปที่ 4.7 แสดงการเลือกแฟ้มข้อมูล Training Data (DataTrain.txt)

เมื่อระบุแฟ้มข้อมูลที่ต้องการเรียบร้อยแล้ว จะต้องทำการเลือกสัญลักษณ์ตัวคั่นแต่ละ

เก็บข้อมูลที่จะวิเคราะห์อยู่ในรูปแบบ Text File หลังจากนั้นกดปุ่ม GENERATE เพื่อสร้าง Tree โดยโปรแกรมจะมีข้อความยืนยันความต้องการสร้าง Decision Tree อีกครั้ง ดังรูป 4.9

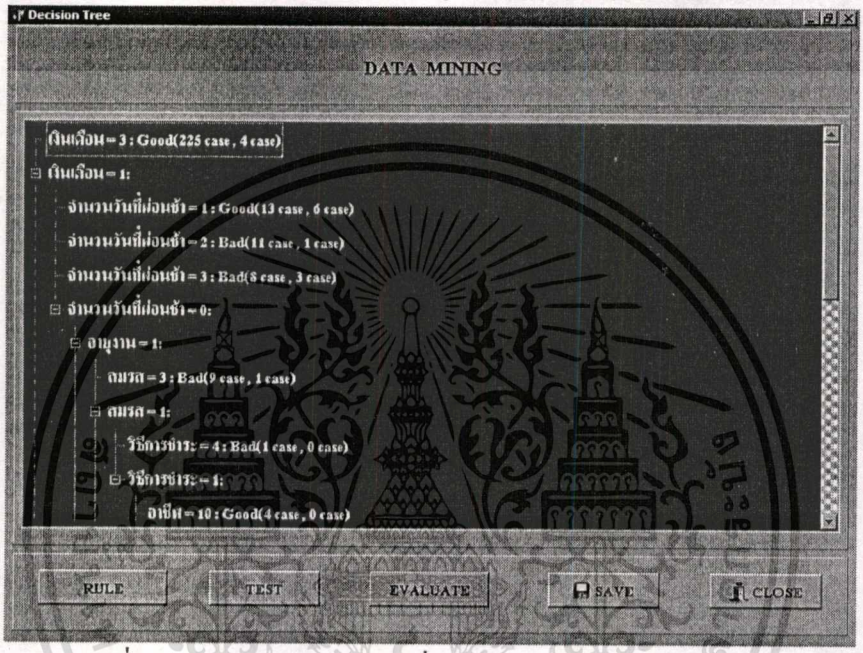


รูปที่ 4.8 แสดงการเลือกสัญลักษณ์ตัวคั่นแต่ละ Attribute

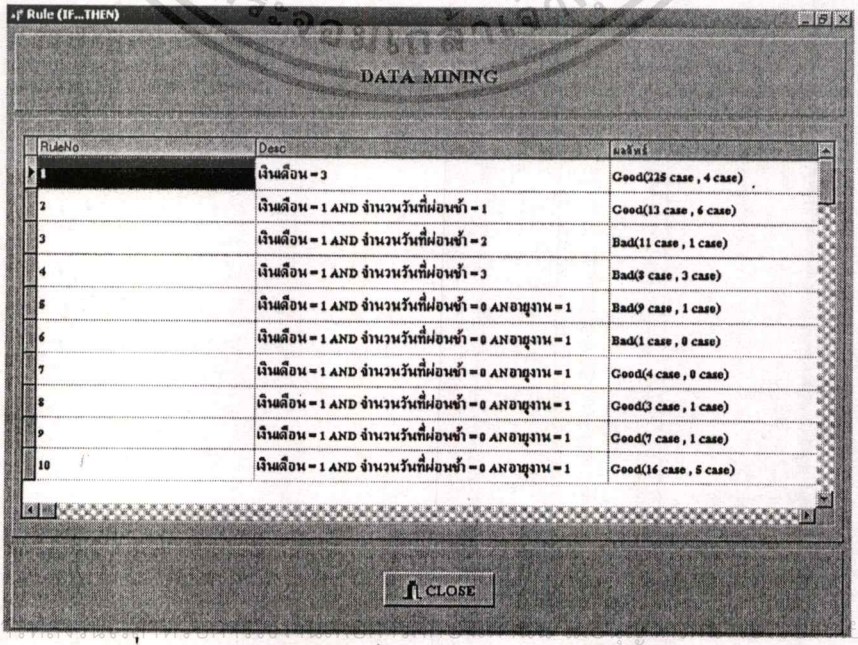


รูปที่ 4.9 แสดงข้อความยืนยันความต้องการสร้าง Decision Tree

หลังจากนั้นโปรแกรมก็จะทำการสร้าง Tree ดังรูปที่ 4.10 และจาก Decision Tree นี้สามารถแสดงผลลัพธ์ออกมาในรูปแบบกฎ IF...THEN ได้อีกรูปแบบ ดังรูปที่ 4.11 และจะเก็บค่าต่างๆ ในรูปแบบกฎ IF...THEN ไว้ในฐานข้อมูล Microsoft Access ชั่วคราว



รูปที่ 4.10 แสดงแบบจำลองที่สร้างในรูปแบบของ Tree



รูปที่ 4.11 แสดงผลลัพธ์ในรูปของกฎ IF...THEN

เมื่อสร้างแบบจำลอง Decision Tree ด้วยข้อมูล Training Data Set เสร็จเรียบร้อยแล้ว จึงทำการประเมินความแม่นยำของ Model โดยกดปุ่ม EVALUATE โปรแกรมจะทำการคำนวณความแม่นยำได้ค่าดังรูปที่ 4.12 ด้วยสูตร ดังนี้

จากสูตร Accuracy = Sensitivity * (pos/(pos+neg)) + Specificity * (neg/(pos+neg)) .....(5)

โดย Sensitivity = t_pos/pos

Specificity = t_neg/neg

เมื่อนำค่า Sensitivity และ Specificity แทนค่าในสมการที่ (5) จะได้

Accuracy = (t_pos/pos) * (pos/(pos+neg)) + (t_neg/neg) * (neg/(pos+neg))

เพราะฉะนั้นจะได้สูตรในการคำนวณค่าความแม่นยำดังสมการที่ (6)

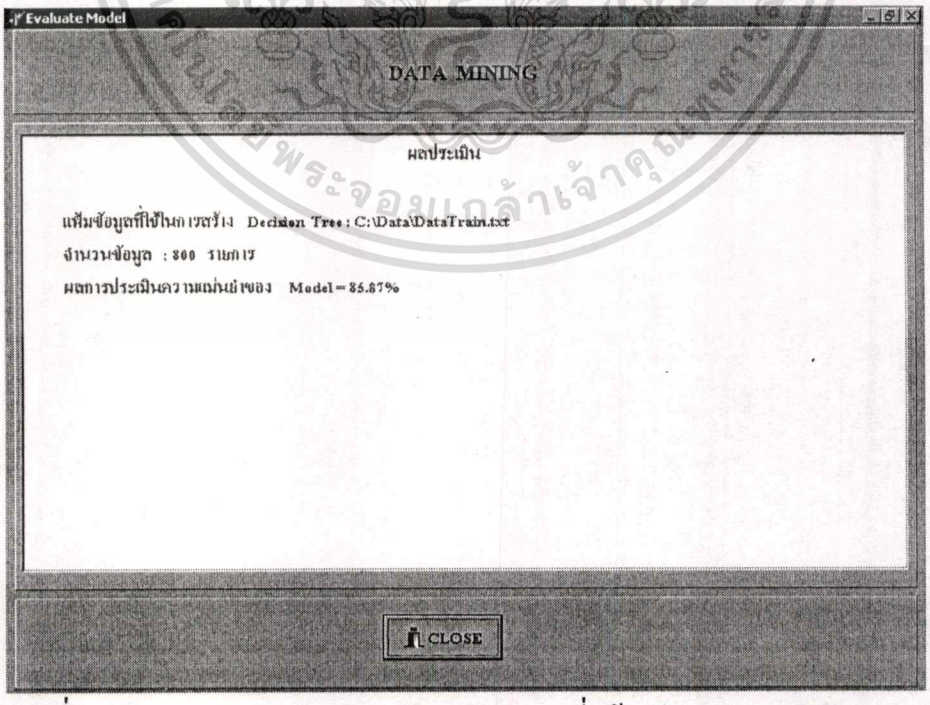
Accuracy = (t_pos/(pos+neg)) + (t_neg/(pos+neg)) .....(6)

โดยที่ t_pos หมายถึงจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive ที่สามารถทำนายกลุ่มได้ถูกต้อง

t_neg หมายถึงจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative ที่สามารถทำนายกลุ่มได้ถูกต้อง

pos หมายถึงข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive

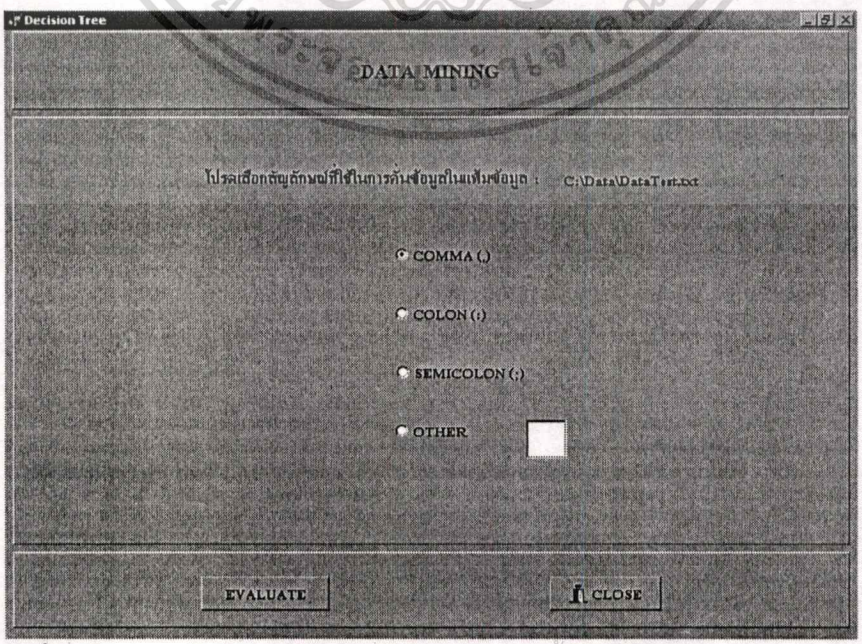
neg หมายถึงข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative



รูปที่ 4.12 แสดงผลของการประเมินแบบจำลองที่สร้างจาก Training Data Set

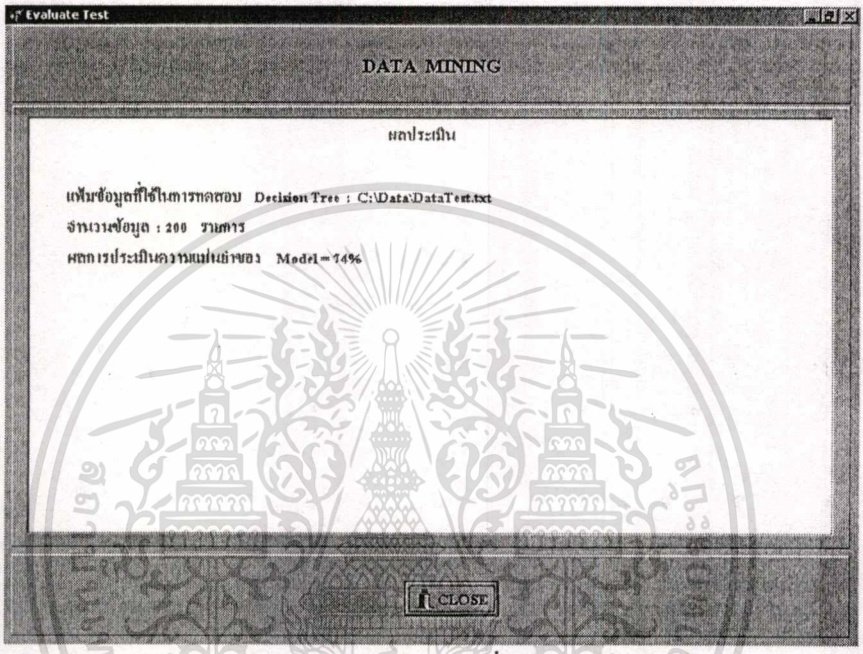
หลังจากประเมินความแม่นยำของแบบจำลองที่เป็น Decision Tree แล้ว จะต้องทำการทดสอบแบบจำลองที่ได้ด้วยข้อมูลชุดที่เป็น Test Data Set ซึ่งมีจำนวน 200 รายการ โดยกดปุ่ม TEST แล้วทำการเลือกแฟ้มข้อมูลที่จะใช้ทดสอบ ดังรูปที่ 4.13 พร้อมทั้งระบุสัญลักษณ์ที่ใช้คั่นแต่ละ Attribute ในแฟ้มข้อมูลสำหรับทดสอบ ดังรูป 4.14

รูปที่ 4.13 แสดงการเลือกแฟ้มข้อมูลสำหรับ Test (DataTest.txt)

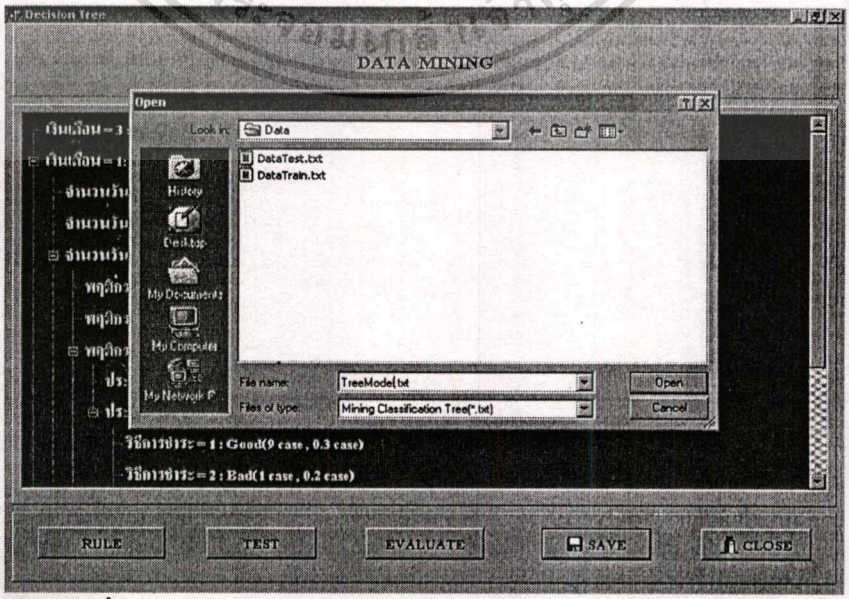


รูปที่ 4.14 แสดงการระบุสัญลักษณ์ที่ใช้คั่น Attribute ในแฟ้มข้อมูลทดสอบ

หลังจากนั้นโปรแกรมก็จะแสดงค่าที่ประเมินความแม่นยำของแบบจำลองที่ทำการทดสอบด้วย Test Data Set ได้ค่าเท่ากับ 74 % ดังรูป 4.15 นอกจากนั้นยังสามารถทำการ Save Decision Tree ที่สร้างได้ โดยการกดปุ่ม SAVE ดังรูป 4.16 โปรแกรมจะทำการเก็บ Decision Tree ไว้ในรูปของ Text File



รูปที่ 4.15 แสดงผลของการประเมินแบบจำลองที่ได้จากการใช้ Test Data Set ทดสอบ



รูปที่ 4.16 แสดงฟังก์ชันการ Save แบบจำลอง (TreeModel.txt)

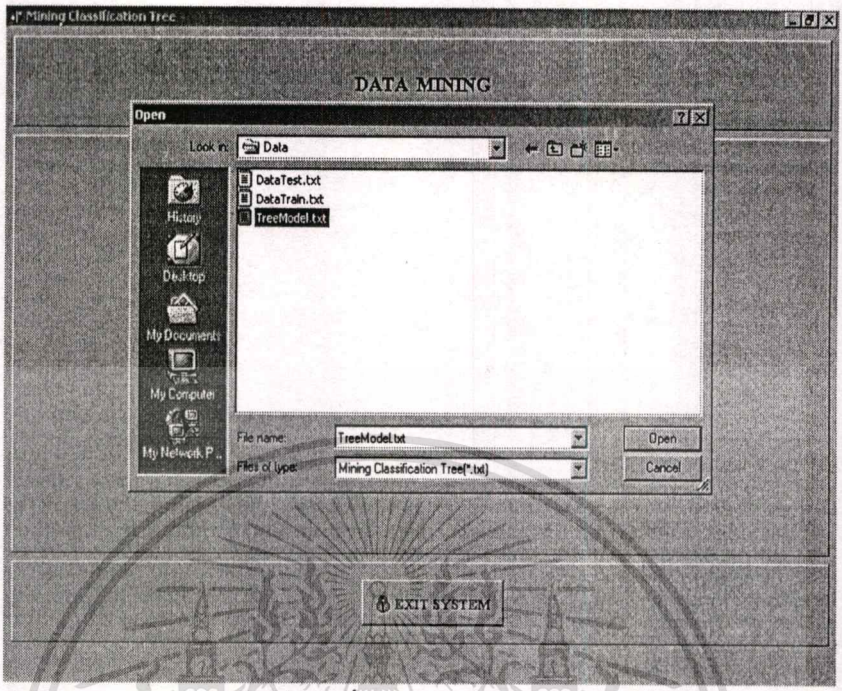
4.5 การวิเคราะห์ผลลัพธ์

จากการนำข้อมูลในอดีตมาวิเคราะห์โดยใช้โปรแกรมที่พัฒนาขึ้นมาเป็นเครื่องมือช่วยในการวิเคราะห์ ซึ่งโปรแกรมจะทำงานตามหลักการของอัลกอริทึม ID3 ได้ผลลัพธ์เป็นแบบจำลองอยู่ในรูปแบบ Decision Tree หรือรูปแบบกฎ IF...THEN เมื่อทำการประเมินความแม่นยำของ Decision Tree ที่สร้างได้จากข้อมูล Training Data Set จำนวน 800 รายการ ปรากฏว่ามีความถูกต้อง 85.87 % และเกิดความผิดพลาด 14.13 % ซึ่งถือว่าเป็นมีความน่าเชื่อถือในระดับที่ยอมรับได้ หลังจากนั้นจึงนำข้อมูลอีกชุดที่เตรียมไว้คือ Test Data Set จำนวน 200 รายการ มาทำการทดสอบ Decision Tree ปรากฏว่าสามารถพยากรณ์ข้อมูลทดสอบได้ถูกต้อง 74 % และพยากรณ์ผิดพลาด 26 % เมื่อพิจารณาผลลัพธ์ทำให้ยอมรับแบบจำลองดังกล่าวได้ ถึงแม้จะยังมีข้อผิดพลาดในการพยากรณ์ ทั้งนี้อาจเนื่องจากปริมาณข้อมูลที่ใช้ไม่มากพอส่งผลให้แบบจำลองที่ได้ไม่สามารถพยากรณ์ได้ถูกต้องสมบูรณ์ 100 % แต่ก็มีความน่าเชื่อถือได้ในระดับหนึ่งพอเป็นแนวทางในการทำนายความเสี่ยงในอนาคต เพื่อติดตามหนี้ของลูกค้าธนาคารต่อไป

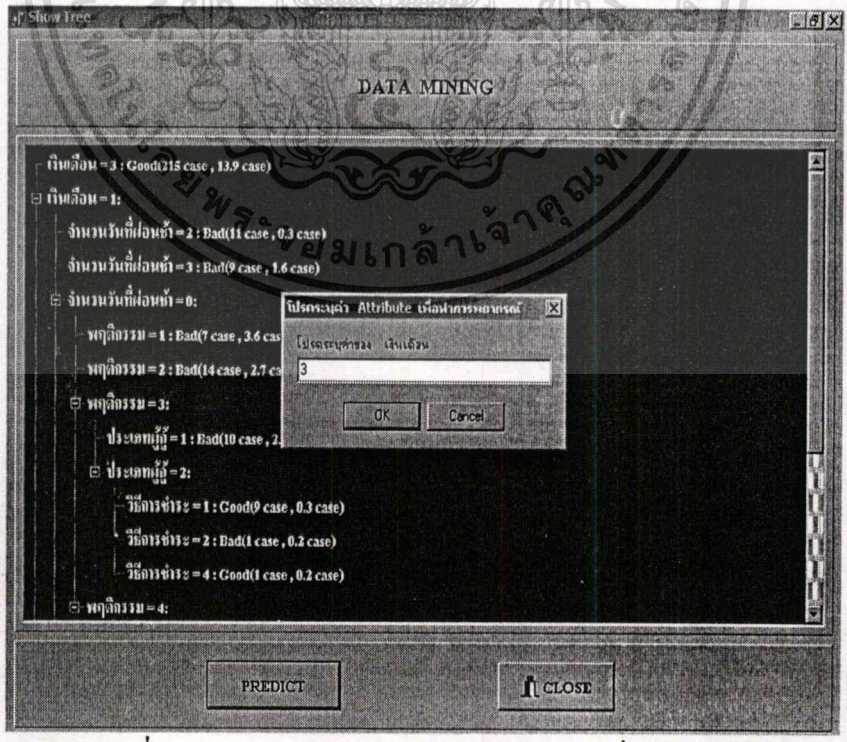
จาก Decision Tree ที่ได้จะเห็นว่าปัจจัยที่มีความสำคัญมากที่สุด (Root Node) คือ เงินเดือนของผู้กู้พบว่าถ้าผู้กู้มีรายได้สูงโอกาสที่จะเกิดเป็นสูญย่อมต่ำ ในขณะที่ผู้กู้มีรายได้ต่อโอกาสเกิดเป็นหนี้สูญจะสูง แต่ถึงอย่างไรก็ตามคงต้องพิจารณาปัจจัยสำคัญอื่นๆ รองลงมาประกอบโดยพิจารณาจากกิ่งของ Tree เช่น จำนวนวันที่ผ่อนชำระล่าช้า, พฤติกรรมผู้กู้, ประเภทผู้กู้ (กู้เดี่ยวหรือกู้ร่วม) เป็นต้น

4.6 การนำไปใช้งานจริง

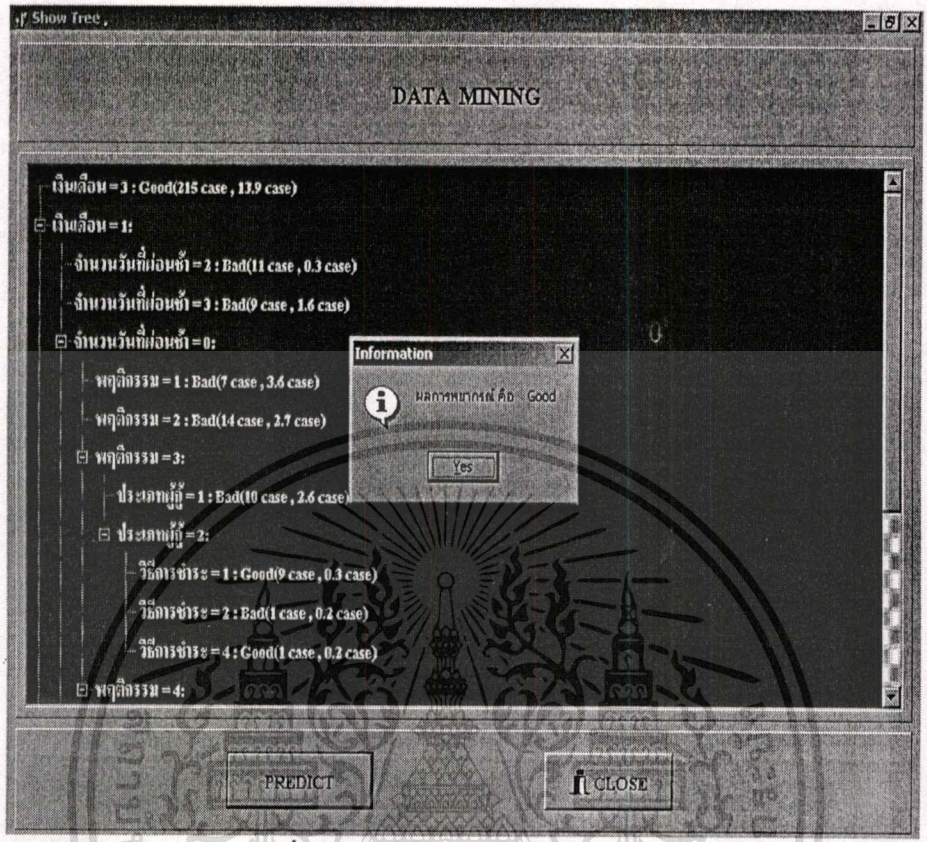
เมื่อต้องการทำการพยากรณ์ล่วงหน้าสำหรับลูกหนี้รายใดๆ ว่ามีโอกาสที่จะเกิดเป็นหนี้สูญหรือไม่ โดยอาศัยแบบจำลองที่สร้างและได้ทำการ Save ไว้แล้ว สามารถทำได้โดยเลือกเมนูหลัก OPEN TREE เพื่อเปิดแฟ้มข้อมูลที่เก็บ Decision Tree ออกมาใช้ในการพยากรณ์ ดังรูป 4.17 ซึ่งโปรแกรมจะแสดงจอภาพให้ระบุค่าของแต่ละ Attribute โดยเริ่มจาก Root Node ของ Tree และไล่ตามกิ่งของ Tree ไปเรื่อยตามค่าที่เราระบุให้ในแต่ละ Attribute ซึ่งจะเป็นส่วนที่กำหนดว่าจะไปยังกิ่งไหนบ้าง จนกระทั่งถึง Leaf Node จะแสดงผลลัพธ์ของการพยากรณ์ ตัวอย่างเช่น ถ้าเราระบุค่าของเงินเดือนผู้กู้ เท่ากับ '3' (หมายถึงผู้กู้มีรายได้มากกว่า 30,000 บาทต่อเดือน) ดังรูป 4.18 ผลลัพธ์การพยากรณ์ปรากฏว่า "GOOD" หมายถึงโอกาสเสี่ยงต่ำการเกิดหนี้สูญต่ำ ดังรูป 4.19



รูปที่ 4.17 แสดงการเปิดแฟ้มข้อมูลที่เก็บ Decision Tree ออกมาใช้ในการพยากรณ์



รูปที่ 4.18 แสดงการระบุค่าของ Attribute เพื่อพยากรณ์



รูปที่ 4.19 แสดงผลลัพธ์การพยากรณ์

บทที่ 5
บทสรุป

5.1 สรุปผลการศึกษา

ในโครงการพิเศษฉบับนี้ได้ทำการวิเคราะห์ความเสี่ยงสินเชื่อธนาคารแห่งหนึ่ง โดยอาศัยกระบวนการดาต้าไมนิ่ง เพื่อป้องกันและแก้ไขปัญหาการเกิดหนี้สูญ ซึ่งจะทำให้การติดตามหนี้เป็นไปอย่างมีประสิทธิภาพ ส่วนอัลกอริทึมที่ใช้ในการวิเคราะห์ในโครงการพิเศษฉบับนี้ ได้เลือกศึกษาอัลกอริทึม ID3 พบว่ามีข้อจำกัดในเรื่องของข้อมูลที่ใช้วิเคราะห์ไม่ควรเป็นข้อมูลที่ต่อเนื่องหรือมีกลุ่มมากจนเกินไป ดังนั้นผู้เขียนจึงได้ทำการเตรียมข้อมูลให้เป็นข้อมูลที่มีลักษณะเหมาะสมกับหลักการของอัลกอริทึมดังกล่าว และได้ทำการพัฒนาโปรแกรม เพื่อช่วยในการสร้างแบบจำลองที่จะใช้ในการพยากรณ์เหตุการณ์ล่วงหน้า และเมื่อประเมินความแม่นยำของ Decision Tree ที่ได้ผลปรากฏว่ามีความแม่นยำในการพยากรณ์ได้ถูกต้อง 85.87 % หลังจากนั้นจึงทำการทดสอบด้วยข้อมูลอีกชุดหนึ่งซึ่งเป็น Test Data Set ปรากฏว่ามีความแม่นยำในการพยากรณ์ได้ถูกต้อง 74 % แสดงว่าผลการพยากรณ์ยังมีโอกาสผิดพลาด ทั้งนี้อาจเนื่องมาจากปริมาณข้อมูลที่ใช้ในการวิเคราะห์มีปริมาณไม่เพียงพอ หรือกลุ่มข้อมูลที่สุ่มมาวิเคราะห์ซึ่งทำการเลือกมาจาก Data Warehouse อาจจะเป็นกลุ่มข้อมูลช่วงที่ขึ้นอยู่กับปัจจัยใดปัจจัยหนึ่งจนเกินไป ทำให้แบบจำลองที่ได้มีลักษณะดังที่สร้าง เพราะถ้าหากสุ่มข้อมูลจาก Data Warehouse อีกกลุ่ม ผลของแบบจำลองอาจจะเปลี่ยนแปลงไป

5.2 ข้อเสนอแนะ

โปรแกรมที่พัฒนาขึ้นมาเป็นเครื่องมือช่วยในการวิเคราะห์มีความยืดหยุ่นน้อย และไม่มีฟังก์ชันในส่วนที่จัดการเตรียมข้อมูล เช่นการหาค่าเฉลี่ย, ค่า Maximum, ค่า Minimum ต่างๆ รวมถึงการจัดกลุ่มข้อมูล เพื่อใช้ในขั้นตอนการจัดเตรียมข้อมูล

นอกจากนั้นในส่วนของข้อมูลที่ใช้วิเคราะห์ควรจะมีปริมาณมากกว่านี้ จะทำให้ได้ผลการวิเคราะห์ที่ถูกต้องและแม่นยำมากขึ้น และเนื่องจากเวลาในการวิเคราะห์โครงการพิเศษฉบับนี้ทำให้ศึกษาและใช้อัลกอริทึม ID3 เพียงอัลกอริทึมเดียวในการวิเคราะห์ ซึ่งอัลกอริทึมนี้เป็นอัลกอริทึมพื้นฐานพบว่ามีข้อจำกัดค่อนข้างมาก ในความเป็นจริงอาจจะลองวิเคราะห์ด้วยอัลกอริทึมอื่นๆ ที่มีการแก้ปัญหาข้อจำกัดของอัลกอริทึม ID3 เช่น อัลกอริทึม C4.5 ที่ปรับปรุงใช้วิเคราะห์

ข้อมูลต่อเนื่อง และข้อมูลที่มีบางค่าหายไปได้ แล้วเปรียบเทียบ Decision Tree ที่ได้ว่าควรจะเลือกอัลกอริทึมใดจึงจะเหมาะสม และมีผลการวิเคราะห์ที่น่าเชื่อถือมากที่สุด

นอกจากนั้นยังมีปัญหาที่สำคัญอีกอย่างหนึ่งของการสร้าง Decision Tree คือ การเกิด "overfits" กล่าวคือในการสร้าง Decision Tree จะมีการแตกออกเป็น Sub node เรื่อยๆ ซึ่งอาจทำให้ Decision Tree ที่ได้เกิดความซับซ้อนจนเกินไป ก่อให้เกิดความยุ่งยากในการทำนาย ซึ่งสาเหตุที่ทำให้เกิด overfit อาจจะเกิดจากข้อมูลที่ใช้วิเคราะห์มีสิ่งรบกวน (noise) วิธีที่จะแก้ปัญหา overfit ดังกล่าวคือ การตัด Sub node ที่ทำให้เกิดความผิดพลาดในการทำนายแล้วแทนที่ด้วย Leaf node ซึ่งวิธีการนี้เรียกว่า "การทำ Pruning" ซึ่งในโปรแกรมที่พัฒนาไม่ได้มีฟังก์ชันการทำ Pruning เนื่องจากเวลาในการพัฒนาค่อนข้างจำกัด ถ้าหากมีผู้สนใจจะทำการศึกษาต่อก็ควรจะทำการ Pruning ด้วย เพื่อให้ Decision Tree ที่สร้างมีประสิทธิภาพในการทำนายมากขึ้น

อย่างไรก็ตามถึงแม้ผลการวิเคราะห์จะมีความน่าเชื่อถือเพียงใด แต่ยอมต้องเกิดความผิดพลาดในการพยากรณ์ไม่มากก็น้อย ซึ่งคงต้องอาศัยความเชี่ยวชาญและความรอบรู้ธุรกิจหรือเรื่องนั้นๆ ประกอบในการช่วยพิจารณาในการตัดสินผลการวิเคราะห์ด้วย

บรรณานุกรม

Jiawei Han and Micheline Kamber. 2001. **Data Mining Concepts and Techniques.** United State of America : Morgan Kaufmann.

Michael J.A. Berry Gordon S. Linoff. 1997. **Data Mining Techniques for Marketing, Sales and Customer Support.** New York : John Wiley and Sons

Michael J.A. Berry Gordon S. Linoff. 1997. **Mastering Data Mining.** New York : John Wiley and Sons

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

ประวัติผู้เขียน

ชื่อ นามสกุล	นางวาสนา เทียมประดิษฐ์
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต สาขาสถิติ
สถานที่สำเร็จการศึกษา	มหาวิทยาลัยเกษตรศาสตร์
ปีการศึกษาที่สำเร็จการศึกษา	ปีการศึกษา 2538

ภาคผนวก

ตัวอย่างข้อมูลที่ใช้ทดสอบก่อนการแปลงข้อมูล

M	3	30	1	2	R3	15	3	8.25	1,005,000	12	N	0	0	COT	CH	2	4	1	30,000	38	10	1
M	3	20	1	1	R3	15	5.1	13.25	650,000	12	N	0	0	COT	CH	1	4	1	12,000	40	12	1
F	1	40	3	1	R1	10	3.5	12.25	920,000	12	N	0	0	COT	CH	1	4	1	45,000	57	22	1
M	1	00	2	1	R3	15	8.5	12.25	850,000	12	N	0	0	COT	CH	1	4	1	32,000	50	17	1
F	2	30	4	2	R1	25	12.6	14	1,200,000	11	N	0	0	COT	CH	2	4	1	50,000	56	20	1
M	3	20	1	2	R1	15	2.5	8.25	90,000	12	N	0	0	COT	CH	2	4	1	8,500	27	4	2
F	3	10	1	2	R3	15	11	13.5	350,000	12	N	0	0	COT	CH	2	4	1	9,000	26	4	2
F	3	20	1	1	R1	10	9.2	12.75	700,000	12	N	0	0	COT	CH	2	4	1	15,000	54	16	2
F	2	10	1	2	R1	5	2.6	7.25	450,000	12	N	0	0	COT	CH	2	4	1	22,000	35	12	2
F	3	20	1	1	R1	9	8.4	12	380,000	12	N	0	0	ATM	CH	1	4	1	32464	50	24	1

ตัวอย่างข้อมูลที่ใช้ทดสอบหลังจากทำการแปลงข้อมูล

2,3,30,1,2,R3,2,1,1,4,2,N,0,0,1,1,2,4,1,2,3,2,1

2,3,20,1,1,R3,2,2,3,3,2,N,0,0,1,1,1,4,1,1,1,2,1

1,1,40,3,1,R1,1,1,3,3,2,N,0,0,1,1,1,4,1,3,2,4,1

2,1,00,2,1,R3,2,3,3,3,2,N,0,0,1,1,1,4,1,3,2,3,1

1,2,30,4,2,R1,3,3,3,4,2,N,0,0,1,1,2,4,1,3,2,3,1

2,3,20,1,2,R1,2,1,1,1,2,N,0,0,1,1,2,4,1,1,2,1,2

1,3,10,1,2,R3,2,2,3,2,2,N,0,0,1,1,2,4,1,1,2,1,2

1,3,20,1,1,R1,1,2,3,3,2,N,0,0,1,1,2,4,1,1,1,3,2

1,2,10,1,2,R1,1,1,1,2,2,N,0,0,1,1,2,4,1,2,2,2,2

1,3,20,1,1,R1,1,3,3,2,2,N,0,0,3,1,1,4,1,3,2,4,1