การวิเคราะห์ความสัมพันธ์ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ โดยเทคนิค Data Mining

Data Mining for Non-Performing Loans

โดย

นายธันยวีร์ สถิรวรกุล

รหัส 45061736

อาจารย์ที่ปรึกษา

รศ. ดร.อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการศึกษากรณีพิเศษ
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 1 ปีการศึกษา 2547
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การวิเคราะห์ความสัมพันธ์ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ โดยเทคนิค Data Mining
นักศึกษา	นาย ธันยวีร์ สถิรวรกุล
อาจารย์ที่ปรึกษา	รศ.ดร. อาริต ธรรมโน
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	การจัดการเทคโนโลยีสารสนเทศ
ปีการศึกษา	2547

บทคัดย่อ

ในสภาพเศรษฐกิจของประเทศไทยในปัจจุบัน ปัญหาของหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) ได้กลายมาเป็นปัญหาที่สำคัญอย่างหนึ่งในการพัฒนาเศรษฐกิจของประเทศไทยในปัจจุบัน ดังนั้น แต่ละสถาบันการเงินจึงต้องพยายามหากลยุทธ์และวิธีการเพื่อลดปัญหาหนี้ที่ไม่ก่อให้เกิดรายได้ให้หมดไป ดังนั้น การนำเอาระบบสารสนเทศเข้ามาแก้ไขปัญหาในส่วนนี้ จึงมีความจำเป็นอย่างยิ่ง โดยเฉพาะความต้องการในการนำเอาเทคนิคที่สามารถค้นหาข้อมูลหรือความสัมพันธ์ที่เราต้องการ ซึ่งจะซ่อนอยู่ในฐานข้อมูลขององค์กรที่มีขนาดใหญ่ได้อย่างมีประสิทธิภาพ จึงได้มีการนำเอาเทคนิค Data Mining เข้ามาช่วยในการวิเคราะห์หารูปแบบความสัมพันธ์และสารสนเทศที่มีประโยชน์ที่ซ่อนอยู่ภายในฐานข้อมูลเหล่านั้นเพื่อนำไปประยุกต์เข้ากับการวางกลยุทธ์ทางด้านธุรกิจการเงินให้รวดเร็วและเกิดประสิทธิภาพสูงสุด ซึ่งโครงการนี้จะนำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่อพยากรณ์ผลลัพธ์ของการปรับปรุงโครงสร้างหนี้และวิเคราะห์ความสัมพันธ์ของปัจจัยแวดล้อมต่างๆที่มีผลต่อการปรับปรุงโครงสร้างหนี้ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ โดยใช้ Predictive Model เพื่อที่จะสามารถนำเอาสารสนเทศที่ได้รับมาวิเคราะห์ถึงแนวโน้มที่ลูกหนี้จะเลื่อนชั้นมาเป็นหนี้ปกติ อีกทั้งหาทางป้องกันและแก้ไขปัญหาของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ให้มีประสิทธิภาพสูงสุด

Title	Data Mining for Non-Performing Loans
Student	Mr. Thunyawee Sathiravorakul
Advisor	Assoc.Prof. Dr. Arit Thammano
Level of Study	Master of Science in Information Technology
Major	Information Technology Management
Academic Year	2004

ABSTRACT

In the present, the problem of NPLs has become a major problem of Thai economy. Thus, financial institutes have to determine the strategies to eliminate this problem. It is necessary to bring information technology to solve this problem, especially the technique which can efficiently seek the hidden information or relationship in the database of large organization. The Data Mining Technique, then is brought to analyze the relationship between these NPLs. Financial institutes also apply this technique to their strategies in order to get rapid response and to achieve the efficiency. This project will present the process and the method of work system development in order to predict the result of debt structure improvement. It also analyze the relationship of surrounding factors which effects debt structure improvement of NPLs by using the predictive model. This model can be used to analyze trends of the shift from NPLs debt to normal debt which can lead to the prevention and the solutions of NPLs problem.

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

บทที่ 1
บทนำ

ในบทนี้จะกล่าวถึงความเป็นมาของปัญหาในการนำเอาเทคนิค Data Mining เข้ามาช่วยในการค้นหาสารสนเทศที่เป็นประโยชน์ต่อองค์กร ซึ่งซ่อนอยู่ในฐานข้อมูลที่มีขนาดใหญ่, วัตถุประสงค์ของโครงการ, ขอบเขตของการศึกษา และขั้นตอนการดำเนินงานของโครงการ รวมไปถึงประโยชน์ที่จะได้รับจากการพัฒนาโครงการนี้

1.1 ความเป็นมาของปัญหา

ในสภาวะเศรษฐกิจปัจจุบัน ธุรกิจต่างๆ มีการแข่งขันกันสูงมาก ทำให้แต่ละองค์กรจำเป็นต้องระดมหาเงินทุนเพื่อมาลงทุนในธุรกิจของตนเอง โดยการกู้ยืมเงินทุนจากแหล่งเงินทุนต่างๆ แต่เมื่อธุรกิจไม่ประสบความสำเร็จ ผลการดำเนินงานไม่ได้เป็นไปตามเป้าหมาย จึงส่งผลกระทบให้ธุรกิจไม่สามารถหาเงินมาชำระหนี้ที่ได้กู้ยืมจากสถาบันการเงินและแหล่งเงินทุนอื่นๆได้ และนั่นก็กลายเป็นที่มาของหนี้ที่ไม่ก่อให้เกิดรายได้ หรือที่รู้จักคุ้นหูว่า NPLs (Non-Performing Loans) โดยปัญหาของหนี้ที่ไม่ก่อให้เกิดรายได้นี้จะมีผลกระทบโดยตรงกับสถาบันการเงินต่างๆที่ปล่อยเงินกู้ไป ทำให้ผลการดำเนินงานของสถาบันการเงินไม่ดีเท่าที่ควร จึงก่อให้เกิดเป็นปัญหาลูกโซ่กระทบไปถึงสภาพเศรษฐกิจของประเทศต่อไป

ปัญหาของหนี้ที่ไม่ก่อให้เกิดรายได้ จึงเป็นปัญหาที่แต่ละสถาบันการเงินต้องให้ความสำคัญในการแก้ไขอย่างมาก โดยสถาบันการเงินจะมีวิธีการแก้ไขโดยมีการเจรจาปรับปรุงโครงสร้างหนี้แก่ลูกหนี้กลุ่มนี้ เพื่อหาวิธีการหรือเงื่อนไขต่างๆที่เหมาะสมแก่ทั้งฝ่ายเจ้าหนี้และลูกหนี้มากที่สุดที่จะทำให้ลูกหนี้สามารถชำระหนี้ได้อย่างมีประสิทธิภาพ ดังนั้น ถ้าหากมีการนำเอาความรู้ทางเทคโนโลยีเข้ามาช่วยในการศึกษาและวิเคราะห์ข้อมูลในส่วนนี้ โดยสามารถพยากรณ์ผลการปรับปรุงโครงสร้างหนี้ของลูกหนี้ว่ามีโอกาสสำเร็จหรือไม่ อีกทั้งจะทำให้ทราบถึงปัจจัยที่มีผลกระทบต่อผลของการปรับโครงสร้างหนี้ว่าปัจจัยใดที่มีผลกระทบต่อลูกหนี้ในแต่ละกลุ่มที่แตกต่างกันไป เพื่อหาทางปรับปรุงแก้ไขปัญหาได้อย่างมีประสิทธิภาพให้แก่ลูกหนี้ได้ ซึ่งเป็นที่มาของโครงการฉบับนี้ โดยจะนำข้อมูลการปรับปรุงโครงสร้างหนี้ของสถาบันการเงินมาใช้ในการวิเคราะห์โดยใช้เทคนิค Data Mining เข้ามาช่วยในการประมวลผลข้อมูล เพื่อพยากรณ์ผลการปรับปรุงโครงสร้างหนี้และศึกษาถึงปัจจัยต่างๆที่มีผลกระทบกับผลการปรับโครงสร้างหนี้ เพื่อประยุกต์

ใช้กับแนวทางในการแก้ไขปัญหาของหนี้ที่ไม่ก่อให้เกิดรายได้ของสถาบันการเงินต่างๆให้มีประสิทธิภาพมากที่สุด

1.2 วัตถุประสงค์

การนำเอาเทคนิคของ Data Mining มาใช้ในการพยากรณ์ผลลัพธ์ของข้อมูลในการปรับปรุงโครงสร้างหนี้ วัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้นี้มาใช้ประกอบในการวางแผนกลยุทธ์ในการป้องกันและแก้ไขหนี้ที่ไม่ก่อให้เกิดรายได้อย่างมีประสิทธิภาพ ซึ่งจะนำไปใช้เป็นแนวทางในการประกอบการตัดสินใจในการดำเนินงานของผู้บริหาร อีกทั้งยังสามารถทราบถึงปัจจัยแวดล้อมที่มีผลกระทบต่อการเปลี่ยนชั้นหนี้ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้

1.3 ขอบเขตการศึกษา

โครงการนี้เป็นการศึกษาเทคนิค Data Mining เพื่อมาประยุกต์ใช้ โดยอาศัยหลักการของ Predictive Model ในการพยากรณ์ผลลัพธ์ของข้อมูลในฐานข้อมูลการปรับโครงสร้างหนี้ของธนาคารพาณิชย์ โดยจะใช้ข้อมูลสะสม ตั้งแต่ มกราคม ปี พ.ศ. 2543 จนถึงสิ้นสุด ณ ธันวาคม ปี พ.ศ. 2546 มาเป็นฐานข้อมูลในการศึกษา ซึ่งจะนำ CHAID Algorithm มาใช้ในการวิเคราะห์หาความสัมพันธ์ของกลุ่มข้อมูล

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1.) ศึกษาแนวคิดและทฤษฎีเบื้องต้นที่เกี่ยวข้องกับเทคนิค Data Mining เพื่อนำมาประยุกต์ใช้ในการพัฒนาโครงการ

2.) ศึกษาทฤษฎี Predictive Model และ Algorithm ต่างๆที่เกี่ยวข้องเพื่อนำมาประยุกต์ใช้ในการศึกษาโครงการ

3.) เก็บรวบรวมข้อมูลที่เกี่ยวข้องในการทำโครงการ

4.) วิเคราะห์และประมวลผลข้อมูลเพื่อพยากรณ์ผลลัพธ์ข้อมูลของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้

5.) สรุปผลการศึกษาข้อมูลของโครงการ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการที่ได้ศึกษาแนวคิดและทฤษฎีของ Data Mining เพื่อนำมาพยากรณ์ผลลัพธ์ของการปรับปรุงโครงสร้างหนี้ของข้อมูลของหนี้ที่ไม่ก่อให้เกิดรายได้ คาดว่าจะได้ให้ประโยชน์ ดังนี้

1.) เพื่อให้เข้าใจถึงแนวคิด ขั้นตอนกระบวนการทำงานและการประมวลผลของเทคนิค Data Mining ที่นำมาใช้ เพื่อสามารถนำไปประยุกต์ใช้กับข้อมูลอื่นๆได้อย่างเหมาะสม

2.) เพื่อให้สามารถพยากรณ์ผลลัพธ์ของการเลื่อนชั้นของลูกหนี้จากหนี้ที่ไม่ก่อให้เกิดรายได้ให้กลับมาเป็นหนี้ปกติได้อย่างถูกต้องแม่นยำ เพื่อนำไปประยุกต์ใช้ประกอบกับการวางกลยุทธ์ในการดำเนินงานอย่างมีประสิทธิภาพสูงสุด

3.) เพื่อให้ทราบถึงปัจจัยที่มีผลกระทบต่อการเลื่อนชั้นของลูกหนี้แต่ละกลุ่มในการปรับปรุงโครงสร้างหนี้ ทำให้สามารถหาแนวทางแก้ไขและป้องกันปัญหาของหนี้ที่ไม่ก่อให้เกิดรายได้ให้หมดไป

บทที่ 2

ทฤษฎีและหลักการของ Data Mining

2.1 ความหมายของ Data Mining

ในอดีต การจะค้นหาข้อมูลที่มีประโยชน์จากฐานข้อมูลนั้นเป็นเรื่องที่ยาก ยิ่งถ้าหากเป็นฐานข้อมูลที่มีขนาดใหญ่มากๆ ก็จะยิ่งต้องใช้เวลาในการค้นหานานมาก จึงทำให้นักพัฒนาระบบต่างคิดค้นวิธีการที่จะทำให้สามารถค้นหาข้อมูลสารสนเทศที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ตลอดจนความสัมพันธ์กันของปัจจัยต่างๆ เพื่อนำมาใช้ประโยชน์ในการวิเคราะห์ การพยากรณ์ ที่แม่นยำถูกต้อง ซึ่งสามารถใช้ประโยชน์ในการกำหนดแนวทางหรือแผนในการปฏิบัติงานขององค์กรนั้นให้มีประสิทธิภาพมากที่สุด

ดังนั้น การที่เราจะค้นหาข้อมูลที่เป็นสารสนเทศที่เราต้องการจากแหล่งข้อมูลดิบที่มีมากมายมหาศาลนั้น เราจำเป็นต้องมีเครื่องมือที่จะช่วยในการค้นหาสารสนเทศเหล่านั้น ซึ่งหนึ่งในนั้นก็คือ เทคนิคของ Data Mining

โดยนิยามของ Data Mining นั้น ก็หมายถึง กระบวนการในการค้นหาเอาข้อมูลสารสนเทศที่ซ่อนอยู่ภายใต้ฐานข้อมูลที่มีอยู่จำนวนมากมาย ซึ่งเก็บอยู่ในระบบฐานข้อมูลขององค์กรออกมาโดยใช้กระบวนการต่างๆในการค้นหาข้อมูลออกมาจากฐานข้อมูล แล้วนำมาตั้งเป็นสมมติฐาน หลังจากนั้นก็นำข้อมูลที่ต้องการทราบมาทำการทดสอบสมมติฐานที่สร้างไว้ในภายหลัง ซึ่งสารสนเทศที่ได้ออกมาต้องมีลักษณะคือ

- เป็นข้อมูลที่ไม่เคยรู้ล่วงหน้ามาก่อน (Unknown) หมายถึง ข้อมูลสารสนเทศที่ได้รับนั้นต้องไม่เคยค้นพบมาก่อนหน้า และไม่สามารถคาดเดาได้ว่าผลที่ได้รับจะออกมาในลักษณะใด
- ต้องเป็นข้อมูลที่มีความถูกต้อง (Valid) หมายถึง สารสนเทศที่ได้รับต้องเป็นสารสนเทศที่มีความถูกต้อง เนื่องจากต้องนำไปใช้ประกอบกับข้อมูลส่วนอื่นๆ ดังนั้นต้องมีความถูกต้อง น่าเชื่อถือ
- สามารถนำไปใช้ประโยชน์ได้ (Actionable) คือ ต้องสามารถนำเอาข้อมูลที่ค้นพบออกมาไปใช้ประโยชน์ด้านอื่นๆได้ เช่น นำมาช่วยตัดสินใจในการวางแผนการตลาด เพื่อสร้างความได้เปรียบทางการแข่งขันในธุรกิจ เป็นต้น

ซึ่งเปรียบเสมือนการขุดหาแร่จากเหมืองแร่ที่มีขนาดใหญ่ กว่าที่จะได้แร่ที่มีค่าอย่างที่ต้องการนั้นต้องผ่านกระบวนการมากมายหลายขั้นตอนในการขุดค้น กลั่นกรอง เพื่อที่จะได้แร่ที่มีค่าออกมา นั่นจึงเป็นที่มาของคำว่า Data Mining

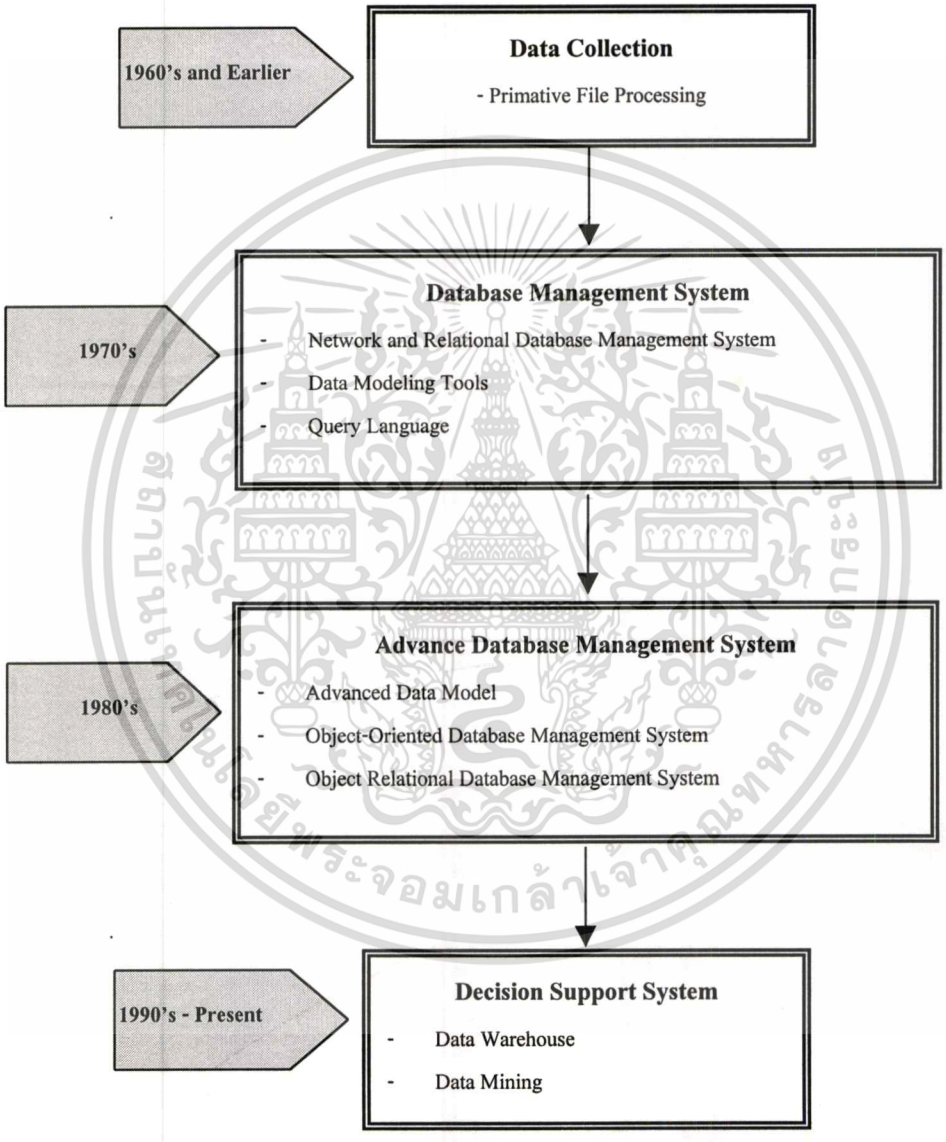
2.2 วิวัฒนาการของ Data Mining

วิวัฒนาการของเทคโนโลยีด้านฐานข้อมูลนั้น ได้มีการพัฒนามาทุกยุคทุกสมัยตั้งแต่ในอดีตจนถึงปัจจุบัน ซึ่งเป็นเทคโนโลยีที่มีความสำคัญมาก เนื่องจากข้อมูลเป็นสิ่งที่สำคัญในการนำมาใช้ประโยชน์ในด้านต่างๆ อีกทั้งแนวโน้มของการเพิ่มขึ้นของข้อมูลก็จะมีแนวโน้มที่สูงขึ้นมาก จึงได้มีการพัฒนาและปรับปรุงวิธีการต่างๆเพื่อที่จะสามารถเก็บรวบรวมและประมวลผลข้อมูลที่มีอยู่อย่างมหาศาลได้อย่างมีประสิทธิภาพ โดยสามารถสรุปวิวัฒนาการของการพัฒนาเทคโนโลยีด้านฐานข้อมูลได้เป็นช่วงเวลา ดังนี้

- ช่วงปี ค.ศ. 1960 ได้มีการเริ่มพัฒนาเทคโนโลยีฐานข้อมูลมาจากระบบ File Processing พื้นฐาน โดยจะพัฒนาเป็นระบบการเก็บข้อมูล, การสร้างฐานข้อมูล (Database), ระบบ IMS และระบบเครือข่าย DBMS
- ช่วงปี ค.ศ. 1970 มีการพัฒนาการการเก็บข้อมูลในรูปแบบของตาราง (Relational Database System) โดยมีการสร้างเครื่องมือต่างๆ ที่ช่วยอำนวยความสะดวกในการจัดการกับข้อมูล อีกทั้งยังมีการคิดค้นภาษาที่ใช้ในการเรียกข้อมูล (Query Language) เพื่ออำนวยความสะดวกในการเข้าถึงข้อมูลในฐานข้อมูล
- ช่วงปี ค.ศ. 1980 เริ่มมีการพัฒนาระบบจัดการให้มีประสิทธิภาพมากขึ้นกว่าในอดีต เพื่อให้จัดเก็บข้อมูลที่มีจำนวนมากและมีความซับซ้อนได้ จึงได้เกิดระบบต่างๆ มากมาย เช่น Object-Oriented Database Management System, Object Relational Database Management System เป็นต้น
- ช่วงปี ค.ศ. 1990 จนมาถึงยุคปัจจุบัน ได้พัฒนาให้สามารถจัดเก็บข้อมูลได้ในหลายรูปแบบ แตกต่างกันทั้งระบบปฏิบัติการ หรือการจัดเก็บฐานข้อมูล ซึ่งเป็นการนำเอาข้อมูลทั้งหมดมารวบรวมและจัดเก็บไว้ในรูปแบบเดียวกัน มีชื่อเรียกว่า Data Warehouse เพื่อเพิ่มความสะดวกในการบริหารจัดการ โดยที่เทคโนโลยี Data Warehouse จะรวมไปถึงการทำ Data Cleansing, Data Integration และ On-Line Analytical Processing (OLAP) ซึ่งเป็นเทคนิคในการวิเคราะห์ข้อมูลในหลายมิตินั้น ได้เกิดขึ้นมาตามลำดับ และเนื่องจากการเพิ่มขึ้นของข้อมูลที่สูงขึ้นอย่างรวดเร็ว ทำให้ฐานข้อมูลมีขนาดที่ใหญ่มาก ซึ่งเกินกว่าที่จะใช้กำลังคนในการบริหารจัดการได้ เป็นผลทำให้มีความจำเป็นที่จะต้องมีการคิดค้นเครื่องมือที่จะมาช่วยในการวิเคราะห์ข้อมูลที่มีจำนวนมากได้อย่าง

รวดเร็วและมีประสิทธิภาพ เพื่อให้ได้ข้อมูลที่เป็นสารสนเทศที่มีประโยชน์ต่อองค์กรออกมา นั่นก็คือ Data Mining

จากที่ได้กล่าวถึงประวัติความเป็นมาและวิวัฒนาการของเทคโนโลยีฐานข้อมูลตั้งแต่ในยุคอดีตจนถึงยุคปัจจุบัน จะสามารถแสดงได้ดังรูปที่ 2.1



รูปที่ 2.1 แสดงวิวัฒนาการของเทคโนโลยีฐานข้อมูล

2.3 ลักษณะของข้อมูลที่จะนำมาทำ Data Mining

ในการทำ Data Mining จะเป็นการนำข้อมูลที่มีนั้นมาใช้งานให้เกิดประโยชน์สูงที่สุด เนื่องจากข้อมูลที่ถูกเก็บไว้ในฐานข้อมูล หากเก็บไว้เฉยๆ จะไม่ก่อให้เกิดประโยชน์ ดังนั้นจึงต้องมีการ

สกัดเอาสารสนเทศไปใช้ให้เกิดประโยชน์สูงสุด ยิ่งในปัจจุบันการวิเคราะห์ข้อมูลจากฐานข้อมูลเดียวอาจให้ความรู้ที่ไม่เพียงพอและลึกซึ้งสำหรับการดำเนินงานภายใต้ภาวะที่มีการแข่งขันสูงและมีการเปลี่ยนแปลงที่รวดเร็วจึงจำเป็นที่จะต้องรวบรวมฐานข้อมูลหลาย ๆ ฐานข้อมูลเข้าด้วยกันเรียกว่า "คลังข้อมูล" (Data Warehouse) ดังนั้นลักษณะของข้อมูลที่จะนำมาใช้ในการทำ Data Mining ก็มีความสำคัญมากเช่นกัน โดยสามารถแยกเป็นประเภทของข้อมูลได้ดังนี้

- Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย Entity-Relationship Model (ER-Model)
- Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ ๆ เดียวกัน
- Transactional Database ประกอบด้วยข้อมูลที่แต่ละ Transaction แทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูปแบบของชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น
- Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่น ๆ เช่น ข้อมูลที่เป็น Object-Oriented, ข้อมูลที่เป็น Text File, ข้อมูลที่เป็น Multimedia, ข้อมูลที่อยู่ในรูปของ web เป็นต้น

นอกจากนี้แล้วข้อมูลที่จะนำมาทำ Data Mining ยังต้องมีลักษณะเฉพาะของข้อมูลที่สามารถจะนำมาทำ Data Mining ดังนี้

- ข้อมูลที่มีขนาดใหญ่เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่า หรือโดยการใช้ Database Management System (DBMS) ในการจัดการฐานข้อมูล
- ข้อมูลที่มาจากหลายแหล่ง โดยอาจจะรวบรวมมาจากหลายระบบปฏิบัติการหรือหลาย DBMS เช่น Oracle , DB2 , MS SQL , MS Access เป็นต้น
- ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการ Mining หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำ Mining แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำ Mining สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลา จึงต้องทำ Mining ใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม
- ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ, ข้อมูลมัลติมีเดีย เป็นต้น ข้อมูลเหล่านี้สามารถนำมาทำ Mining ได้เช่นกันแต่ต้องใช้เทคนิคการทำ Data Mining ขั้นสูง

2.4 ขั้นตอนการทำงานของ Data Mining

การใช้เทคนิค Data Mining นั้น จะประกอบด้วยขั้นตอนในการทำงานโดยสังเขป ซึ่งสามารถแบ่งได้เป็น 5 ขั้นตอน ดังนี้

1. การกำหนดวัตถุประสงค์ของธุรกิจในการทำ Data Mining (Business Objective Determination)

เป็นการกำหนดขอบเขตแนวทางของการทำ Data Mining ให้ชัดเจน เพื่อให้ตอบสนองต่อการแก้ไขปัญหาของธุรกิจ (Business Problem) ได้อย่างเหมาะสม ซึ่งจะต้องทำการวิเคราะห์ข้อมูลทางธุรกิจเพื่อให้ทราบว่า วัตถุประสงค์ของการทำ Data Mining ครั้งนี้คืออะไร เพื่อจะได้กำหนดสมมติฐานและแบบจำลองในการทดลองได้อย่างเหมาะสม

2. การเตรียมข้อมูลในการทำ Data Mining (Data Preparation)

เป็นขั้นตอนในการทำข้อมูลดิบที่เราได้รับมา ซึ่งจะอยู่ในรูปแบบที่หลากหลายแตกต่างกันไปให้อยู่ในรูปที่พร้อมจะใช้งาน เพื่อให้ผลลัพธ์ที่ได้จากการทำ Data Mining มีความถูกต้องแม่นยำมากยิ่งขึ้น ซึ่งขั้นตอนนี้จะเป็นขั้นตอนที่ต้องใช้เวลามากที่สุด เนื่องจากปริมาณข้อมูลที่มีจำนวนมาก และข้อมูลที่ได้รับส่วนใหญ่จะมาจากหลากหลายแหล่ง รูปแบบของข้อมูลจึงมีความแตกต่างกันมาก เราจึงจำเป็นที่จะต้องใช้เวลาในการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกัน เพื่อให้พร้อมที่จะใช้งานก่อน โดยขั้นตอนการเตรียมข้อมูลนี้ จะแบ่งออกเป็น 3 ขั้นตอน ได้ดังนี้

- การคัดเลือกข้อมูล (Data Selection) เป็นการคัดเลือกเฉพาะข้อมูลที่เราต้องการใช้และนำข้อมูลที่ไม่ต้องการออกไป เพื่อเป็นการจัดเตรียมเฉพาะข้อมูลที่จำเป็นในการใช้วิเคราะห์จริงๆเก็บไว้เท่านั้น เพื่อให้เป็นไปตามวัตถุประสงค์ที่ได้กำหนดไว้ในตอนต้น นอกจากนี้ยังเป็นการลดขนาดของข้อมูลไม่ให้ใหญ่เกินไป เพื่อความสะดวกรวดเร็วในการทำ Data Mining
- การกรองข้อมูล (Data Preprocessing) เป็นการเตรียมข้อมูลที่ได้จากการคัดเลือกเบื้องต้นให้อยู่ในรูปที่สมบูรณ์ เนื่องจากบางครั้งข้อมูลที่ได้รับมาอาจจะยังไม่ถูกต้องสมบูรณ์ ซึ่งสามารถพิจารณาได้เป็นหลายลักษณะ ดังนี้

  - ข้อมูลที่ขาดหายไป (Missing Value) หมายถึง ข้อมูลที่มีค่าของข้อมูลไม่ครบถ้วนในข้อมูลชุดนั้น ซึ่งอาจจะมีสาเหตุมาจากความผิดพลาดในการป้อนข้อมูลของตัวบุคคล เช่น การกรอกข้อมูลไม่ครบ ทำให้ข้อมูลขาดความสมบูรณ์ในการนำมาใช้ ดังนั้นควรมีการแก้ไขข้อมูล โดยอาจจะตัดข้อมูลดังกล่าวออกไปเลย หรือกำหนดค่าเข้าไปจากการใช้ค่าเฉลี่ยของข้อมูลในชุดนั้น หรืออาจกำหนดค่าใหม่ที่แสดงไว้แต่จะไม่มีผลในการนำมาวิเคราะห์ เป็นต้น

- ข้อมูลที่มีความผิดพลาด (Noisy Data) หมายถึง ข้อมูลที่มีค่าของข้อมูลผิดเพี้ยนไปจากข้อมูลในชุดนั้น มีค่าผิดปกติ แตกต่างจากข้อมูลในกลุ่มนั้น โดยอาจจะเกิดจากความผิดพลาดของมนุษย์เอง หรือว่าโปรแกรมที่นำมาใช้งานมีความผิดพลาด ซึ่งวิธีในการแก้ไขคือ ตรวจสอบดูว่าข้อมูลตัวใดเป็น Noisy Data แล้วตัดทิ้งออกไป เพื่อไม่ให้ข้อมูลดังกล่าวมีผลกระทบต่อผลการทำ Data Mining ทำให้ผลลัพธ์ที่ได้ผิดพลาด ไม่น่าเชื่อถือ แต่ถ้าข้อมูลที่ขาดไปมีจำนวนมาก การตัดทิ้งอาจมีผลกระทบได้ ดังนั้นเราอาจใช้วิธีบันทึกส่วนที่หายไปด้วยค่าเฉลี่ย (สำหรับข้อมูลที่เป็น Categorical อาจบันทึกด้วยค่าฐานนิยม หรือบันทึกเป็น Unknown แทน)

- การแปลงข้อมูล (Data Transformation) เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้ในการวิเคราะห์ตามลักษณะของ Algorithm แต่ละแบบที่จะนำมาใช้ในการวิเคราะห์ เพื่อให้การอ่านผลวิเคราะห์สามารถเข้าใจได้ง่าย ไม่ซับซ้อน เช่น ถ้า Algorithm ที่เราจะนำมาใช้ในการประมวลผลนั้น ต้องใช้ข้อมูลที่เป็นตัวเลข (Numeric) เราก็ต้องแปลงข้อมูลที่เป็นข้อความ (Categorical) ให้อยู่ในรูปแบบตัวเลข โดยอาจจะกำหนดให้เลข 1, 2 แทนข้อความที่ 1, 2 ตามลำดับ เป็นต้น

3. การทำ Data Mining (Data Mining)

จะเป็นขั้นตอนของการนำเอาข้อมูลที่เราได้เตรียมไว้ในรูปแบบที่ถูกต้องแล้วมาผ่าน Algorithm ที่เหมาะสมที่เราได้เลือกไว้ในการทำ Data Mining เพื่อประมวลผลออกมา โดย Data Mining จะใช้วิธีการทำซ้ำๆจนกว่าจะสามารถหาความสัมพันธ์หรือรูปแบบที่เราต้องการออกมา ซึ่งก็จะได้ผลลัพธ์ที่ถูกต้องตรงตามวัตถุประสงค์ที่เราได้กำหนดเอาไว้ในตอนต้น

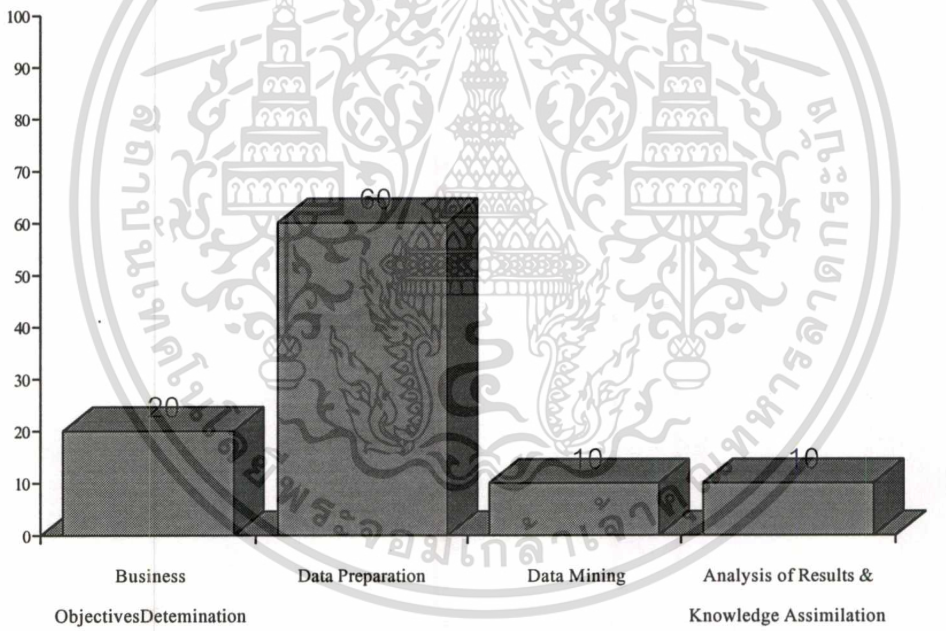
4. การวิเคราะห์ผลลัพธ์ที่ได้ (Results Analysis)

เป็นการนำเอาผลลัพธ์ที่ได้จากการประมวลผลผ่าน Algorithm มาวิเคราะห์ แปลความหมายว่าผลลัพธ์ที่ได้มีลักษณะเป็นอย่างไร มีความถูกต้องมากน้อยแค่ไหน และสามารถนำมาใช้ได้ตามวัตถุประสงค์ที่ต้องการได้หรือไม่ ซึ่งการทำงานในขั้นตอนนี้จำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลและการวิเคราะห์ทางธุรกิจ รวมทั้งพื้นฐานความเข้าใจในข้อมูลที่นำมาใช้ประมวลผล เนื่องจาก การทำ Data Mining นั้นจะได้ค่าออกมาเป็นตัวเลข หรือเป็นช่วงตัวเลข เพราะข้อมูลที่นำมาใช้ประมวลผลนั้น ส่วนมากจะอยู่ในรูปของตัวเลข ดังนั้นการนำผลจากการทำ Data Mining มาใช้ประโยชน์นั้น ต้องอาศัยความรู้ ความเข้าใจด้านข้อมูลและธุรกิจเป็นหลักในการแปลความผลลัพธ์ให้ถูกต้อง

5. การนำเอาข้อมูลสารสนเทศที่ได้มาประยุกต์ใช้กับธุรกิจ (Knowledge Assimilation)

จะเป็นการรวบรวมเอาผลวิเคราะห์ที่ได้ทั้งหมดจากขั้นตอนที่แล้วมาใช้ให้เกิดประโยชน์สำหรับธุรกิจนั้นๆ ซึ่งอาจจะต้องนำเอาความรู้ในด้านอื่นๆเข้ามาประยุกต์ใช้ควบคู่กันไปด้วย และอาจจะทำให้เกิดข้อมูลสารสนเทศใหม่ๆขึ้นอีกด้วย ดังนั้นในขั้นตอนนี้จะเป็นการนำเสนอแนวทางในการนำเอารูปแบบของผลลัพธ์ที่ได้จากการวิเคราะห์ไปใช้ให้เกิดประโยชน์สูงสุด และเป็นการนำเสนอแนวความคิดทางธุรกิจที่ได้ค้นพบใหม่อีกด้วย

ดังนั้น จากการศึกษาขั้นตอนต่างๆ ในการทำ Data Mining ข้างต้น ทำให้เราทราบถึงลักษณะการทำงานในแต่ละขั้นตอน ซึ่งทำให้เราสามารถแบ่งช่วงเวลาในการทำงานออกมา โดยแสดงออกมาในรูปแผนภูมิที่ 2.2



รูปที่ 2.2 แสดงอัตราส่วนการทำงานในแต่ละขั้นตอนของการทำ Data Mining

2.5 เทคนิคของ Data Mining

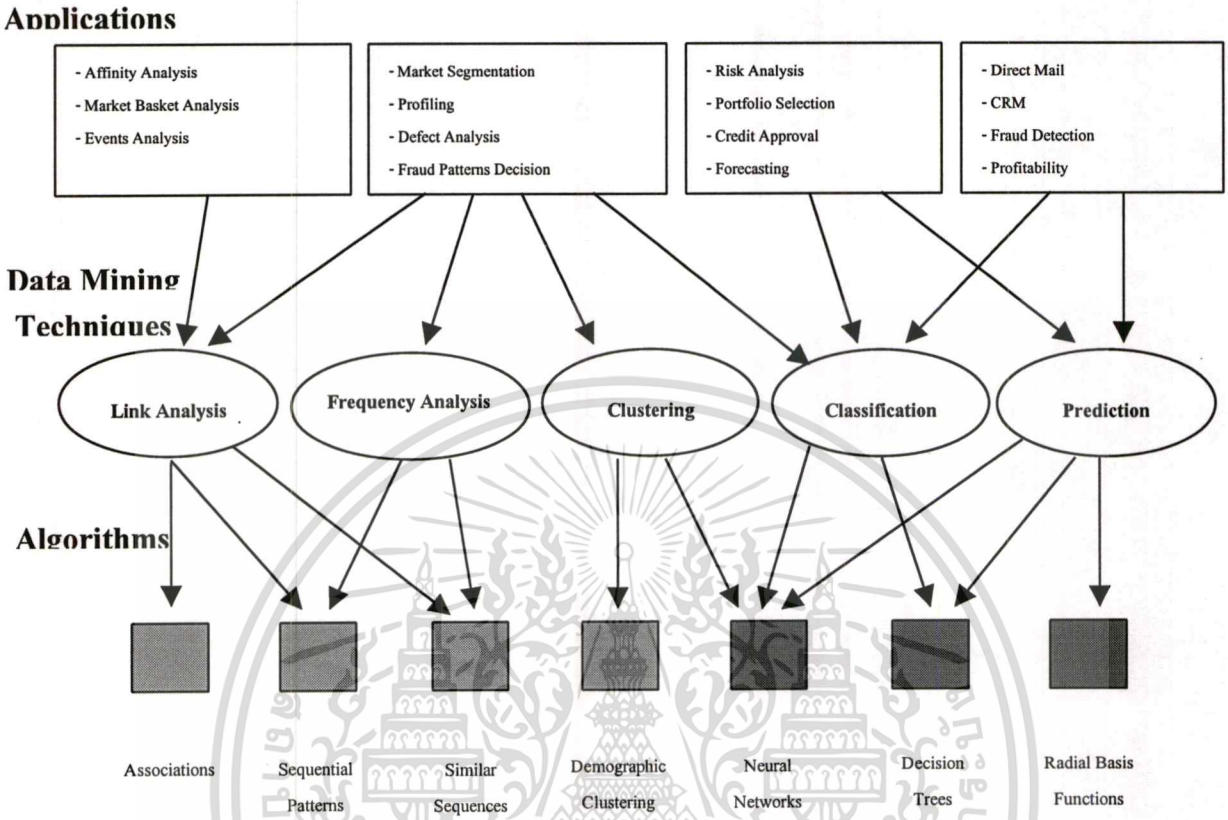
เทคนิคของการทำ Data Mining ที่เรานิยมใช้กันนั้น สามารถแบ่งออกกว้างๆ ได้เป็น 2 ลักษณะ คือ

- Discovery Data Mining เป็นการค้นหารูปแบบของข้อมูล โดยไม่จำเป็นต้องทราบรูปแบบของข้อมูลล่วงหน้า ซึ่งสามารถนำมาประยุกต์ใช้ได้กับหลายเทคนิค ยกตัวอย่างเช่น

- Data Clustering หรือ Data Segmentation เป็นการแยกกลุ่มของข้อมูลในฐานข้อมูลตามลักษณะคุณสมบัติที่เหมือนกันให้อยู่ในกลุ่มเดียวกัน โดยไม่ทราบล่วงหน้าว่าจะแบ่งข้อมูลออกเป็นจำนวนกี่กลุ่ม
- Link Analysis เป็นการค้นหาความสัมพันธ์ระหว่าง Record หรือกลุ่มของ Record ในฐานข้อมูลเดียวกัน ว่ามีความสัมพันธ์กันในลักษณะใด
- Frequency Analysis เป็นเทคนิคที่ใช้เฉพาะในการวิเคราะห์กลุ่มของข้อมูลที่มีการเรียงลำดับกันได้

- Predictive Data Mining เป็นกระบวนการสร้างโมเดลการทำนายผลของตัวแปรที่เราต้องการทราบในฐานข้อมูล โดยใช้การสังเกตจากรูปแบบความสัมพันธ์ของข้อมูลในอดีต ซึ่งจะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้แล้วจึงนำไปทดสอบโมเดลที่สร้างขึ้น ซึ่งตัวอย่างของเทคนิคที่ใช้ ได้แก่
  - Classification เป็นกระบวนการจัดหมวดหมู่ของข้อมูล โดยการนำเอาข้อมูลที่กำหนดไว้ล่วงหน้ามาใส่เข้าไปในโมเดลที่ได้เตรียมไว้เพื่อคาดการณ์ว่าผลลัพธ์ที่ออกมามีค่าเป็นอย่างไรในตัวแปรที่เป็นเป้าหมาย ซึ่งวิธีที่นิยมใช้กันมากก็คือ Tree Induction และ Neural Induction
  - Value Prediction เป็นการพยากรณ์ค่าของตัวแปรที่มีความต่อเนื่องจากตัวแปรอื่นในฐานข้อมูล ซึ่งจะใช้ทำนายค่าที่เป็นตัวเลข เช่น การทำนายราคาหุ้น เป็นต้น โดยมีวิธีที่น่าสนใจในการพยากรณ์คือ Linear Regression และ Nonlinear Regression

ซึ่งจากการแบ่งหมวดหมู่ลักษณะการทำงานของแต่ละเทคนิค จะสามารถยกตัวอย่างงานที่ที่เหมาะสมกับเทคนิคแต่ละประเภทได้ดังรูปที่ 2.3



รูปที่ 2.3 แสดง Applications และ Algorithm ของ Data Mining

สำหรับโครงการนี้ จะนำเสนอกระบวนการของการพยากรณ์ (Predictive Model) แบบ Classification ในการพยากรณ์ผลการปรับปรุงโครงสร้างหนี้ของลูกหนี้ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ ซึ่งจะใช้เทคนิคของ Tree Decision ในการวิเคราะห์ เนื่องจาก เทคนิค Tree Decision จะสามารถเรียนรู้ถึงรูปแบบของข้อมูล โดยสามารถแสดงให้เห็นภาพที่ชัดเจนออกมาในรูปของแผนภูมิต้นไม้ที่จะค่อยๆแตกออกมาเป็นกิ่งก้าน ทำให้สามารถแปลความหมายของความสัมพันธ์ของข้อมูลให้เข้าใจได้ง่าย การนำไปใช้เพื่อให้บรรลุวัตถุประสงค์ของธุรกิจก็จะทำได้อย่างมีประสิทธิภาพมีความถูกต้องมากขึ้น

บทที่ 3

เทคนิคการพยากรณ์ (Predictive Model)

3.1 การสร้างแบบจำลองพยากรณ์

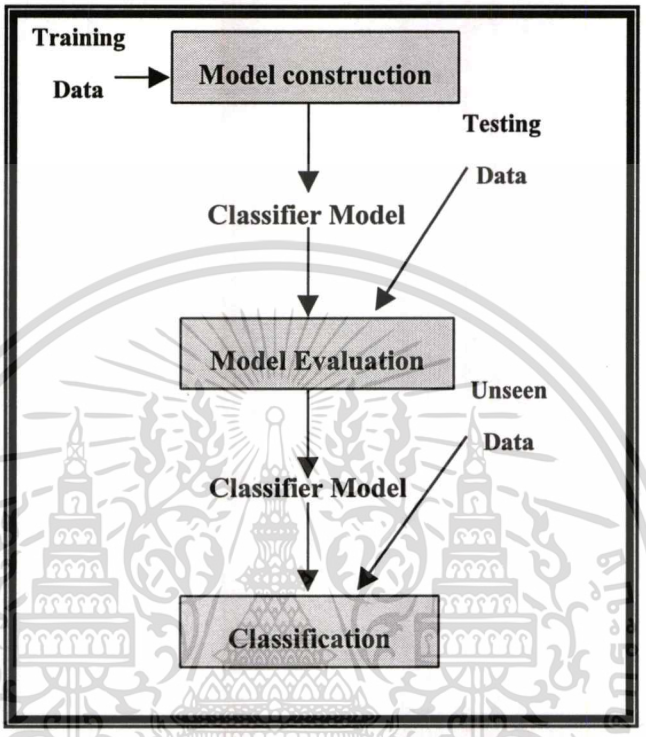
จากบทความข้างต้นที่พูดถึงเทคนิคการทำ Data Mining ที่มีหลากหลายรูปแบบ โดยในบทนี้เราจะมาเน้นในเรื่องของการพยากรณ์ผลลัพธ์โดยใช้ Data Mining

Predictive Data Mining หมายถึง การค้นหารูปแบบความสัมพันธ์ของฐานข้อมูลที่มีอยู่เป็นจำนวนมาก เพื่อนำมาใช้ในการพยากรณ์ผลลัพธ์ หรือนำมาใช้ในการคำนวณ เพื่อช่วยในการตัดสินใจในอนาคตได้อย่างถูกต้องและแม่นยำ ซึ่งเทคนิคในการ Predictive จะเป็นการศึกษา Decision Criteria ของข้อมูลที่เก็บไว้ในอดีต เหมือนกับการเรียนรู้จากประสบการณ์ของมนุษย์ที่จะจดจำเหตุการณ์ที่เคยเกิดขึ้นว่ามีลักษณะแบบใด เพื่อตัดสินใจ ถ้าหากเกิดเหตุการณ์ในลักษณะนี้ขึ้นอีกก็จะสามารถตัดสินใจได้รวดเร็วขึ้น โดยเทคนิคนี้ จะทำนายถึงความเป็นไปได้ ซึ่งจะใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ คือเราจะใช้เทคนิคนี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อตัดสินใจเลือกลักษณะข้อมูลที่ต้องการ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้ แล้วจึงนำไปทำนายผลลัพธ์ในกลุ่มข้อมูลที่ต้องการทราบ ซึ่งวิธีนี้ เรียกว่า Supervised Learning ดังนั้นข้อมูลที่มีอยู่ต้องสมบูรณ์ จึงจะทำให้ผลลัพธ์ออกมาถูกต้อง เพราะเราต้องนำข้อมูลในอดีตมาสร้างแบบจำลอง โดยในการทำงานจะแบ่งออกเป็น 2 ขั้นตอน คือ

- Training Phase คือ ขั้นตอนการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ความสัมพันธ์ของข้อมูลในอดีตที่เก็บไว้ในฐานข้อมูล ซึ่งจะใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมดในการสร้างแบบจำลอง
- Testing Phase คือ ขั้นตอนที่ใช้ทำการทดสอบแบบจำลองที่สร้างขึ้นมาจาก Training Phase ว่ามีความถูกต้องและมีความน่าเชื่อถือมากน้อยเพียงใด โดยจะนำข้อมูลส่วนที่เหลืออีก 20% จากการแบ่งไว้มาใช้ทดสอบแบบจำลองที่สร้างขึ้น

โดย Predictive Model ที่นำมาใช้ในโครงการนี้ คือ เทคนิคการทำ Classification จะเป็นกระบวนการสร้างโมเดลที่จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก, ดี, ปานกลาง และไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือการแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ เป็นต้น โดยพิจารณาจากข้อมูลที่มีอยู่ ซึ่งวิธีที่นิยมใช้กันมากก็คือ Tree

Induction และ Neural Induction โดยกระบวนการ Classification นี้แบ่งขั้นตอนการทำงานเป็น 3 ขั้นตอน ดังรูปที่ 3.1



รูปที่ 3.1 แสดงกระบวนการทำ Classification

3.2 โครงสร้างแบบต้นไม้ (Decision Tree)

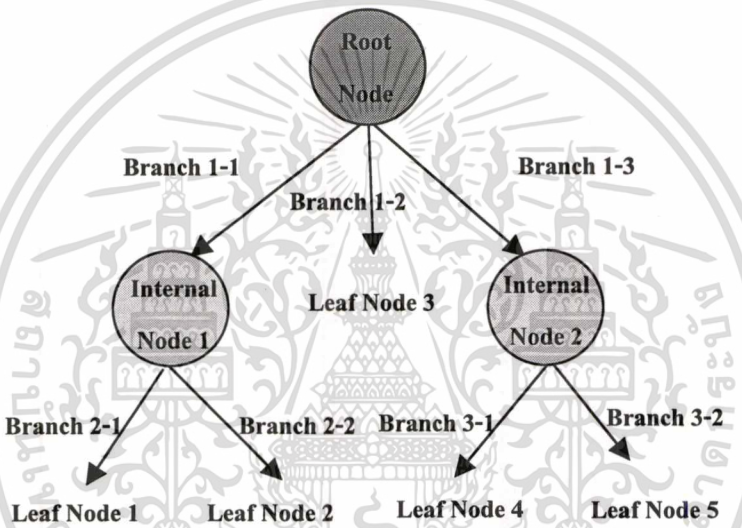
ในโครงการนี้ จะเลือกใช้เทคนิคการ Classification แบบวิธี Decision Tree ในการทำการวิเคราะห์ เนื่องจากผลลัพธ์ที่ได้จากการใช้วิธีนี้ จะสามารถตีความหมายของผลพยากรณ์ได้ง่าย และยังสามารถทำความเข้าใจกระบวนการใช้งานได้ง่าย อีกทั้งยังสามารถหาสาเหตุที่มาที่ไปของผลลัพธ์ได้ในรูปของ If-Then Rules โดยหลักการของ Tree Decision คือการแตกผลลัพธ์ของตัวแปรที่เรานำมาใช้ในการประมวลผลออกเป็นลำดับชั้น ลักษณะเหมือนแผนภูมิโครงสร้างขององค์กร โดยที่แต่ละ Node จะแสดงถึง Attribute ของข้อมูล แต่ละกิ่งแสดงถึงผลในการประมวลผล และ Leaf Node แสดงถึงผลลัพธ์ที่เราต้องการทราบ ซึ่งได้กำหนดไว้แล้ว

ซึ่งเทคนิคของ Tree Decision นั้น จะสามารถรองรับข้อมูลที่มีลักษณะของข้อมูลได้หลายลักษณะ ดังนี้

- Nominal เป็นลักษณะข้อมูลที่เป็นข้อมูลตัวเลข เช่น 1, 2, 3, 10, 20 เป็นต้น

- Ordinal เป็นข้อมูลที่มีลักษณะของข้อมูลที่สามารถแยกออกเป็นประเภทต่างๆได้ เช่น อากาศเย็น , อากาศร้อน, อากาศอบอุ่น เป็นต้น
- Interval เป็นข้อมูลที่มีลักษณะเป็นค่าต่อเนื่อง หรือค่าเฉลี่ย เช่น อุณหภูมิ เป็นต้น

โดยจะแสดงผลในรูปแผนภูมิที่จะแสดงการแตกผลลัพธ์ของตัวแปรแต่ละตัวออกมาเรื่อยๆ จนสุดท้ายจะได้คำตอบที่ต้องการ เหมือนลักษณะของกิ่งของต้นไม้ที่แตกออกมา ดังรูปที่ 3.2 เป็นรูปตัวอย่างของ Decision Tree



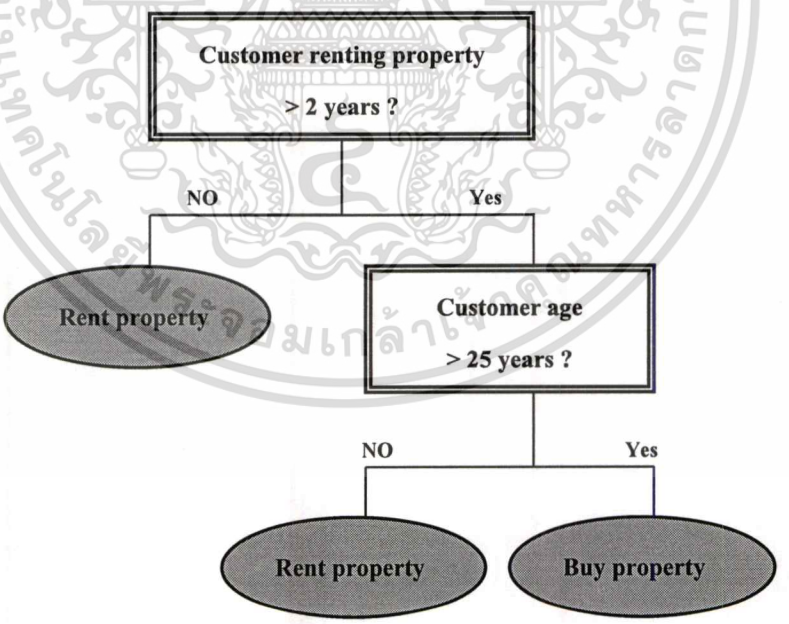
รูปที่ 3.2 แสดงตัวอย่างการแสดงผลของ Decision Tree

จากรูปจะประกอบไปด้วย Node ต่างๆ จุดที่เริ่มต้นของแผนภูมิจะเรียกว่า Root Node หลังจากนั้นจะแตกข้อมูลออกเป็นกิ่งต่างๆ (Branch) ตามทางเลือกของ Node ต่างๆ ซึ่งกระบวนการนี้จะดำเนินการต่อไปเรื่อยๆจนกระทั่งได้ผลลัพธ์สุดท้ายของตัวแปรที่เป็นเป้าหมาย (Target Attribute) เรียกว่า Leaf Node ซึ่งจะเก็บค่าของคำตอบไว้ที่ Node นี้ แต่ในกรณีที่ปริมาณข้อมูลมีจำนวนมาก ทำให้ทางเลือกในการแตกของข้อมูลเป็นไปในลักษณะที่แตกแขนงออกไปหลายทาง อาจจะมีการแตกเอาทางเลือกที่ไม่มีความสำคัญออกมาด้วยเนื่องจากอาจเป็นผลมาจากข้อมูลบางส่วนที่อาจจะเป็นข้อมูลที่มีความผิดพลาด (Noisy Data) ซึ่งแผนภูมิที่ได้จะทำการวิเคราะห์ได้ยาก จึงจำเป็นต้องมีกระบวนการตัดแต่งกิ่งของคำตอบให้เข้าใจได้ง่ายที่สุด โดยจะคัดเลือกเอาทางเลือกที่มีความเป็นไปได้น้อยที่สุดออกไป เราจะเรียกขั้นตอนนี้ว่า การแต่งกิ่ง (Tree Pruning) เพื่อเป็นการคัดเอาผลลัพธ์ที่ไม่ดีออกไป ทำให้ผลของการวิเคราะห์ข้อมูลมีความน่าเชื่อถือมากที่สุด

ยกตัวอย่าง สมมติว่าบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะดูแลอาคารต่าง ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา (Branch), ข้อมูลพนักงาน (Staff), ข้อมูลทรัพย์สิน (Property) และข้อมูลลูกค้า (Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer_Rental), รายการให้เช่า (Rentals), รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีการประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า

" 40 % ของลูกค้าที่เช่าบ้านนานกว่าสองปี และมีอายุเกิน 25 ปี จะซื้อบ้านเป็นของตนเอง โดยกรณีเช่นนี้เกิดขึ้น 35 % ของลูกค้าผู้เช่าบ้านของบริษัท"

ดังรูปที่ 3.3 แสดงให้เห็นถึง Decision Tree สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า



รูปที่ 3.3 แสดงตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าจะซื้อ

จากตัวอย่างพฤติกรรมการเช่าและซื้อบ้านข้างต้น ลองมาดูตัวอย่างที่เป็นรูปธรรมมากขึ้น โดยจะมีตาราง Business_Info แสดงถึงรายการทั้งหมดเกี่ยวกับลูกค้าบ้านเช่าของบริษัทโดยมี

รายละเอียดเกี่ยวกับอายุ และระยะเวลาเช่า รวมทั้งการซื้อบ้านของลูกค้าแต่ละราย ดังนี้

ตารางที่ 3.1 ตารางแสดงรายละเอียดของลูกค้าบ้านเช่า

Age	Rent_Period	Buy
23	3	No
36	1.5	No
20	1.5	No
27	2	Yes
20	1	No
50	2.5	Yes
36	1	No
36	2	Yes
22	2.5	No

SQL สำหรับ Decision Tree ของตัวอย่างนี้แบ่งเป็น 2 ชุด สำหรับปัจจัยแต่ละอย่าง

1. SQL สำหรับ Root Node ดังนี้

```
SELECT      B.Rent_Period, B.Buy, COUNT (*)
FROM        Business_Info B
WHERE       B.Rent_Period > 2
GROUP BY  B.Rent_Period , B.Buy
```

ผลลัพธ์ของ SQL นี้คือ

ตารางที่ 3.2 ตารางแสดงผลลัพธ์ของการแตกกิ่งในกลุ่มแรก

Rent_Period	Buy	Yes	No
1	0	2	
1.5	0	2	
2	2	0	
2.5	1	1	
3	0	1	

2. SQL สำหรับ node ที่เป็น Child ทางขวาของ Root คือ

```
SELECT      B.Age, B.Buy, COUNT(*)
FROM        Business_Info B
WHERE       B.Age > 25
GROUP BY    B.Age, B.Buy
```

ผลลัพธ์ของ SQL นี้คือ

ตารางที่ 3.3 ตารางแสดงผลลัพธ์ในการแตกกิ่งในกลุ่มที่ 2

Rent_Period	Buy	Yes	No
20	0	2	
22	0	1	
23	2	1	
27	1	0	
36	1	2	
50	1	0	

ผลลัพธ์ที่ได้จากแต่ละ Node ของ Decision Tree เรียกว่า AVC Set (Attribute Value, Class Label) จากตัวอย่างข้างต้นจะเห็นว่ามี 2 AVC Sets เพื่อใช้ในการจัดกลุ่มลูกค้า เป็นต้น

ดังนั้นในการใช้เทคนิค Decision Trees จึงต้องขึ้นอยู่กับลักษณะของข้อมูลที่นำมาใช้ในการประมวลผลด้วย ซึ่งเทคนิคนี้จะประกอบไปด้วย Algorithm หลายประเภท ยกตัวอย่างเช่น CHAID (Chi-Square Automatic Interaction Detection), CART (Classification And Regression Tree), ID3, QUEST (Quick Unbiased, Statistical, Efficient, Statistical Tree) เป็นต้น สำหรับโครงการนี้จะทำการเลือกศึกษา CHAID Algorithm เพื่อหาความสัมพันธ์ของกลุ่มข้อมูลว่ามีลักษณะเป็นอย่างไร สามารถทำนายผลการพยากรณ์ได้แม่นยำแค่ไหน

3.3 CHAID Algorithm

สำหรับ Algorithm ที่จะนำมาทำการศึกษาในโครงการนี้ คือ CHAID Algorithm ซึ่งจะเป็น Algorithm ที่มีความนิยมในการนำมาใช้งาน Data Mining ประกอบกับข้อมูลที่เป็นตัวเลข ดังนั้นในการหาความสัมพันธ์ของ CHAID Algorithm จะออกแบบมาเพื่อจัดการแยกแยะค่าตัวแปร Variable โดยเฉพาะ โดยสามารถใช้ได้ทั้งข้อมูลที่เป็นตัวเลขปกติ หรือตัวเลขต่อเนื่อง ซึ่ง CHAID Algorithm

จะทำการแยกชุดข้อมูลโดยพิจารณาจากความสัมพันธ์ที่ได้จากการคำนวณ โดยจะแบ่งจากชุดข้อมูลรวมเป็นซับเซ็ตย่อยระดับ Child Node ตามจำนวนที่กำหนดซึ่งจะทำจนกว่าตัวแปรที่มีไม่มีความสัมพันธ์กับชุดข้อมูลที่เป็นเป้าหมายแล้วก็จะหยุดกระบวนการ ผลที่ได้ก็จะเป็นทางเลือกที่ใช้ในการพยากรณ์ โดยจะมีองค์ประกอบในการใช้ CHAID Algorithm ในการวิเคราะห์ดังนี้

- ตัวแปรที่ใช้ในการพยากรณ์
- ตัวแปรที่เป็นเป้าหมาย
- การกำหนดค่า CHAID Parameters ที่จะใช้ในการวิเคราะห์ชุดข้อมูล

การทำงานของ CHAID Algorithm จะใช้การคำนวณของ Chi-Squared (χ^2), degree of freedom (d.f.) และ p-value ซึ่งจะมีสูตรดังนี้

สมมติมีตารางความถี่จากการทดลอง ซึ่งจะใช้อธิบายวิธีคำนวณของ Chi-Squared ได้ดังนี้

ตารางที่ 3.4 ตารางแสดงความถี่จากการทดลองของ (χ^2)

	A_1	A_2	A_3	...	A_c	
B_1	O_{11}	O_{12}	O_{12}	...	O_{1c}	R_1
B_2	O_{21}	O_{22}	O_{23}	...	O_{2c}	R_2
.	.	.	.		.	.
.	.	.	.		.	.
B_r	O_{r1}	O_{r2}	O_{r3}		O_{rc}	R_r
	C_1	C_2	C_3	...	C_c	N

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(O_{ij} - E_{ij} \right)^2}{E_{ij}}$$

O_{ij} = ค่าของข้อมูลที่เก็บได้

E_{ij} = ค่าของเหตุการณ์ที่น่าจะเกิดขึ้น

ซึ่งค่าของเหตุการณ์ที่น่าจะเกิดจะหาได้ โดยการคูณความน่าจะเป็นแต่ละค่าด้วยผลรวมของจำนวนความถี่ทั้งหมด N คือ

$$E_{ij} = N \cdot \Pr\left(A_i \mathrm{I} B_j \right)$$

ส่วนสูตรที่ใช้ในการ degree of freedom (V) ของข้อมูลจะสามารถคำนวณหาได้จากสูตรดังต่อไปนี้

$$V=(R-1)(C-1)$$

R = จำนวนประเภทของข้อมูลของตัวแปร B

C = จำนวนประเภทของข้อมูลของตัวแปร A

หลังจากนั้นจะนำค่าของ Chi-Squared (χ^2) และ degree of freedom (V) มาใช้ในการคำนวณหาค่าของ p-value ต่อไป โดยการเปิดค่าในตาราง p-value ดังนั้น จะขอยกตัวอย่างของการหาค่าความสัมพันธ์โดยใช้ CHAID Algorithm ได้ดังนี้

ยกตัวอย่างว่ามีข้อมูลที่มีค่าตัวแปร 3 ตัว คือ X1, X2 และ Y เป็น Target Attribute ซึ่งสามารถใช้ CHAID Algorithm มาพิจารณาหาค่าความสัมพันธ์ โดยจะมีข้อมูล ดังนี้

ตัวแปร Y จะมีค่า 4 ประเภท คือ 1, 2, 3, 4

ตัวแปร X1 จะมีค่า 4 ประเภท คือ 1, 2, 3, 4

ตัวแปร X2 จะมีค่า 3 ประเภท คือ 0, 1, 2

ขั้นตอนการทำงานของ CHAID Algorithm จะทำตามขั้นตอนต่อไปนี้

- ขั้นที่ 1 ต้องทำการคำนวณหาการกระจายตัวของ Target Attribute ที่ Root Node ซึ่งในที่นี้คือ ตัวแปร Y ตามข้อมูลในตารางที่ 3.5

ตารางที่ 3.5 ตารางแสดงการกระจายตัวของตัวแปร Y ใน Root Node

Cat	%	N
1	35.00	35
2	8.00	8
3	35.00	35
4	22.00	22
Total	(100.00)	100

- ขั้นที่ 2 เปรียบเทียบค่าความแตกต่างที่น้อยที่สุดทีละคู่ของกลุ่มตัวแปร X เทียบกับ Target Attribute โดยดูจากคู่ที่มีค่า p-value มากที่สุด ซึ่งเกี่ยวข้องกับการกระจายของ Target

Attribute ในแต่ละ Node โดยในตัวอย่างของขั้นตอนนี้ จะแสดงความสัมพันธ์ระหว่างตัวแปร X1 และตัวแปร Y.ใน Node ตามตารางที่ 3.6

ตารางที่ 3.6 ตารางแสดงค่าความสัมพันธ์ระหว่างตัวแปร X1 และตัวแปร Y

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
2	12	2	15	13	42
3	0	1	0	1	2
4	0	0	1	4	5
Column Total	35	8	35	22	100

(χ^2) = 25.63559, d.f. = 9 (p = 0.00234955)

ซึ่งจากตารางข้างต้น จะขอแสดงการคำนวณค่า Chi-Square (χ^2) และ degree of freedom เพื่อสร้างความเข้าใจในการคำนวณ โดยจะยกตัวอย่างการหาค่า Chi-Square (χ^2) จากข้อมูลในตารางที่ 3.6 ได้ดังนี้

โดยต้องเริ่มจากการหาคำนวณค่าของเหตุการณ์ที่น่าจะเกิดขึ้น (E_{ij}) ในแต่ละกรณีที่เกิดขึ้น เช่น กรณีของตัวแปรที่ Y = 1, X1 = 1 จะเท่ากับ 17.85 (= 35*51/100) หรือกรณีของตัวแปรที่ Y = 4, X1 = 4 จะเท่ากับ 1.1 (= 22*5/100) เป็นต้น โดยจะหาในทุกกรณี หลังจากนั้นก็จะนำไปแทนค่าในสูตรการหาค่า Chi-Square (χ^2) ก็จะได้ค่าเท่ากับ 25.63559

ส่วนค่า degree of freedom ก็แทนค่าตามสูตร โดยที่ R = 4, C = 4 ก็จะได้ค่าเท่ากับ 9 ตามตัวอย่าง หลังจากนั้นก็นำค่าที่ได้ทั้งสองค่าไปหาค่า p-value ต่อไป

จากตารางที่ 3.6 จะสามารถคำนวณค่า Chi-Square (χ^2), d.f. และ p-value ตามลำดับ โดยสามารถแสดงผลลัพธ์ได้ตามตารางที่ 3.7 – 3.12

ตารางที่ 3.7 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 2

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
2	12	2	15	13	42
Column Total	35	7	34	17	93

(χ^2) = 9.193281, d.f. = 3 (p = 0.02682849)

ตารางที่ 3.8 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 3

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
3	0	1	0	1	2
Column Total	23	6	19	5	53

$(\chi^2) = 8.079281$, d.f. = 3 (p = 0.0456149)

ตารางที่ 3.9 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 4

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
4	0	0	1	4	5
Column Total	23	5	20	8	56

$(\chi^2) = 19.72078$, d.f. = 3 (p = 0.0001939266)

ตารางที่ 3.10 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 2 และ 3

X1 / Y	1	2	3	4	Row Total
2	12	2	15	13	42
3	0	1	0	1	2
Column Total	12	3	15	14	44

$(\chi^2) = 7.23356$, d.f. = 3 (p = 0.0648145)

ตารางที่ 3.11 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 2 และ 4

X1 / Y	1	2	3	4	Row Total
2	12	2	15	13	42
4	0	0	1	4	5
Column Total	12	2	16	17	47

$(\chi^2) = 4.962482$, d.f. = 3 (p = 0.1745651)

ตารางที่ 3.12 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 3 และ 4

X1 / Y	2	3	4	Row Total
3	1	0	1	2
4	0	1	4	5
Column Total	1	1	5	7

(χ^2) = 3.08, d.f. = 2 (p = 0.2143811)

- ขั้นตอนที่ 3 ในกระบวนการคำนวณของ CHAID Algorithm นี้ จะต้องคำนวณหาค่าของ p-value ของความสัมพันธ์ในกลุ่มของตัวแปรต่างๆที่สัมพันธ์กันในแต่ละกรณี เพื่อนำมาเปรียบเทียบกับค่า Alpha Level α_{merge} (0.05 คือค่าที่กำหนดไว้ในการเปรียบเทียบ) ซึ่งจากการพิจารณาค่าของ p-value ในแต่ละกรณีต่างๆ จะเห็นว่า ค่าความสัมพันธ์ของตัวแปรในกรณีที่ 3 และ 4 (ตารางที่ 3.12) ที่มีค่า p-value เท่ากับ 0.2143811 ซึ่งมีค่า p-value มากที่สุดจากการเปรียบเทียบกับค่า Alpha Level α_{merge} ดังนั้น จะทำให้สามารถยุบรวมกลุ่มของกรณีที่ 3 และ 4 เป็นกลุ่มเดียวกันได้ ทำให้ผลลัพธ์จากการยุบกลุ่ม จะสามารถจัดกลุ่มข้อมูลของตัวแปร X1 และ Y ได้ใหม่ ดังตารางที่ 3.13 – 3.15

ตารางที่ 3.13 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 2

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
2	12	2	15	13	42
Column Total	35	7	34	17	93

(χ^2) = 9.193281, d.f. = 3 (p = 0.02682849)

ตารางที่ 3.14 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 3, 4

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
3, 4	0	1	1	5	7
Column Total	23	6	20	9	58

(χ^2) = 20.25577, d.f. = 3 (p = 0.0001502343)

ตารางที่ 3.15 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 2 และ 3, 4

X1 / Y	1	2	3	4	Row Total
2	12	2	15	13	42
3, 4	0	1	1	5	7
Column Total	12	3	16	18	49

(χ^2) = 6.408565, d.f. = 3 (p = 0.09333908)

- ขั้นตอนที่ 4 ก็จะทำตามกระบวนการนี้ซ้ำต่อไปเรื่อยๆ โดยจะหาค่า p-value ของความสัมพันธ์ของตัวแปรในแต่ละคู่ว่าคู่ใดที่มีค่า p-value สูงที่สุด เมื่อเทียบกับค่า Alpha Level α merge โดยจะพบว่าค่า p-value ของความสัมพันธ์ของตัวแปรในกรณีที่ 2 และ 3, 4 (ตารางที่ 3.15) ที่มีค่า p-value เท่ากับ 0.09333908 ซึ่งมีค่ามากที่สุดจากการเปรียบเทียบกับค่า Alpha Level α merge ดังนั้นก็จะต้องทำการยุบรวมกลุ่มของกรณีที่ 2 และ 3, 4 เข้าด้วยกันเป็นกลุ่มเดียวกันได้อีก ซึ่งจากการยุบรวมกลุ่มในครั้งนี้จะทำให้เราสามารถหยุดการยุบรวมกลุ่มได้ เนื่องจากสามารถแบ่งกลุ่มของข้อมูลของตัวแปร X1 ได้แล้ว
- ขั้นตอนที่ 5 เราจะต้องทำการคำนวณเพื่อที่จะปรับค่าของ p-value สำหรับกลุ่มที่ได้ยุบรวมกันใหม่ โดยใช้ Bonferroni Multiplier โดยใช้ข้อมูลการยุบรวมกลุ่ม จากตารางที่ 3.16

ตารางที่ 3.16 ตารางแสดงค่าความสัมพันธ์ระหว่าง X1 และ Y ในกรณีที่ 1 และ 2, 3, 4

X1 / Y	1	2	3	4	Row Total
1	23	5	19	4	51
2, 3, 4	12	3	16	18	49
Column Total	35	8	35	22	100

(χ^2) = 13.08861, d.f. = 3 (p = 0.004448843)

โดยค่าของ (χ^2), p-value จากข้อมูลความสัมพันธ์ของตัวแปรในกรณีที่ 1 และ 2, 3, 4 ที่ได้คำนวณไว้ข้างต้น จะต้องทำการปรับโดยการใช้ Bonferroni Multiplier โดยที่ลักษณะของข้อมูลจะเป็น Nominal โดยสูตรในการคำนวณจะมีดังนี้

$$B_{free} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{r!(r-i)!}$$

c = จำนวนกลุ่มข้อมูลของตัวแปร X1

r = จำนวนของกลุ่มที่ยุบรวมกัน

ซึ่งจากการคำนวณตามสูตรข้างต้น จะทำให้ค่าของ p-value มีการปรับใหม่ เท่ากับ 0.031142 (= 0.0004448843 × 7)

- ขั้นตอนที่ 6 จะต้องทำการคำนวณการกระจายตัวของตัวแปร X2 ด้วย โดยการทำตามขั้นตอนที่ 2 – 5 ซ้ำอีกครั้ง แต่เปลี่ยนจากตัวแปร X1 เป็นตัวแปร X2 ซึ่งแสดงรายละเอียดค่าความสัมพันธ์ระหว่าง X2 และ Y ตามตารางที่ 3.17

ตารางที่ 3.17 ตารางแสดงค่าความสัมพันธ์ระหว่างตัวแปร X2 และตัวแปร Y

X2 / Y	1	2	3	4	Row Total
0	33	8	34	22	97
1	1	0	1	0	2
2	1	0	0	0	1
Column Total	35	8	35	22	100

(χ^2) = 2.768778, d.f. = 6 (p = 0.8372577)

จากตารางที่ 3.18 จะสามารถคำนวณค่า Chi-Square (χ^2), d.f. และ p-value ตามลำดับ โดยสามารถแสดงผลลัพธ์ได้ตามตารางที่ 3.18 – 3.20

ตารางที่ 3.18 ตารางแสดงค่าความสัมพันธ์ระหว่าง X2 และ Y ในกรณีที่ 0 และ 1

X2 / Y	1	2	3	4	Row Total
0	33	8	34	22	97
1	1	0	1	0	2
Column Total	34	8	35	22	99

(χ^2) = 0.8881097, d.f. = 3 (p = 0.8282962)

ตารางที่ 3.19 ตารางแสดงค่าความสัมพันธ์ระหว่าง X2 และ Y ในกรณีที่ 0 และ 2

X2 / Y	1	2	3	4	Row Total
0	33	8	34	22	97
2	1	0	0	0	1
Column Total	34	8	34	22	98

$(\chi^2) = 1.901759$, d.f. = 3 (p = 0.5930452)

ตารางที่ 3.20 ตารางแสดงค่าความสัมพันธ์ระหว่าง X2 และ Y ในกรณีที่ 1 และ 2

X2 / Y	1	3	Row Total
1	1	1	2
2	1	0	1
Column Total	2	1	3

$(\chi^2) = 0.75$, d.f. = 1 (p = 0.3864762)

- ขั้นตอนที่ 7 จะทำในลักษณะเดียวกับกรณีของตัวแปร X1 คือ จะพิจารณาค่า p-value ของแต่ละกรณีเพื่อเทียบกับค่า Alpha Level α_{merge} ว่าค่า p-value ของกรณีใดที่มีค่ามากที่สุดเมื่อเปรียบเทียบกัน ซึ่งจะพบว่าค่า p-value ของความสัมพันธ์ของตัวแปรในกรณีที่ 0 และ 1 (ตารางที่ 3.19) ที่มีค่า p-value มากที่สุด เท่ากับ 0.8282962 ทำให้สามารถยุบรวมกลุ่มของกรณีที่ 0 และ 1 ได้ และจะเป็นการหยุดการยุบรวมกลุ่ม เนื่องจากสามารถแยกกลุ่มของตัวแปรได้อย่างชัดเจนแล้ว
- ขั้นตอนที่ 8 เราจะต้องทำการคำนวณเพื่อที่จะปรับค่าของ p-value ใหม่อีกครั้ง สำหรับกลุ่มที่ได้ยุบรวมกันใหม่ โดยใช้ Bonferroni Multiplier โดยใช้ข้อมูลการยุบรวมกลุ่มของ X2 และ Y ในกรณีที่ 0, 1 และ 2 จากตารางที่ 3.21

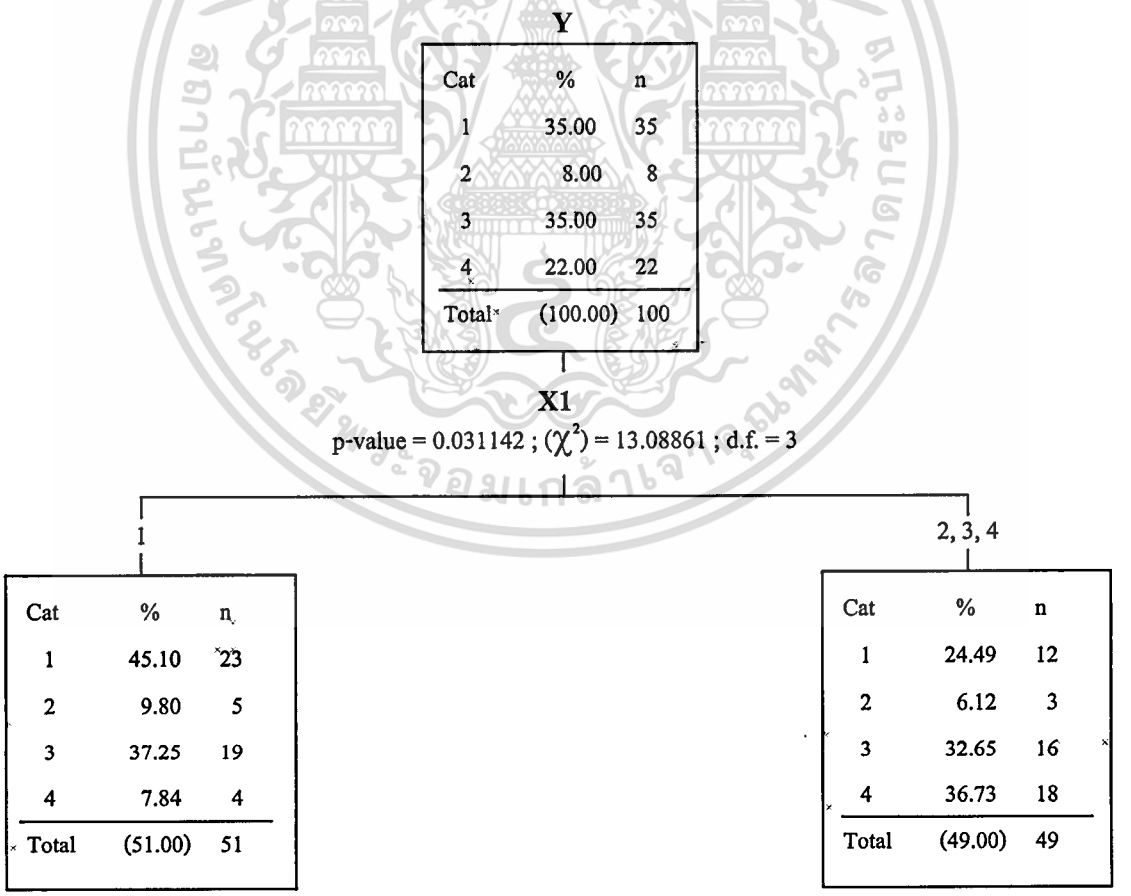
ตารางที่ 3.21 ตารางแสดงค่าความสัมพันธ์ระหว่าง X2 และ Y ในกรณีที่ 0, 1 และ 2

X2 / Y	1	2	3	4	Row Total
0, 1	34	8	35	22	99
2	1	0	0	0	1
Column Total	35	8	35	22	100

$(\chi^2) = 1.875902$, d.f. = 3 (p = 0.5985587)

ซึ่งจากการคำนวณตามสูตร จะทำให้ค่าของ p-value ของตัวแปร X2 มีการปรับใหม่ เท่ากับ 1.79568 (= 0.5985587 × 3) เหมือนกรณีของตัวแปร X1

- ขั้นตอนที่ 9 เป็นขั้นตอนสุดท้ายในการแตก Node เพื่อสร้าง Tree สำหรับ CHAID Algorithm โดยใช้พื้นฐานการยุบรวมกลุ่ม ซึ่งจะพิจารณาจากค่า p-value ที่ได้ทำการปรับค่าใหม่แล้วของแต่ละตัวแปรมาเทียบกับค่า Alpha Level αmerge (ในตัวอย่างนี้ จะเท่ากับ 0.05) แล้วเลือกเอาค่าที่น้อยที่สุดที่น้อยกว่าค่า Alpha Level αmerge มา ซึ่งจากตัวอย่างนี้จะเลือกตัวแปร X1 ในการแตกกิ่งต่อไป เนื่องจาก ค่า p-value ใหม่ของตัวแปร X1 มีค่าน้อยกว่าตัวแปร X2 (p-value ของ X1 = 0.031142, p-value ของ X2 = 1.79568) เมื่อเทียบกับค่า Alpha Level αmerge โดยหลังจากที่เลือกเอาตัวแปร X1 มาใช้ ก็สรุปได้ว่าจะแตกจาก Root Node ออกเป็น 2 Internal Node โดยที่ Internal Node 1 จะเป็นกรณีที่ X1 = 1 จะมีกรณีได้ 35 กรณี ส่วน Internal Node 2 ก็จะเป็นกรณีที่ X1 = 2, 3 หรือ 4 จะมีกรณีได้ในส่วนที่เหลืออีก 65 กรณี ดังรูปที่ 3.4



รูปที่ 3.4 แสดงแผนผัง Tree ของข้อมูลจากตัวอย่าง โดยการใช้ CHAID Algorithm

ดังนั้น การสร้าง Tree จะทำตามขั้นตอนและกระบวนการของ Algorithm ที่เราเลือกมาใช้งานต่อไปเรื่อยๆ จนกระทั่งแตกกิ่งจนได้คำตอบตามกฎเกณฑ์ที่ได้กำหนดไว้ในการสร้าง Tree หรือหยุดสร้าง Tree เมื่อได้ผลลัพธ์ตามที่ต้องการ ซึ่งผลที่เราต้องการก็จะแสดงอยู่ใน Leaf Node ซึ่งเป็น Node สุดท้ายของ Tree ที่สร้างขึ้น

สำหรับการศึกษาในบทนี้ จะทำให้เราทราบถึงการสร้างแบบจำลองในการพยากรณ์ (Predictive Model) โดยใช้การสร้างแผนภูมิต้นไม้ (Decision Tree) เพื่อทำนายผลลัพธ์ของข้อมูลที่เราต้องการ ซึ่งในบทนี้จะเน้นไปถึง Algorithm ที่นำมาใช้ในการศึกษาโครงการนี้ นั่นคือ CHAID Algorithm โดยได้อธิบายถึงหลักการคิดพื้นฐานของ Algorithm ว่ามีกระบวนการทำงานอย่างไร ซึ่งจะใช้เป็นประโยชน์ในการทำโครงการต่อไป

บทที่ 4
วิธีดำเนินการศึกษา

ในบทนี้จะกล่าวถึงกระบวนการและขั้นตอนต่างๆในการนำเอาข้อมูลของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาประมวลผลผ่านทางโปรแกรมสำเร็จรูปที่ได้เตรียมไว้ โดยเริ่มจากการเตรียมข้อมูล, การนำข้อมูลมาทำ Data Mining ไปจนถึงการแสดงผลลัพธ์ที่ได้จากการทำ Data Mining

4.1 การกำหนดวัตถุประสงค์และขอบเขตของการศึกษา

ขั้นตอนแรกของการทำ Data Mining ก็คือ การกำหนดวัตถุประสงค์และขอบเขตของการดำเนินงานว่าเป็นอย่างไร ซึ่งวัตถุประสงค์ของการศึกษาโครงการนี้ คือ ต้องการทำการพยากรณ์ผลของการเลื่อนชั้นของลูกหนี้ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ให้มาเป็นหนี้ปกติได้อย่างแม่นยำ โดยใช้หลักการของ CHAID Algorithm มาใช้ในการวิเคราะห์ชุดข้อมูล โดยผ่านการใช้โปรแกรมสำเร็จรูปในการประมวลผล แล้วนำผลการศึกษาที่ได้มาทำการวิเคราะห์และตีความหมายต่อไป

นอกจากนี้ยังสามารถศึกษาถึงตัวแปรต่างๆของข้อมูลที่นำมาใช้ว่า ตัวแปรใดมีผลกระทบต่อผลการเลื่อนชั้นของลูกหนี้มากน้อยเพียงใด ทำให้ผู้บริหารสามารถวิเคราะห์และเลือกวางแผนกลยุทธ์ได้อย่างถูกต้องเพื่อผลการปรับปรุงโครงสร้างหนี้ที่มีโอกาสสำเร็จสูงสุด

4.2 การเตรียมข้อมูล

เป็นขั้นตอนที่จะใช้เวลานานที่สุดในการทำ Data Mining โดยข้อมูลที่จะนำมาใช้ในการประมวลผลในโครงการนี้ เป็นข้อมูลที่ได้จากสถาบันการเงินแห่งหนึ่งโดยเป็นข้อมูลของลูกหนี้แต่ละรายที่อยู่ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ ซึ่งเป็นข้อมูลตั้งแต่ มกราคม 2543 จนถึง ณ สิ้นปี 2546 โดยวัตถุประสงค์ของการนำข้อมูลประเภทนี้มาใช้เนื่องจากต้องการศึกษาดูถึงปัจจัยที่มีผลกระทบต่อโอกาสในการเลื่อนชั้นหนี้ของลูกหนี้จากกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาเป็นลูกหนี้ปกติ เพื่อนำมาใช้ประกอบการตัดสินใจด้านกลยุทธ์ขององค์กร และทดสอบโมเดลที่สร้างขึ้นมาใช้ในการพยากรณ์ผลว่ามีความแม่นยำในการพยากรณ์มากน้อยเพียงใด ซึ่งถ้าหากโมเดลที่นำมาใช้สามารถทำนายผลได้อย่างถูกต้องในระดับความเชื่อมั่นที่สูงมาก ก็จะสามารถนำมาใช้เป็นแนวทางในการประมาณการว่าโอกาสในการปรับปรุงโครงสร้างหนี้ได้สำเร็จมีจำนวนเท่าไร

จากข้อมูลที่นำมาประมวลผลนั้น ได้นำเอาข้อมูลลูกหนี้ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาจำนวน 8,000 ราย มีจำนวน Attribute ทั้งหมด 10 Attribute โดยจะแสดงรายละเอียดของแต่ละ Attribute ได้ ดังนี้

- Region เป็น Attribute ของข้อมูลที่แสดงถึงภูมิภาคที่อยู่ของลูกหนี้ว่าตั้งอยู่ในภูมิภาคใดของประเทศ โดยจะแยกเป็น 5 กลุ่ม ดังตารางที่ 4.1

ตารางที่ 4.1 ตารางแสดงความหมายของค่าใน Attribute Region

ค่าใน Attribute	ความหมาย
BKK	กรุงเทพมหานคร
Central	ภาคกลาง
North	ภาคเหนือ
NE	ภาคตะวันออกเฉียงเหนือ
South	ภาคใต้

- New Depart เป็น Attribute ของข้อมูลที่ระบุว่าลูกหนี้แต่ละรายสังกัดอยู่กับส่วนงานใดขององค์กร จะแบ่งได้เป็น 3 หน่วยงาน ดังตารางที่ 4.2

ตารางที่ 4.2 ตารางแสดงความหมายของค่าใน Attribute New Depart

ค่าใน Attribute	ความหมาย
Dep1	หน่วยงานที่ 1
Dep2	หน่วยงานที่ 2
Dep3	หน่วยงานที่ 3

- Method เป็น Attribute ที่ระบุถึงวิธีที่ใช้ในการปรับปรุงโครงสร้างหนี้ของลูกหนี้ที่องค์กรกำหนดไว้ให้ โดยมีทั้งหมด 9 วิธี ดังตารางที่ 4.3

ตารางที่ 4.3 ตารางแสดงความหมายของค่าใน Attribute Method

ค่าใน Attribute	ความหมาย
M1	ลดเงินต้นและ / หรือ ดอกเบี้ยค้างรับ
M2	ลดอัตราดอกเบี้ยในสัญญาปรับโครงสร้างหนี้

ค่าใน Attribute	ความหมาย
M3	แปลงหนี้เป็นทุน (รับหุ้นสามัญของลูกหนี้แทนภาระหนี้)
M4	ขยายระยะเวลารับชำระหนี้ (จากเดิมที่เป็นระยะยาวอยู่แล้ว)
M5	ปรับหนี้ระยะสั้นเป็นหนี้ระยะยาว
M6	ให้ระยะเวลาปลอดหนี้ (เงินต้นและ / หรือ ดอกเบี้ย)
M7	รับโอนทรัพย์สินที่เป็นหลักประกันหนี้
M8	รับโอนทรัพย์สินที่ไม่ใช่หลักประกันหนี้
M9	รับโอนทรัพย์สินโดยมีสัญญาให้สิทธิลูกหนี้ขอโอนกลับคืน
Oth	วิธีอื่นๆที่นอกเหนือจาก 9 วิธีที่กำหนดไว้

- Times เป็น Attribute ที่ระบุจำนวนครั้งที่ทำการปรับปรุงโครงสร้างหนี้ของลูกหนี้ว่าเคยทำการปรับปรุงโครงสร้างหนี้กับองค์กรมาแล้วกี่ครั้ง โดยในข้อมูลจะระบุไว้ 5 ครั้ง ดังตารางที่ 4.4

ตารางที่ 4.4 ตารางแสดงความหมายของค่าใน Attribute Times

ค่าใน Attribute	ความหมาย
Time1	ลูกหนี้ปรับโครงสร้างหนี้ครั้งที่ 1
Time2	ลูกหนี้ปรับโครงสร้างหนี้ครั้งที่ 2
Time3	ลูกหนี้ปรับโครงสร้างหนี้ครั้งที่ 3
Time4	ลูกหนี้ปรับโครงสร้างหนี้ครั้งที่ 4
Time5	ลูกหนี้ปรับโครงสร้างหนี้ครั้งที่ 5

- Bus Type เป็น Attribute ของข้อมูลที่จะแสดงถึงประเภทของธุรกิจของลูกหนี้ที่อยู่ในกลุ่มข้อมูล โดยจะระบุไว้ทั้งหมด 12 ประเภทธุรกิจ ดังตารางที่ 4.5

ตารางที่ 4.5 ตารางแสดงความหมายของค่าใน Attribute Bus Type

ค่าใน Attribute	ความหมาย
T1	ธุรกิจประเภทการเกษตร ประมงและป่าไม้
T2	ธุรกิจประเภทการเหมืองแร่และย่อยหิน

ค่าใน Attribute	ความหมาย
T3	ธุรกิจประเภทการอุตสาหกรรม
T4	ธุรกิจประเภทการก่อสร้าง
T5	ธุรกิจประเภทการค้าส่งและค้าปลีก
T6	ธุรกิจประเภทการส่งสินค้าออก
T7	ธุรกิจประเภทการนำสินค้าเข้า
T8	ธุรกิจประเภทการธนาคารและธุรกิจการเงิน
T9	ธุรกิจประเภทเกี่ยวกับอสังหาริมทรัพย์
T10	ธุรกิจประเภทการสาธารณูปโภค
T11	ธุรกิจประเภทการบริการ
T12	ธุรกิจประเภทการอุปโภคบริโภคส่วนบุคคล

- OS เป็น Attribute ของข้อมูลที่แสดงถึงจำนวนของยอดหนี้ของลูกหนี้แต่ละรายในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ ว่ามีจำนวนเท่าไร โดยในการนำข้อมูลยอดหนี้มาใช้นี้ จะทำการแบ่งกลุ่มของยอดหนี้ไว้เป็นช่วงๆ ซึ่งสามารถแบ่งออกเป็น 5 กลุ่ม ดังตารางที่ 4.6

ตารางที่ 4.6 ตารางแสดงความหมายของค่าใน Attribute OS

ค่าใน Attribute	ความหมาย
Group1	ลูกหนี้ที่มียอดหนี้ต่ำกว่า 1 ล้านบาท
Group2	ลูกหนี้ที่มียอดหนี้ระหว่าง 1 - 5 ล้านบาท
Group3	ลูกหนี้ที่มียอดหนี้ระหว่าง 5 – 20 ล้านบาท
Group4	ลูกหนี้ที่มียอดหนี้ระหว่าง 20 – 100 ล้านบาท
Group5	ลูกหนี้ที่มียอดหนี้ตั้งแต่ 100 ล้านบาทขึ้นไป

- Int เป็น Attribute ของข้อมูลที่แสดงถึงอัตราดอกเบี้ยที่ใช้ในการปรับปรุงโครงสร้างหนี้ของลูกหนี้ที่ได้ระบุไว้ในสัญญาการปรับปรุงโครงสร้างหนี้ โดยกำหนดไว้ 7 ประเภท ดังตารางที่ 4.7

ตารางที่ 4.7 ตารางแสดงความหมายของค่าใน Attribute Int

ค่าใน Attribute	ความหมาย
Rate1	อัตราดอกเบี้ย MLR
Rate2	อัตราดอกเบี้ย MLR + Spread
Rate3	อัตราดอกเบี้ย MLR - Spread
Rate4	อัตราดอกเบี้ย GLR
Rate5	อัตราดอกเบี้ย GLR + Spread
Rate6	อัตราดอกเบี้ย GLR - Spread
Rate7	อัตราดอกเบี้ยอื่นๆที่นอกเหนือจากที่กำหนด

ซึ่งอัตราดอกเบี้ย MLR (Minimum Loan Rate) หมายถึง อัตราดอกเบี้ยสำหรับลูกค้ารายใหญ่ชั้นดีแบบมีระยะเวลา ส่วนอัตราดอกเบี้ย GLR (General Lending Rate) หมายถึง อัตราดอกเบี้ยเงินให้กู้ยืมมาตรฐานทั่วไป

- Class เป็น Attribute ที่แสดงถึงระดับชั้นของลูกหนี้ที่เข้ามาทำสัญญาปรับปรุงโครงสร้างหนี้ ซึ่งจะแยกเป็น 3 ระดับ ดังตารางที่ 4.8

ตารางที่ 4.8 ตารางแสดงความหมายของค่าใน Attribute Class

ค่าใน Attribute	ความหมาย
SS	ชั้นต่ำกว่ามาตรฐาน
D	ชั้นสงสัย
DL	ชั้นสงสัยจะสูญ

- Pay Period เป็น Attribute ของข้อมูลที่แสดงถึงงวดการผ่อนชำระหนี้ของลูกหนี้แต่ละรายว่ามีรอบระยะเวลาที่ต้องผ่อนชำระเป็นระยะเวลาเท่าใด โดยแบ่งเป็น 4 ประเภท ดังตารางที่ 4.9

ตารางที่ 4.9 ตารางแสดงความหมายของค่าใน Attribute Pay Period

ค่าใน Attribute	ความหมาย
1M	งวดชำระ 1 เดือนต่อครั้ง
3M	งวดชำระ 3 เดือนต่อครั้ง

ค่าใน Attribute	ความหมาย
6M	งวดชำระ 6 เดือนต่อครั้ง
12M	งวดชำระ 12 เดือนต่อครั้ง

- Result เป็น Attribute เป้าหมายของโมเดลที่เราต้องการสร้าง โดยที่แสดงให้ทราบว่าลูกหนี้สามารถเลื่อนชั้นจากลูกหนี้ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาเป็นลูกหนี้ปกติได้สำเร็จหรือไม่ โดยจะมีค่าแสดงไว้ 2 ค่า คือ Yes และ No

โดยข้อมูลที่ได้มาแต่ละ Attribute จะมีความหลากหลายของข้อมูลมาก จึงจำเป็นที่จะต้องมีการกำหนดและแปลงข้อมูลให้อยู่ในลักษณะเดียวกัน เพื่อให้ผลการพยากรณ์มีความถูกต้องแม่นยำมากที่สุด และสามารถตีความได้อย่างมีประสิทธิภาพ เช่น ข้อมูลของยอดหนี้ จะเป็นตัวเลขที่กระจายกันมาก ก็จะทำการแบ่งเป็นช่วงของตัวเลข เพื่อเป็นการกำหนดให้ข้อมูลเป็นหมวดหมู่ หรือข้อมูลบาง Attribute ที่มีรูปแบบที่แปลกแตกต่างไป ก็จะทำการแปลงข้อมูลเหล่านั้นให้มีรูปแบบที่เหมือนกันตรงตามที่ได้กำหนดไว้ ซึ่งพร้อมที่จะนำมาใช้งานกับโปรแกรมสำเร็จรูปได้ เป็นต้น

หลังจากที่ได้ทำการคัดเลือกเอาเฉพาะข้อมูลที่เราต้องการนำมาใช้ในการวิเคราะห์จริงๆ แล้ว ต่อไปก็จะทำการกรองข้อมูลโดยจะเป็นการกำจัดข้อมูลที่มีความผิดพลาด (Noisy Data) และข้อมูลที่ขาดหายไป (Missing Data) ให้หมดไป เมื่อทำการกรองข้อมูลเสร็จแล้วก็พร้อมที่จะนำเอาข้อมูลมาใช้ ซึ่งจะได้ข้อมูลที่มีลักษณะและรายละเอียด ดังตารางต่อไปนี้

ตารางที่ 4.10 ตารางแสดงค่าและประเภทของข้อมูล

ชื่อ Attribute	ค่าใน Attribute	ประเภทของข้อมูล
OS	Group1, Group2, Group3, Group4, Group5	Text
Region	BKK, Central, North, NE, South	Text
New Depart	Dep1, Dep2, Dep3	Text
Method	M1, M2, M3, M4, M5, M6, M7, M8, M9, Oth	Text
Bus Type	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12	Text
Times	Time1, Time2, Time3, Time4, Time5	Text
Int	Rate1, Rate2, Rate3, Rate4, Rate5, Rate6, Rate7	Text
Pay Period	1M, 3M, 6M, 12M	Text
Class	SS, D, DL	Text
Result	Yes, No	Text

4.3 การทำ Data Mining

ในการทำ Data Mining ของโครงการนี้ จะทำการศึกษา CHAID Algorithm โดยจะใช้โปรแกรมสำเร็จรูปที่ชื่อว่า Answer Tree ในการสร้างโมเดลเพื่อพยากรณ์ผลการปรับโครงสร้างหนี้ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ ซึ่งโปรแกรมสำเร็จรูปอันนี้จะสามารถรับข้อมูลได้ 3 ลักษณะ คือ เป็นข้อมูล Worksheet จาก Excel, ข้อมูลที่เป็น SPSS File และ ข้อมูลที่เป็น Text ซึ่งเราจะนำข้อมูลเข้าในลักษณะของ Worksheet ที่เป็น Excel ที่ได้ทำการแปลงข้อมูลพร้อมที่จะนำเข้าแล้ว ซึ่งลักษณะวิธีการสร้าง Tree ของโปรแกรม Answer Tree จะมี 4 แบบ คือ

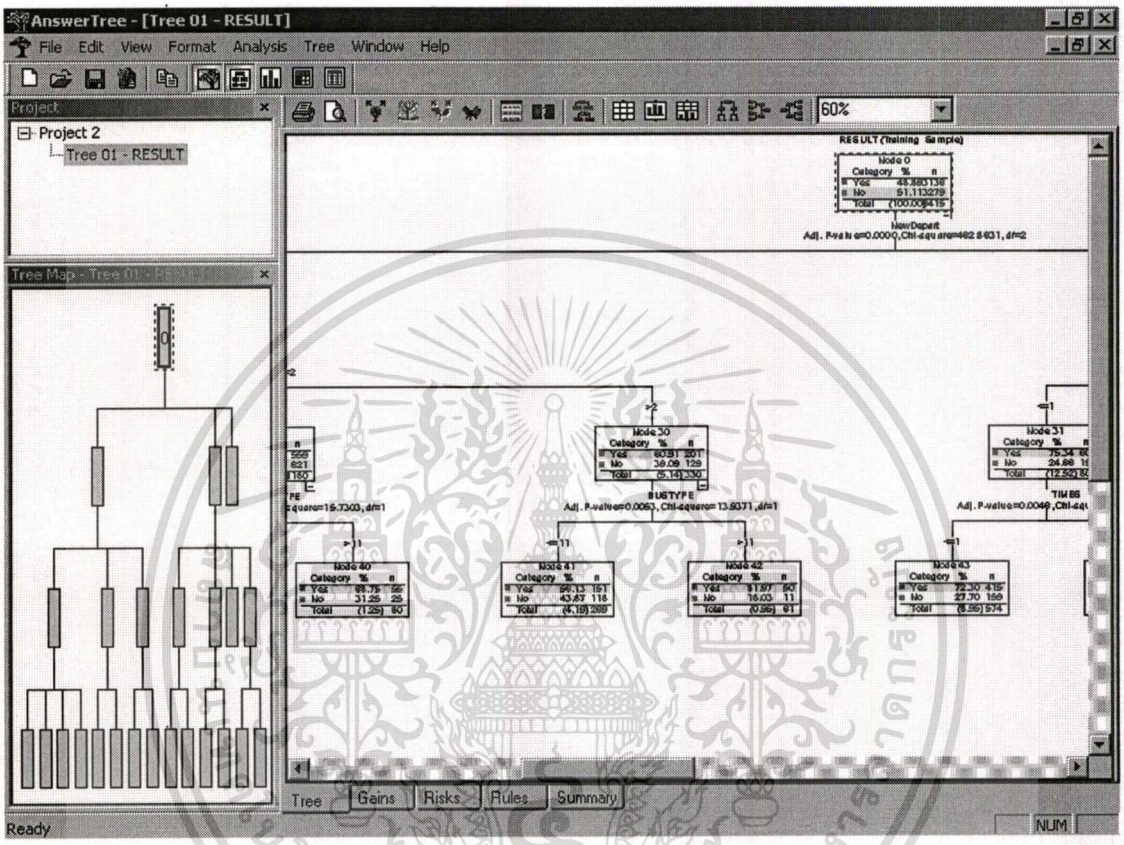
- Merging เป็นการหาความสัมพันธ์ของตัวแปรที่ใช้ในการพยากรณ์กับข้อมูลที่เป็น Target Attribute โดยโปรแกรมจะยุบรวมตัวแปรที่ไม่มีความสำคัญเป็นกลุ่มหนึ่ง และตัวแปรที่มีความสำคัญเป็นอีกกลุ่มหนึ่ง
- Splitting เป็นการเลือกจุดที่จะแตกกิ่งของ Tree โดยอาจใช้ตัวแปรประชากรซึ่งถูกเลือกโดยการเปรียบเทียบกับประชากรทั้งหมด
- Stopping จะใช้กฎเกณฑ์ ซึ่งถูกกำหนดว่าเมื่อไรจึงจะหยุดการสร้าง Tree
- Pruning เป็นการยุบกิ่งซึ่งเมื่อแตกออกไปแล้วมีค่าหรือกลุ่มข้อมูลที่แสดงข้อมูลน้อยและนำมาใช้ในการพยากรณ์โดยที่ไม่เกิดประโยชน์

โดยในโปรแกรมสำเร็จรูป Answer Tree นี้จะมี Algorithm ที่จะใช้ในการทำ Data Mining ให้เลือกใช้อยู่ 4 Algorithm ดังนี้

- CHAID (Chi-squared Automatic Interaction Detector) เป็นการหาค่าความสัมพันธ์โดยใช้วิธีการของ Chi-squared Statistics เพื่อใช้ในการแตกกิ่งที่ดีที่สุด
- Exhaustive CHAID เป็นการปรับปรุงการทำงานของ CHAID เพื่อให้สามารถหาความสัมพันธ์ของชุดข้อมูลเพื่อใช้ในการแตกกิ่งอย่างละเอียดมากขึ้น
- CART (Classification and Regression Trees) จะสร้าง Binary Tree โดยอาศัยการคำนวณของ Ginni Index และ Twoing
- QUEST (Quick, Unbiased, Efficient Statistical Tree) เป็นวิธีการในการคำนวณอย่างรวดเร็ว เพื่อหลีกเลี่ยงการขัดแย้งระหว่างตัวแปรที่ใช้ในการพยากรณ์

ซึ่งผลที่ได้จาก Algorithm ทั้ง 4 ประเภทนี้จะแสดงผลลัพธ์ในรูปแบบที่เหมือนกัน คือ จะทำการอธิบายความสัมพันธ์ของกลุ่มข้อมูลเพื่อนำมาใช้ในการพยากรณ์ (Predictive) หรือใช้ในการแบ่งกลุ่ม (Classification) ซึ่ง Algorithm จะทำการเลือกค่าของตัวแปรที่ดีที่สุดในการดำเนินการโดยกระบวนการจะอาศัยการทำซ้ำหลายๆครั้งจนกว่าจะได้ Tree ที่คิดว่าเหมาะสมในการอธิบายค่าความสัมพันธ์ของกลุ่มของข้อมูล แม้ว่า Algorithm ทั้ง 4 แบบนี้ จะมีความแตกต่างทั้งกระบวนการ

คิด วิธีการในการค้นหาความสัมพันธ์และคุณลักษณะต่างๆ ซึ่งในโปรแกรมสำเร็จรูป Answer Tree นี้ เราจะเลือกใช้ CHAID Algorithm มาใช้ในการสร้าง Tree ก็จะได้ผลลัพธ์ออกมาดังรูปที่ 4.1



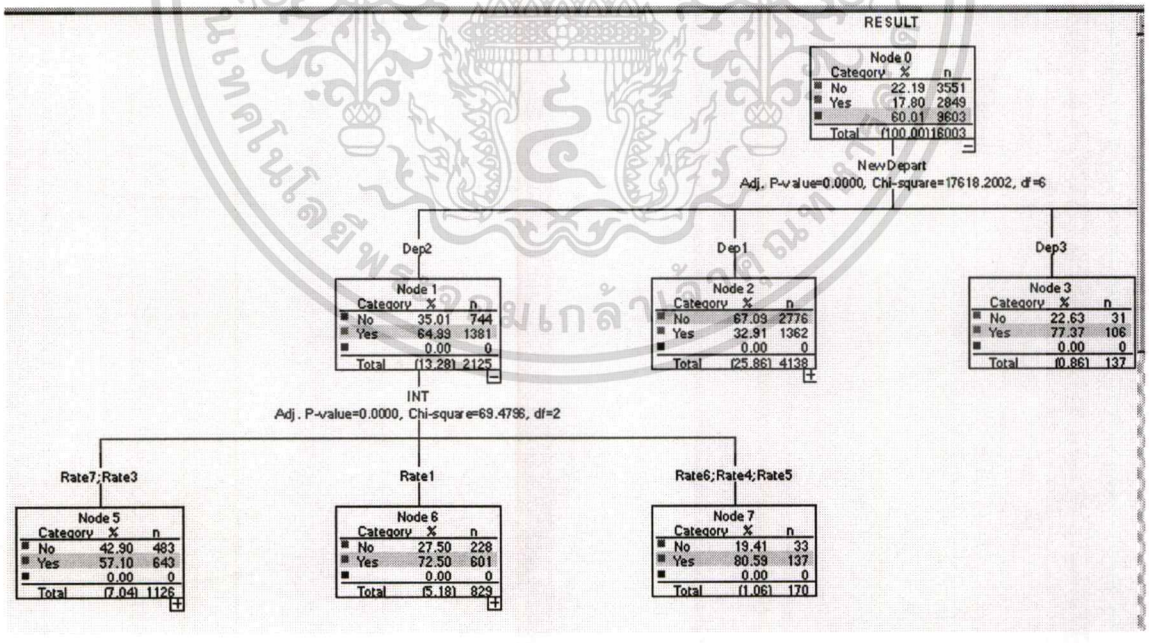
รูปที่ 4.1 แสดงผล Tree ที่ได้จากโปรแกรม Answer Tree

จากการวิเคราะห์ข้อมูลของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) จำนวน 8,000 ราย โดยใช้ CHAID Algorithm นั้น ข้อมูลที่นำมาใช้นั้นไม่ได้มีการกำหนด Validation ในโปรแกรม แต่จะแบ่งกลุ่มข้อมูลออกเป็น 2 ชุด โดยชุดแรกจะใช้ในการ Training Data เพื่อให้โปรแกรมได้เรียนรู้ความสัมพันธ์ของข้อมูลที่นำมาใช้สร้างโมเดลนี้ จำนวน 6,400 รายหรือคิดเป็นร้อยละ 80 ของข้อมูลทั้งหมด ส่วนข้อมูลชุดที่สองจะใช้ในการ Testing Data เพื่อทดสอบว่า Tree ที่ได้มาจากการ Training นั้น มีความถูกต้อง น่าเชื่อถือหรือไม่ เป็นการทดสอบโมเดลที่สร้างขึ้นมาอีกครั้งหนึ่ง โดยใช้ข้อมูลในการ Testing นี้ จำนวน 1,600 ราย หรือคิดเป็นร้อยละ 20 ของข้อมูลทั้งหมด โดยจะพิจารณาลักษณะของผลการพยากรณ์ที่ได้ออกเป็น 2 ลักษณะ คือ

- กลุ่มของลูกหนี้ที่สามารถชำระหนี้ได้ตามเงื่อนไข (Attribute Result = Yes) โดยกลุ่มนี้จะเป็นกลุ่มลูกหนี้ที่สามารถจ่ายชำระหนี้ได้ตามเงื่อนไขจนสามารถเลื่อนชั้นจากกลุ่มหนี้ที่ไม่

ก่อให้เกิดรายได้มาเป็นลูกหนี้ปกติได้ โดยการพิจารณานี้จะกำหนดค่าความเชื่อมั่นของแต่ละกรณีให้มีค่าความเชื่อมั่นไม่น้อยกว่า 75% เนื่องจาก เป็นค่าความเชื่อมั่นที่มีโอกาสเกิดขึ้นได้สูงในระดับหนึ่ง และการสร้างโมเดลนี้ ก็เพื่อค้นหาลักษณะรูปแบบความสัมพันธ์ของตัวแปรของข้อมูลที่น่าจะเกิดขึ้น เพื่อเป็นแนวทางในการพิจารณาเงื่อนไขที่มีความน่าสนใจจากการอ่านผลจาก Tree ที่ได้สร้างขึ้น ดังนั้น จะสามารถนำเสนอแนวทางที่มีความน่าสนใจออกเป็นกรณีตัวอย่างที่น่าสนใจได้ 4 กรณี ดังนี้

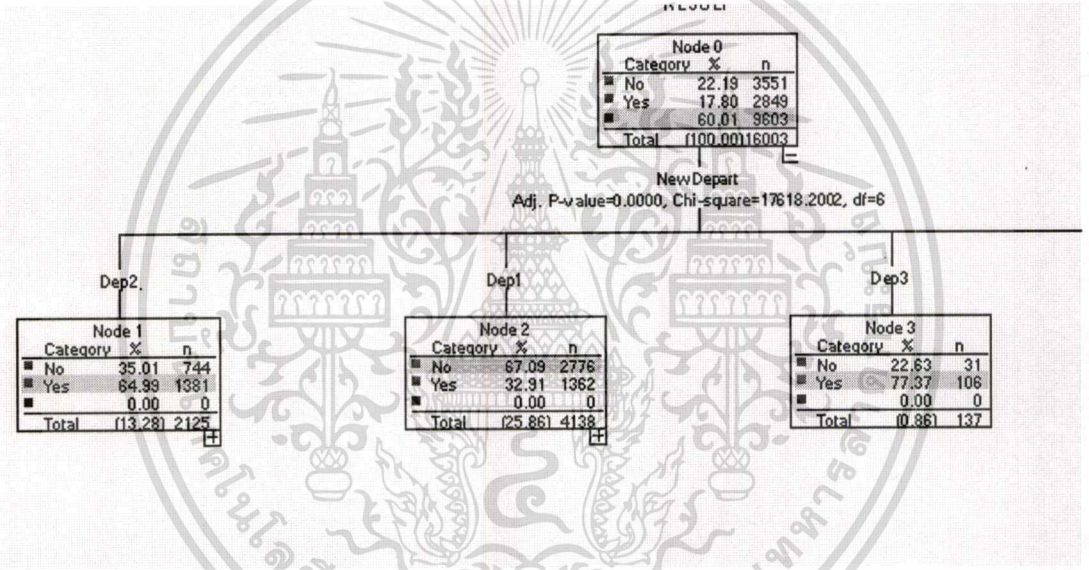
- กรณีที่ 1 จะเลือก Attribute New Depart เป็นตัวหลักที่ใช้ในการแบ่งข้อมูล โดยเลือกค่าที่เป็น Dep2 และ Attribute Int เป็น Rate4, Rate5, Rate6 โดยมีความหมายว่า ถ้าลูกหนี้ในกรณีนี้เป็นลูกหนี้ที่อยู่ในความดูแลของหน่วยงานที่ 2 และใช้อัตราดอกเบี้ยที่ผ่อนชำระเป็น GLR เป็นเกณฑ์หลักในการผ่อนชำระ จะมีโอกาสที่ลูกหนี้สามารถชำระหนี้ได้ตามเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้สำเร็จสูงถึง 80.59% ของจำนวนข้อมูลใน Node นั้น (มีค่าความเสี่ยงที่จะพยากรณ์ผิดพลาด เท่ากับ 1.06%) ซึ่งสามารถเขียนในรูปของ If-Then Rule ได้ว่า If *NewDepart = Dep2 and Int = Rate4, Rate5, Rate6* Then *Result = Yes* โดยแสดงผลได้ดังรูปที่ 4.2



รูปที่ 4.2 แสดงรูป Tree ที่ได้ในกรณีที่ 1

- กรณีที่ 2 เป็นกรณีที่โปรแกรมเลือกเอาค่าของ Attribute New Depart มาเป็นตัวหลักในการแบ่งข้อมูลในชั้นแรกเหมือนกับกรณีที่ 1 โดยเลือกค่าที่เป็น Dep3 เพียง

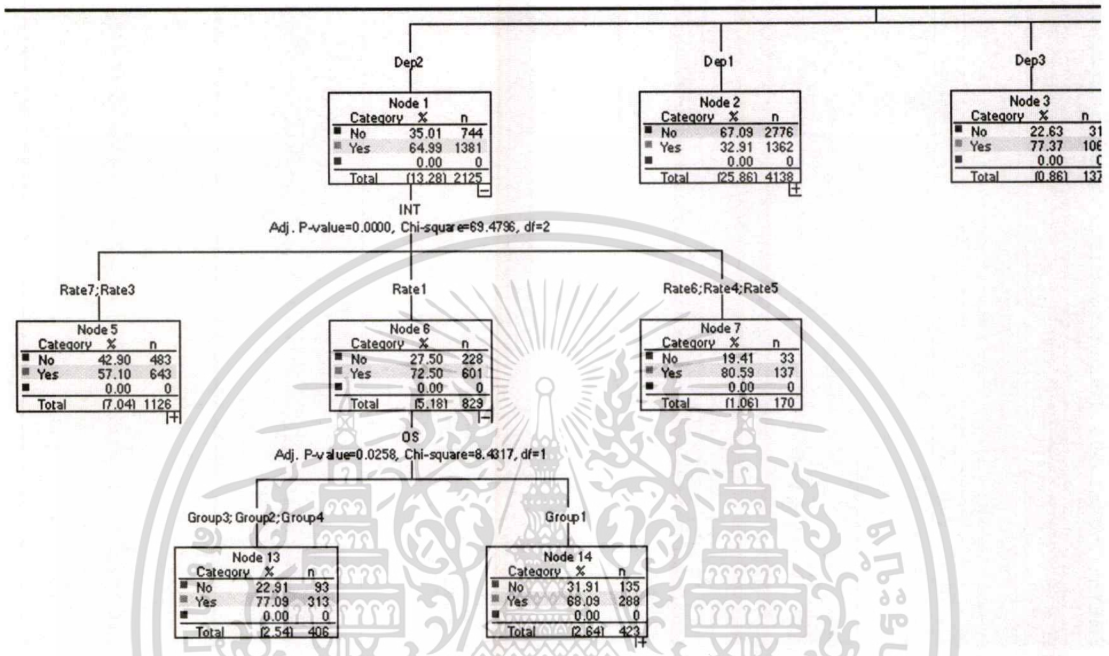
ค่าเดียว อาจเนื่องมาจากข้อมูลที่แตกต่อไปจาก Node นี้ ไม่มีความสำคัญหรือไม่ได้ให้ความหมายที่มีความสำคัญในการแปลความจาก Tree ที่สร้างขึ้น โปรแกรมจึงทำการแต่งกิ่ง (Pruning) ทำให้ไม่มีการแตกกิ่งต่อจาก Node นี้ ซึ่งโมเดลนี้จะแปลความหมายได้ว่า ถ้าลูกหนี้ในกรณีนี้อยู่ในความดูแลของหน่วยงานที่ 3 แล้วนั้นจะมีโอกาสที่ลูกหนี้สามารถชำระหนี้ได้ตามเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้ได้สำเร็จสูงถึง 77.37% ของจำนวนข้อมูลใน Node นั้น (มีค่าความเสี่ยงที่จะพยากรณ์ผิดพลาด เท่ากับ 0.86%) %) ซึ่งสามารถเขียนในรูปของ If-Then Rule ได้ว่า If *NewDepart* = *Dep3* Then *Result* = *Yes* โดยแสดงผลได้ดังรูปที่ 4.3



รูปที่ 4.3 แสดงรูป Tree ที่ได้ในกรณีที่ 2

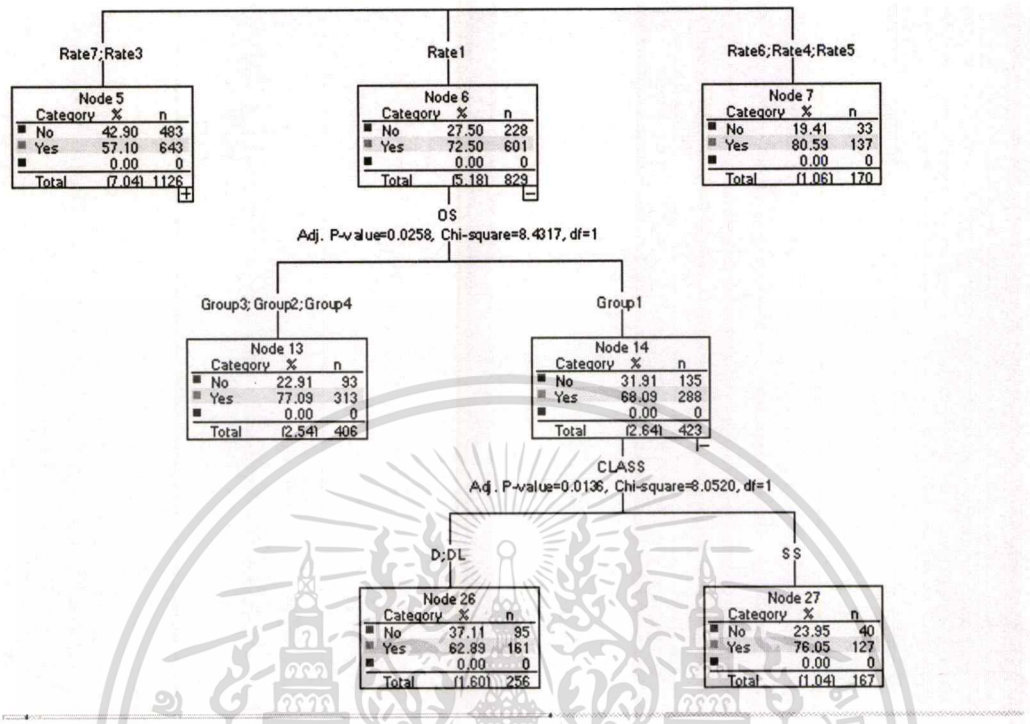
- กรณีที่ 3 เป็นกรณีที่โปรแกรมจะเลือกเอาค่าของ Attribute New Depart ที่เป็น Dep2, เลือกค่า Attribute Int ที่เป็น Rate1 และเลือกค่า Attribute OS เป็น Group2, Group3, Group4 ซึ่งโมเดลนี้จะสามารถแปลความหมายได้ว่า ถ้าลูกหนี้ NPLs ที่อยู่ในกรณีนี้อยู่ในความดูแลของหน่วยงานที่ 2, ใช้อัตราดอกเบี้ยในการผ่อนชำระเป็น MLR และมียอดหนี้ที่ 1 ล้านบาทถึง 100 ล้านบาท จะมีโอกาสที่ลูกหนี้สามารถชำระหนี้ได้ตามเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้สำเร็จสูงถึง 77.09% ของจำนวนข้อมูลใน Node นั้น (มีค่าความเสี่ยงที่จะพยากรณ์ผิดพลาด เท่ากับ 2.54%) %) ซึ่งสามารถเขียนในรูปของ If-Then Rule ได้ว่า If *NewDepart* =

Dep2 and Int = Rate1 and OS = Group2, Group3, Group4 Then *Result = Yes*
โดยจะสามารถแสดงผลได้ในรูปของ Tree ได้ดังรูปที่ 4.4



รูปที่ 4.4 แสดงรูป Tree ที่ได้ในกรณีที่ 3

- กรณีที่ 4 เป็นกรณีที่โปรแกรมจะเลือกเอาค่าของ Attribute New Depart เป็น Attribute หลักเหมือนแต่ละกรณีที่ผ่านมาโดยจะเลือกค่าที่เป็น Dep2, เลือกค่า Attribute Int ที่เป็น Rate1, เลือกค่า Attribute OS เป็น Group1 และเลือกค่า Attribute Class เป็น SS ซึ่งโมเดลนี้จะสามารถแปลความหมายได้ว่า ถ้าลูกหนี้ NPLs ที่อยู่ในกรณีนี้อยู่ในความดูแลของหน่วยงานที่ 2, ใช้อัตราดอกเบี้ยในการผ่อนชำระเป็น MLR , มียอดหนี้ต่ำกว่า 1 ล้านบาทและเป็นลูกหนี้ที่มีชั้นหนี้อยู่ในระดับชั้นต่ำกว่ามาตรฐาน จะมีโอกาสที่ลูกหนี้สามารถชำระหนี้ได้ตามเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้ได้สำเร็จสูงถึง 76.05% ของจำนวนข้อมูลใน Node นั้น (มีค่าความเสี่ยงที่จะพยากรณ์ผิดพลาด เท่ากับ 1.04%) ซึ่งสามารถเขียนในรูปของ If-Then Rule ได้ว่า If *NewDepart = Dep2 and Int = Rate1 and OS = Group1 and Class = SS* Then *Result = Yes* โดยจะสามารถแสดงผลได้ในรูปของ Tree ได้ดังรูปที่ 4.5

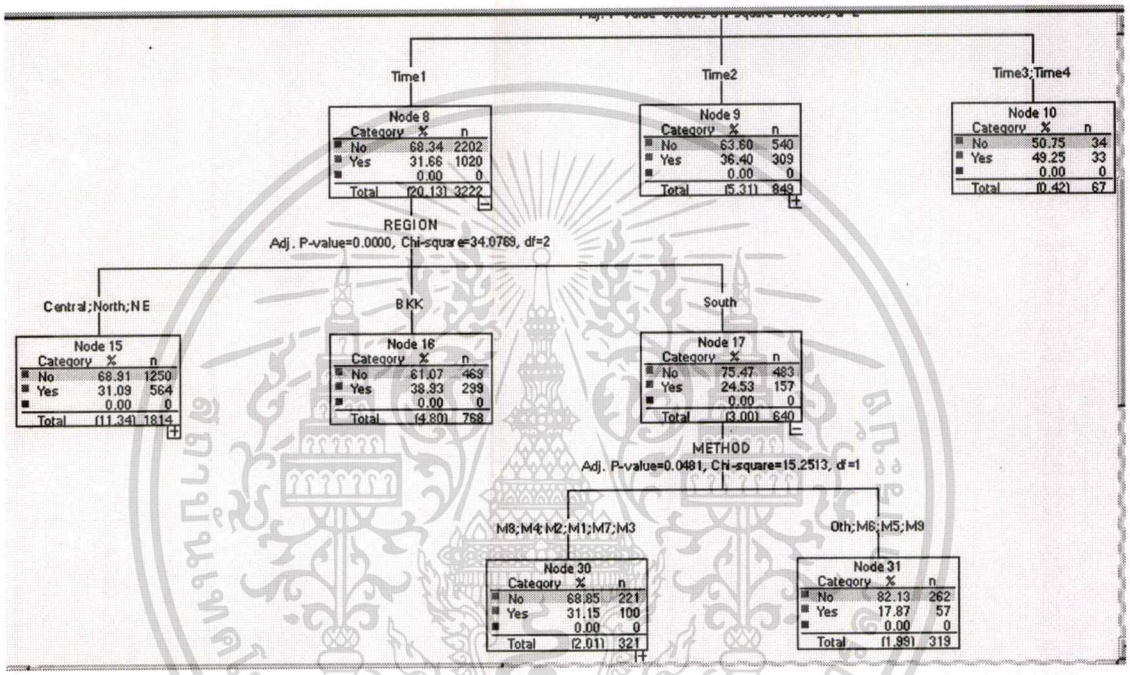


รูปที่ 4.5 แสดงรูป Tree ที่ได้ในกรณีที่ 4

- กลุ่มของลูกหนี้ที่ไม่สามารถทำตามเงื่อนไขที่ได้ระบุไว้ในสัญญาปรับปรุงโครงสร้างหนี้ได้สำเร็จ (Attribute Result = No) ซึ่งจะมีผลทำให้การเลื่อนชั้นของลูกหนี้ไม่สามารถเปลี่ยนจากหนี้ที่ไม่ก่อให้เกิดรายได้ไปเป็นหนี้ปกติได้ โดยจะใช้เกณฑ์ในการพิจารณาที่เหมือนกัน คือ จะกำหนดค่าความเชื่อมั่นของผลพยากรณ์ในแต่ละกรณีให้มีค่าไม่น้อยกว่า 75% ดังนั้นจากการอ่านผลพยากรณ์จาก Tree ที่สร้างขึ้นมานี้ จะสามารถนำเสนอแนวทางที่มีความน่าสนใจได้เพียงแค่กรณีเดียว สามารถสรุปได้ดังนี้

  - กรณีนี้ จะเลือก Attribute New Depart เป็นตัวหลักที่ใช้ในการแบ่งข้อมูล โดยเลือกค่าที่เป็น Dep1, Attribute Times เป็น Time1, Attribute Region เป็น South และ Attribute Method เป็น M5, M6, M9, Oth ซึ่งโมเดลนี้จะสามารถแปลความหมายได้ว่า ถ้าลูกหนี้ NPLs ที่อยู่ในกรณีนี้อยู่ในความดูแลของหน่วยงานที่ 1, เป็นลูกหนี้ที่ทำการปรับปรุงโครงสร้างหนี้เป็นครั้งแรก, ลูกหนี้อาศัยอยู่ในภาคใต้และใช้วิธีการปรับปรุงโครงสร้างหนี้แบบปรับหนี้สั้นเป็นหนี้ระยะยาว หรือให้ระยะเวลาปลอดหนี้ หรือสามารถรับโอนทรัพย์สินโดยมีสิทธิที่จะขอโอนกลับคืนได้ หรือใช้วิธีอื่นๆที่นอกเหนือจาก 9 วิธีที่ระบุไว้ จะพยากรณ์ว่าลูกหนี้ไม่สามารถชำระหนี้

ได้ตามเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้ได้สูงถึง 82.13% ของจำนวนข้อมูลใน Node นั้น (มีค่าความเสี่ยงในการพยากรณ์ผิดพลาด เท่ากับ 1.99%) ซึ่งสามารถเขียนในรูปของ If-Then Rule ได้ว่า If *NewDepart = Dep1 and Times = Time1 and Region = South and Method = M5, M6, M9, Oth* Then *Result = No* โดยจะสามารถแสดงผลได้ในรูปของ Tree ได้ดังรูปที่ 4.6



รูปที่ 4.6 แสดงรูป Tree ที่ได้ในกรณีที่ลูกหนี้ไม่สามารถทำตามเงื่อนไขได้

ซึ่งจากการที่เราได้นำข้อมูลของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาประมวลผลโดยใช้ CHAID Algorithm ผ่านโปรแกรมสำเร็จรูป Answer Tree นั้น ผลลัพธ์ที่ได้ออกมาจะแสดงให้เห็นว่า โปรแกรมสามารถสร้างโมเดลที่พยากรณ์ผลได้ถูกต้อง 67.11% หรือเป็นจำนวน 4,295 ราย ที่เหลืออีก 32.89% คิดเป็นจำนวน 2,105 ราย จะเป็นในส่วนที่พยากรณ์ไม่ถูกต้อง โดยจากจำนวนข้อมูลที่นำมาใช้ในการ Training Data จำนวน 6,400 ราย จะเป็นการทำนายผลว่าลูกหนี้สามารถปรับโครงสร้างหนี้ได้สำเร็จ (Attribute Result = Yes) จำนวน 1,271 ราย และลูกหนี้ไม่สามารถชำระหนี้ได้ตามสัญญาปรับปรุงโครงสร้างหนี้ได้ (Attribute Result = No) จำนวน 3,024 ราย

จากการนำข้อมูลที่ได้แบ่งไว้เพื่อมาทำการ Testing Data พบว่าสามารถทำนายผลการพยากรณ์ได้ 769 ราย หรือคิดเป็น 48.0625% และไม่สามารถพยากรณ์ผลได้ถูกต้องตามผลของข้อ

มูลที่นำมาทดสอบเป็นจำนวน 828 ราย หรือคิดเป็น 51.75% นอกจากนี้ยังมีในส่วนของข้อมูลที่ไม่สามารถพยากรณ์ผลลัพธ์ได้เลยแค่ 3 ราย คิดเป็น 0.1875%

ซึ่งจากการทดสอบนำข้อมูลกลุ่มตัวอย่างไปแทนค่าใน Tree ที่สร้างขึ้นเพื่อหาสาเหตุของความผิดพลาด โดยจะสรุปในส่วนของกลุ่มที่พยากรณ์ผิดพลาดและกลุ่มที่ไม่สามารถหาค่าของการพยากรณ์ได้ดังนี้

- สาเหตุของกลุ่มที่การพยากรณ์ผิดพลาด อาจจะมีสาเหตุเนื่องมาจากข้อมูลที่นำมาใช้ในการทดสอบโมเดลนั้นไม่ถูกต้อง ทำให้เมื่อนำมาทดสอบพยากรณ์ตาม Tree ที่สร้างขึ้น จึงให้ผลพยากรณ์ที่ผิดพลาด
- สาเหตุของกลุ่มที่ไม่สามารถหาค่าของการพยากรณ์ได้ อาจจะมีสาเหตุเนื่องมาจากข้อมูลที่เราแบ่งไว้เพื่อ Training Data มีค่าของข้อมูลใน Attribute ไม่ครอบคลุมทุกค่าที่มีอยู่ ดังนั้นถ้าหากข้อมูลที่เรานำมาทดสอบโมเดลมีค่านอกเหนือจากข้อมูลที่นำมาสร้างโมเดล ก็จะทำให้ไม่สามารถแสดงค่าของการพยากรณ์ได้

บทที่ 5
บทสรุป

ในบทนี้จะกล่าวสรุปถึงผลวิเคราะห์ที่ได้จากนำข้อมูลของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้มาผ่านกระบวนการในการสร้างโมเดล Tree Decision เพื่อทำนายผลการปรับโครงสร้างหนี้ของลูกหนี้ โดยทำการศึกษาจาก CHAID Algorithm ว่ามีลักษณะเป็นอย่างไร มีความถูกต้องและความน่าเชื่อถือเพียงใด มีปัญหาและข้อผิดพลาดในการศึกษาโครงการนี้หรือไม่ อย่างไร และสามารถนำผลที่ได้ไปใช้ประโยชน์ได้อย่างไร

5.1 บทสรุป

จากการศึกษาโครงการฉบับนี้ พบว่าผลลัพธ์ที่ได้จากการสร้างโมเดลโดยใช้ Algorithm ดังกล่าวมาข้างต้น โดยผ่านโปรแกรมสำเร็จรูป Answer Tree นั้น จะได้ผลลัพธ์ออกมา คือ สามารถใช้ทำนายผลลัพธ์ออกมาที่ความถูกต้องที่ประมาณ 40-50% แต่จะมีส่วนที่ผิดพลาดหลายจุด อาจเนื่องมาจากข้อมูลที่นำมาใช้ในการสร้างโมเดลอาจจะยังไม่ถูกต้อง เพราะอาจจะมีความผิดพลาดอยู่ในขั้นตอนของการบันทึกข้อมูล ทำให้ผลในการพยากรณ์ที่ได้จะมีความผิดพลาดอยู่มาก

แต่จะพบว่าปัจจัยที่มีผลต่อการปรับโครงสร้างหนี้ของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ว่าจะมีโอกาสสำเร็จมากน้อยแค่ไหนนั้น ที่เห็นได้เด่นชัดจากการทดลองคือ ปัจจัยทางด้านอัตราดอกเบี้ยที่กำหนดไว้ในสัญญาปรับปรุงโครงสร้างหนี้ (Attribute Int) โดยจะมีผลต่อการปรับโครงสร้างหนี้มาก นอกจากนี้ยังมีปัจจัยอื่นที่มีส่วนสำคัญอีก นั่นก็คือ หน่วยงานที่ดูแล (Attribute New Depart) ซึ่งก็จะมีส่วนในการทำให้การปรับโครงสร้างหนี้สำเร็จได้ ส่วนปัจจัยอื่นๆที่ได้นำมาใช้ในการทดลองครั้งนี้ จะไม่ค่อยมีผลต่อการปรับโครงสร้างหนี้ หรืออาจจะมีก็เพียงแค่ส่วนน้อย

5.2 ปัญหาและอุปสรรค

ปัญหาและอุปสรรคที่พบในการศึกษาโครงการนี้ คือ ข้อมูลที่นำมาใช้ในการทดลองนี้ อาจจะไม่มีความทันสมัยเพียงพอ เมื่อนำมาใช้กับสภาพเศรษฐกิจที่มีการเปลี่ยนแปลงตลอดเวลาในปัจจุบัน เช่น อัตราดอกเบี้ยที่นำมาใช้ ณ วันที่เก็บข้อมูลของการทำสัญญาปรับปรุงโครงสร้างหนี้

นั้น อาจจะไม่เท่ากับอัตราดอกเบี้ยในปัจจุบันเนื่องจากสภาพเศรษฐกิจมีการเปลี่ยนแปลงตลอดเวลา ทำให้การนำผลการพยากรณ์ที่ได้มาใช้ในการวางกลยุทธ์อาจให้ผลที่มีประสิทธิภาพได้ไม่เต็มที่ นอกจากนี้ปัจจัยต่างๆของข้อมูลที่คัดเลือกมาใช้ในการสร้างโมเดล อาจจะไม่เหมาะสม เนื่องจากการเลือกปัจจัยที่ไม่มีผลต่อการปรับปรุงโครงสร้างหนี้มาใช้ในการสร้างโมเดล ทำให้ผลที่ได้มีข้อผิดพลาดให้เห็นได้ หรืออาจจะยังขาดข้อมูลจากแหล่งอื่นๆที่ไม่สามารถเปิดเผยได้ ทำให้ผลการพยากรณ์ยังมีความถูกต้อง น่าเชื่อถืออยู่ในระดับที่ไม่สูงมากนัก อีกทั้งโปรแกรมสำเร็จรูปที่นำมาศึกษาในครั้งนี้เป็นโปรแกรมที่เป็น Freeware จึงทำให้ฟังก์ชั่นต่างๆในการกำหนดค่าเงื่อนไขหรือกฎเกณฑ์ต่างๆ อาจไม่มีความละเอียดเพียงพอ ทำให้ประสิทธิภาพของการใช้งานโปรแกรมยังไม่เต็มที่

5.3 แนวทางการนำไปใช้ประโยชน์

ผลลัพธ์ที่ได้จากการศึกษาโครงการฉบับนี้ จะสามารถนำมาใช้ประโยชน์ในการพยากรณ์แนวโน้มของลูกหนี้ที่อยู่ในกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ว่ามีโอกาสปรับปรุงโครงสร้างหนี้ได้สำเร็จมากน้อยแค่ไหน ทำให้สามารถจัดกลุ่มของลูกหนี้ได้ว่ากลุ่มไหนต้องดูแลอย่างใกล้ชิด, กลุ่มไหนอยู่ในกลุ่มที่มีความเสี่ยงน้อยในการปรับโครงสร้างหนี้ไม่สำเร็จ เป็นต้น ซึ่งหากผลการทำนายบอกว่าโอกาสปรับสำเร็จมีน้อย ก็จะต้องให้การดูแลเอาใจใส่เป็นพิเศษ เพื่อจะทำให้ลูกหนี้ในกลุ่มนี้สามารถกลับมาชำระจนกลายเป็นลูกหนี้ชั้นดีได้ เพื่อประโยชน์ขององค์กรต่อไป อีกทั้งยังสามารถทำการประมาณการล่วงหน้าถึงเป้าหมายที่จะต้องกำหนดไว้ในแผนการทำงานขององค์กรได้อย่างมีประสิทธิภาพ

นอกจากนี้ยังสามารนำผลที่ได้มาประกอบกับการวางแผนกลยุทธ์ในการปรับโครงสร้างหนี้ได้ คือ เราอาจจะนำเอาเงื่อนไขที่น่าสนใจที่ได้จากการอ่านค่า Tree มาประกอบกับการกำหนดเงื่อนไขในสัญญาปรับปรุงโครงสร้างหนี้ให้มีความเหมาะสมกับกลุ่มลูกหนี้แต่ละกลุ่ม เพื่อเพิ่มโอกาสในความสำเร็จของการปรับปรุงโครงสร้างหนี้ เช่น อาจทำการเลือกเอาค่าของอัตราดอกเบี้ยที่ระบุในเงื่อนไขที่มีโอกาสทำสำเร็จสูงในโมเดลมาใช้ในการตั้งเงื่อนไขในการเซ็นสัญญาปรับโครงสร้างหนี้ของลูกหนี้ หรืออาจจะเลือกให้หน่วยงานที่มีประสิทธิภาพในด้านนี้เป็นผู้ดูแลลูกหนี้ในกลุ่มที่ต้องดูแลเป็นพิเศษ โดยเลือกหน่วยงานมาจากเงื่อนไขที่มีโอกาสทำสำเร็จสูงในโมเดล เป็นต้น แต่ทั้งนี้ต้องดูประกอบกับข้อมูลในส่วนอื่นๆด้วย

ดังนั้น การศึกษาโครงการฉบับนี้ เป็นเพียงการทดลองเบื้องต้นในการพยากรณ์ผลการปรับปรุงโครงสร้างหนี้ของกลุ่มข้อมูลของหนี้ที่ไม่ก่อให้เกิดรายได้ โดยผลที่ได้อาจจะยังมีความถูกต้องอยู่ในระดับที่ไม่สามารถนำไปใช้ในการพยากรณ์ได้อย่างแม่นยำ แต่ก็สามารถใช้เป็นแนวทางที่น่า

สนใจในการทำการศึกษาและพัฒนาต่อไป เพื่อใช้ในการกำหนดกลยุทธ์ต่างๆ เพื่อแก้ไขปัญหาของกลุ่มหนี้ที่ไม่ก่อให้เกิดรายได้ให้ลดน้อยลง ซึ่งหากสามารถทำได้ก็จะมีประโยชน์ต่อองค์กรและภาพรวมของสภาพเศรษฐกิจของประเทศต่อไป

บรรณานุกรม

Agrawal, R. and Srikant, R. 1994. **Fast Algorithms for Mining Association Rules.** [Online] Available : Http://citeseer.nj.nec.com

Berry , M. J. and Linoff, G. 1997. **Data Mining Techniques For Marketing , Sales and Customer Support.** Wiley Computer and Son Inc.

Han, J. and Kamber, M. 2003. **Data Mining : Concepts and Techniques.** [Online]. Available: Http://www.cs.cfu.ca

Kusiak, A. **Association Rules : The Apriori Algorithm.** [Online] . Available : Http://www.icaen.uiowa.edu

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

DBMS, Data Mining Solutions Supplement. **Association and Sequencing.** [Online]. Available : Http://www.dbmsmag.com

ภาคผนวก ก

ขั้นตอนการใช้งานโปรแกรมสำเร็จรูป Answer Tree

ขั้นตอนการใช้งานโปรแกรมสำเร็จรูป Answer Tree

การใช้งานโปรแกรมสำเร็จรูป Answer Tree ในการสร้าง Tree จะมีขั้นตอนในการทำงานต่างๆ ดังต่อไปนี้

ขั้นตอนที่ 1 ต้องสร้างโครงการใหม่ โดยการเลือกไปที่คำสั่ง File แล้วเลือกที่ New Project ดังตัวอย่างในรูปที่ ก.1

ขั้นตอนที่ 2 นำเข้าข้อมูลที่เตรียมไว้เพื่อใช้ในการสร้างโมเดล ซึ่งโปรแกรมสามารถรับข้อมูลเข้าได้ 3 ประเภท โดยเลือกนำเข้าข้อมูลที่เป็นที่เป็น Excel ดังตัวอย่างในรูปที่ ก.2

ขั้นตอนที่ 3 จะมีหน้าจอให้เลือกไฟล์ข้อมูลที่ต้องการนำเข้ามาใช้งาน โดยเลือกไปที่ชื่อไฟล์ที่ได้เตรียมไว้ พร้อมทั้งชื่อ Sheet ที่เราต้องการ ดังตัวอย่างรูปที่ ก.3 และ ก.4 ตามลำดับ

ขั้นตอนที่ 4 เป็นการเลือก Algorithm ที่จะใช้ในการทำ Data Mining โดยในโปรแกรมนี้จะมี Algorithm ให้เลือกใช้อยู่ 4 Algorithm ให้เราเลือกใช้ CHAID Algorithm มาใช้ในการสร้าง Tree ดังตัวอย่างในรูปที่ ก.5

ขั้นตอนที่ 5 จะเป็นการกำหนด Target Attribute ที่เป็นเป้าหมายของโมเดล ซึ่งจะใช้ Attribute Result เป็น Target Attribute และ Attribute อื่นๆก็จะเป็นตัวแปร (Variable) ที่ใช้ในการสร้าง Tree ดังตัวอย่างในรูปที่ ก.6

ขั้นตอนที่ 6 เป็นการกำหนด Validation ของข้อมูล คือ การทดสอบโมเดลที่เราได้สร้างขึ้นมา โดยการแบ่งข้อมูลออกเป็น 2 ส่วนคือ ส่วนแรกจะใช้ในการ Training ข้อมูลเพื่อให้โปรแกรมได้ทำการเรียนรู้ความสัมพันธ์ของกลุ่มข้อมูลเพื่อสร้างโมเดลออกมา และส่วนที่สองจะใช้ในการ Testing โมเดลที่ได้สร้างขึ้นว่ามีความถูกต้อง เป็นไปในทิศทางเดียวกันหรือไม่ โดยสามารถกำหนดเป็นเปอร์เซ็นต์ที่ต้องการได้ ดังตัวอย่างในรูปที่ ก.7

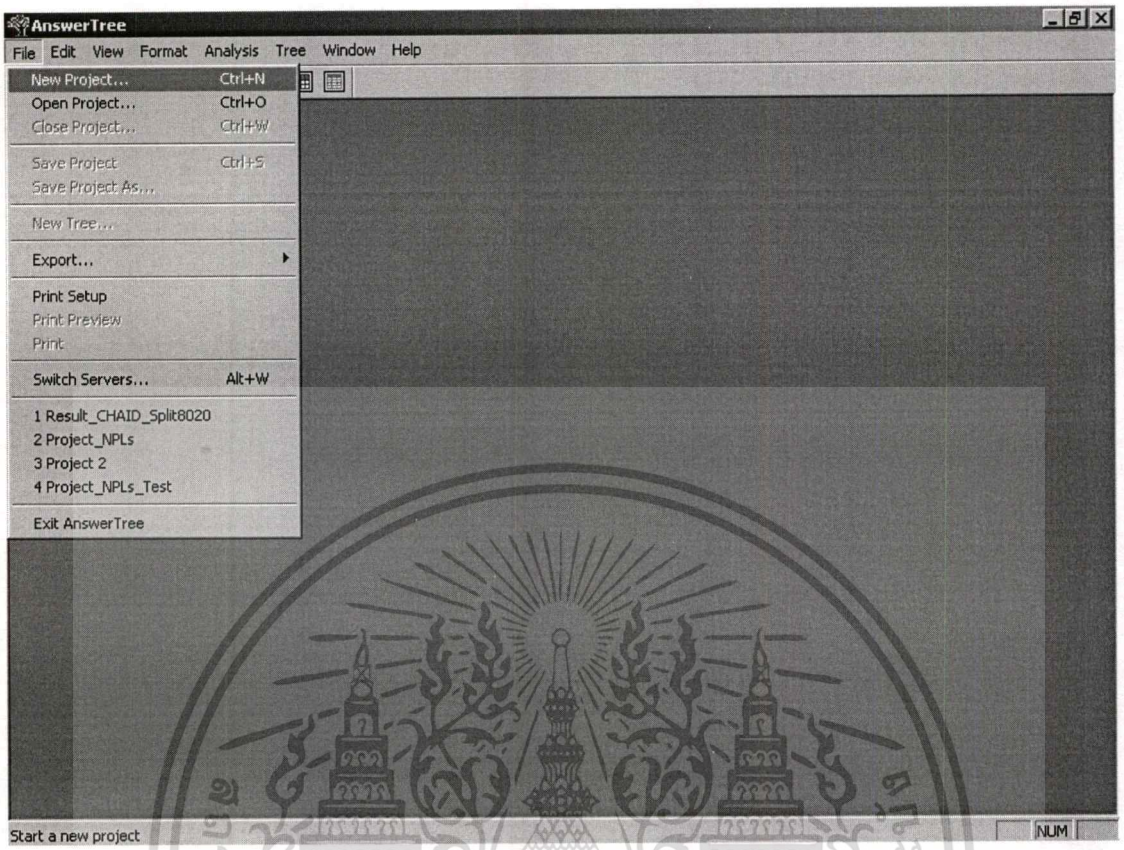
ขั้นตอนที่ 7 จะเป็นการกำหนด Option ต่างๆของโปรแกรม Answer Tree ในการกำหนดเงื่อนไขของการแตกกิ่งของ Tree ที่เราสร้างขึ้น ว่าจะให้แตกกิ่งกี่ระดับ ซึ่งจะรวมถึงการกำหนด Parent Node และ Child node ด้วย ดังตัวอย่างในรูปที่ ก.8 และ ก.9 ตามลำดับ

ขั้นตอนที่ 8 โปรแกรมก็จะทำการประมวลผลเพื่อหาค่าความสัมพันธ์ของกลุ่มข้อมูลโดยใช้ Rule ต่างๆที่กำหนดไว้มาสร้าง Tree ซึ่งเมื่อดำเนินการเสร็จแล้วก็จะได้ Root Node ออกมาเพียง Node เดียว ดังตัวอย่างในรูปที่ ก.10

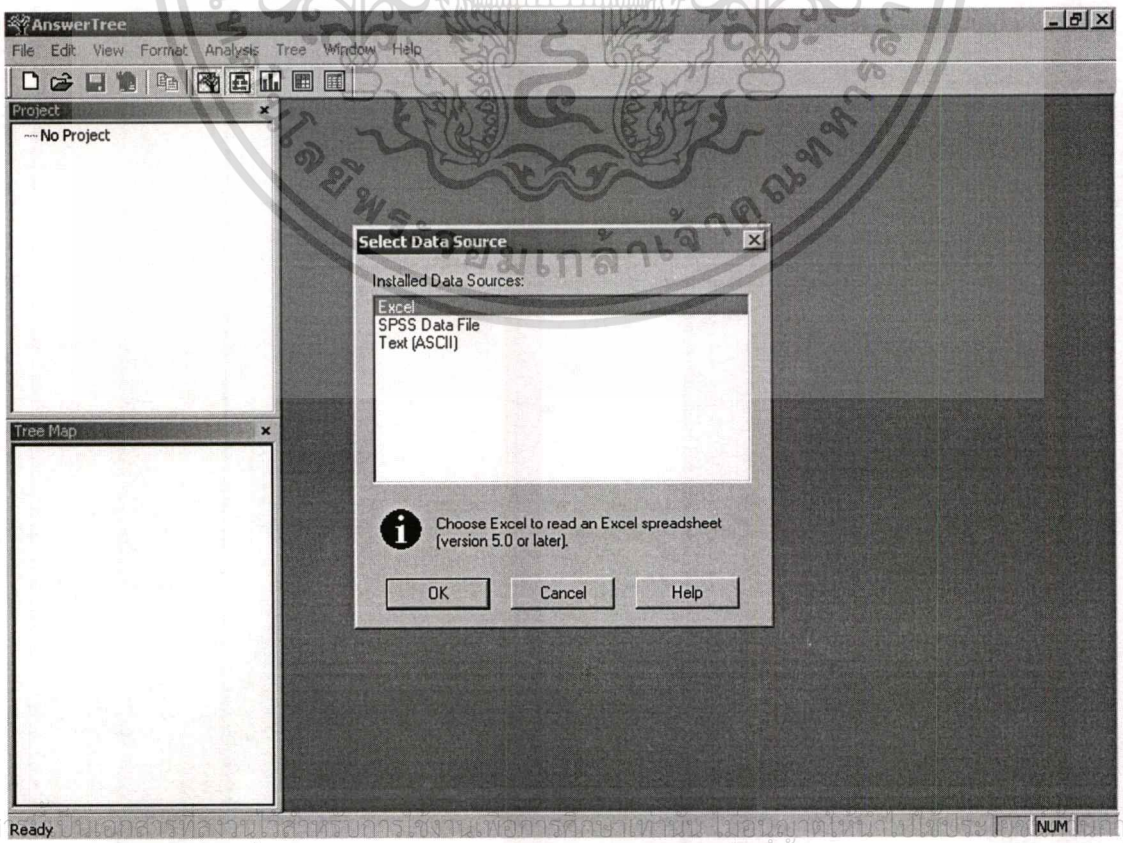
ขั้นตอนที่ 9 ต้องทำการ Grow Tree เพื่อที่จะแตกกิ่งของ Tree ออกมาให้เสร็จสมบูรณ์ตามที่เราได้กำหนดค่า โดยเลือกไปที่คำสั่ง Tree แล้วเลือกไปที่ Grow Tree ดังตัวอย่างในรูปที่ ก.11

ขั้นตอนที่ 10 โปรแกรมจะสร้าง Tree ที่เสร็จสมบูรณ์ออกมาดังตัวอย่างในรูปที่ ก.12

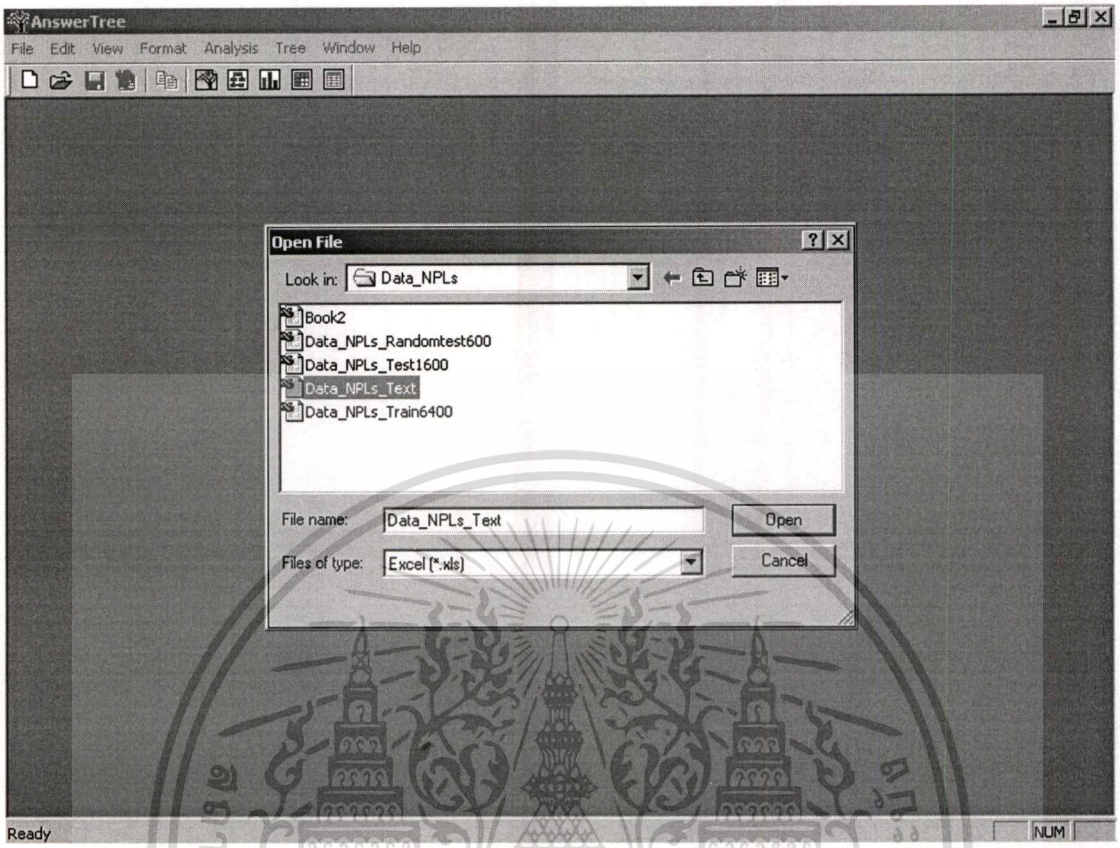
ขั้นตอนที่ 11 จาก Tree ที่เสร็จสมบูรณ์ โปรแกรมสามารถแสดงข้อมูลที่ได้จากการสร้าง Tree และข้อมูลที่แสดงค่าความเสี่ยงในการพยากรณ์ผิดพลาดของโมเดล ดังตัวอย่างในรูปที่ ก.13 และ ก.14 ตามลำดับ



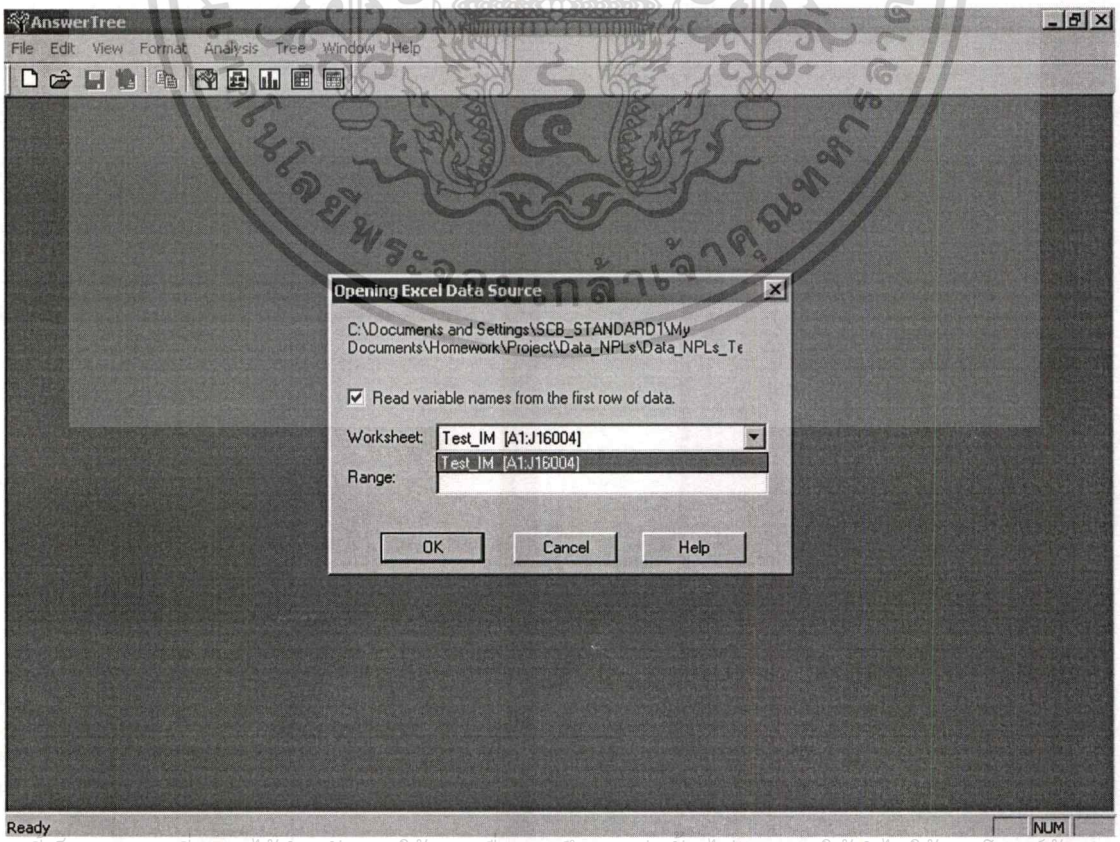
รูปที่ ก.1 แสดงหน้าจอการสร้าง New Project



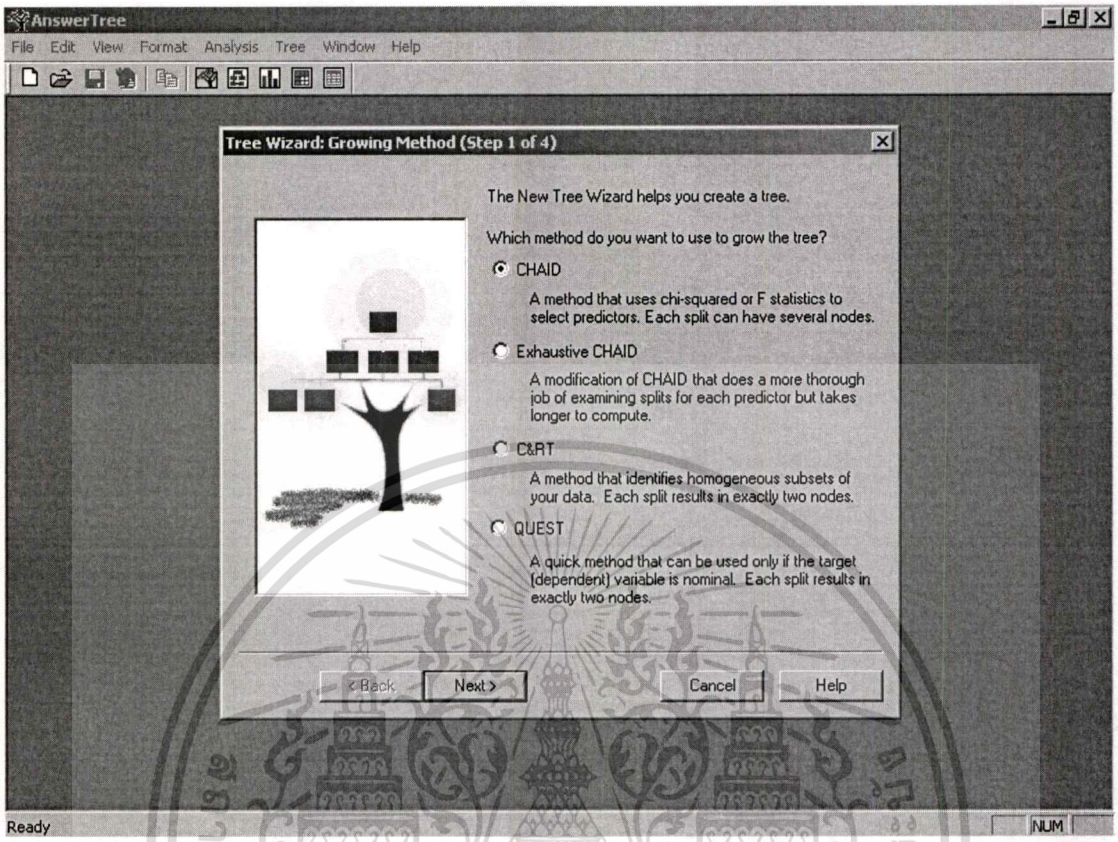
รูปที่ ก.2 แสดงหน้าจอของประเภทของข้อมูลที่สามารถนำเข้าได้



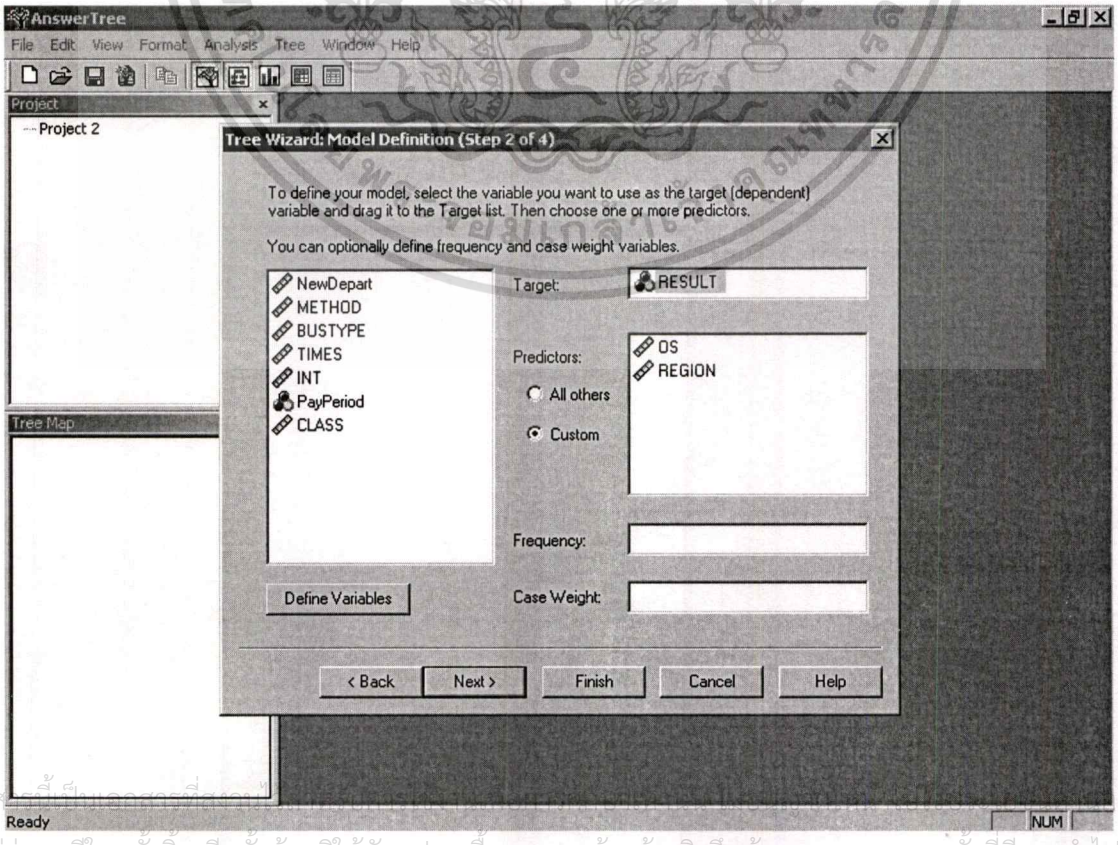
รูปที่ ก.3 แสดงหน้าจอการเลือกไฟล์ที่จะนำเข้าโปรแกรม



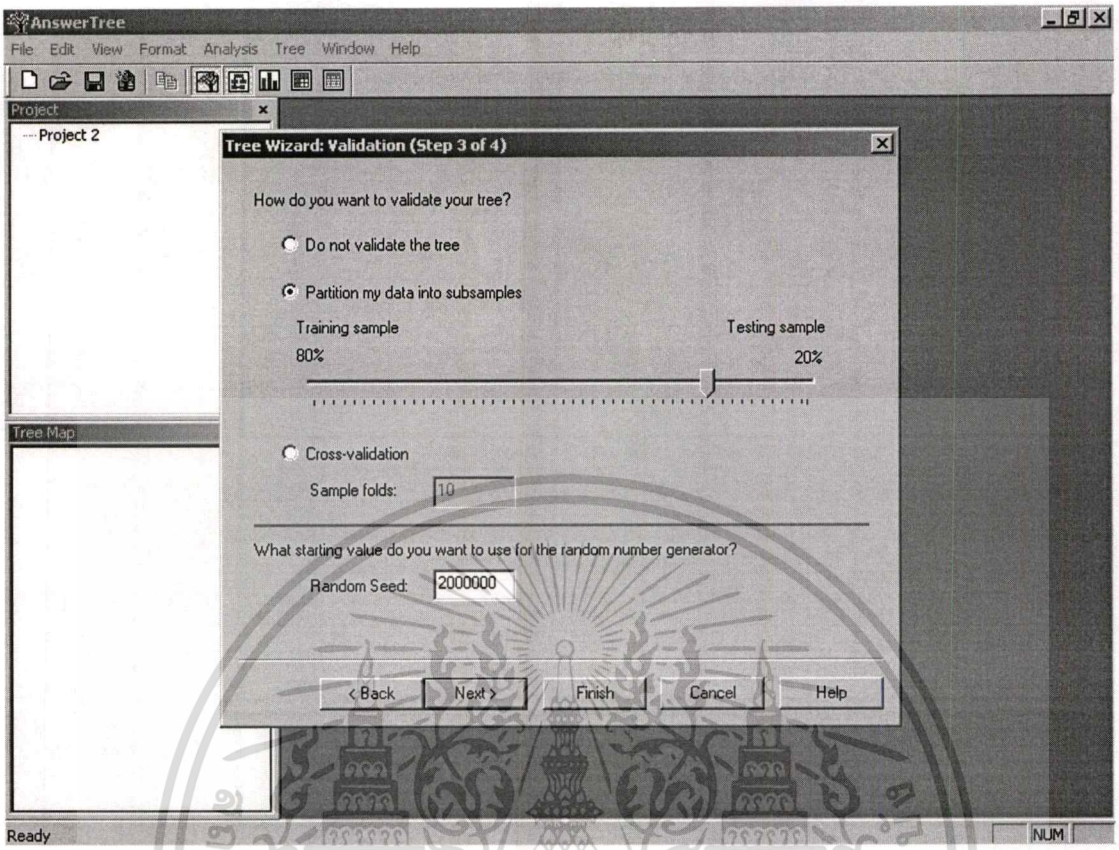
รูปที่ ก.4 แสดงหน้าจอการเลือก Sheet ข้อมูลที่ต้องการใช้



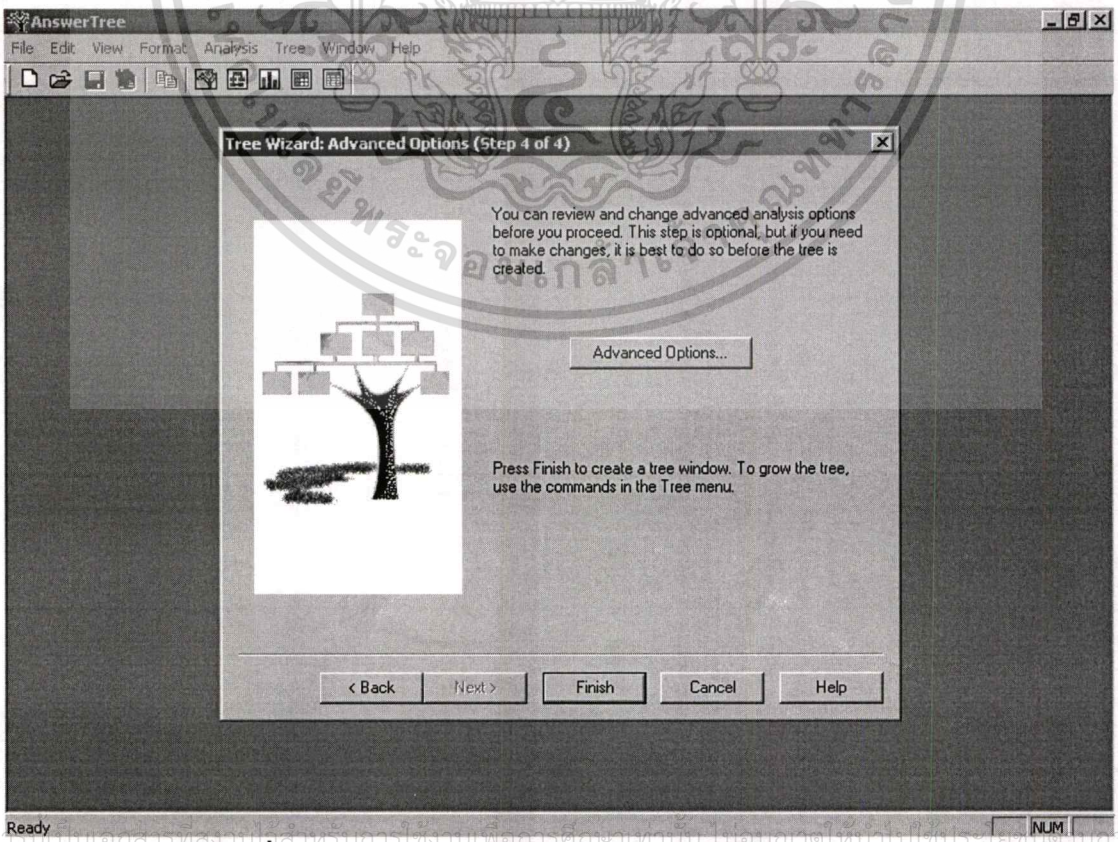
รูปที่ ก.5 แสดงหน้าจอการเลือก Algorithm ที่ใช้สร้างโมเดล



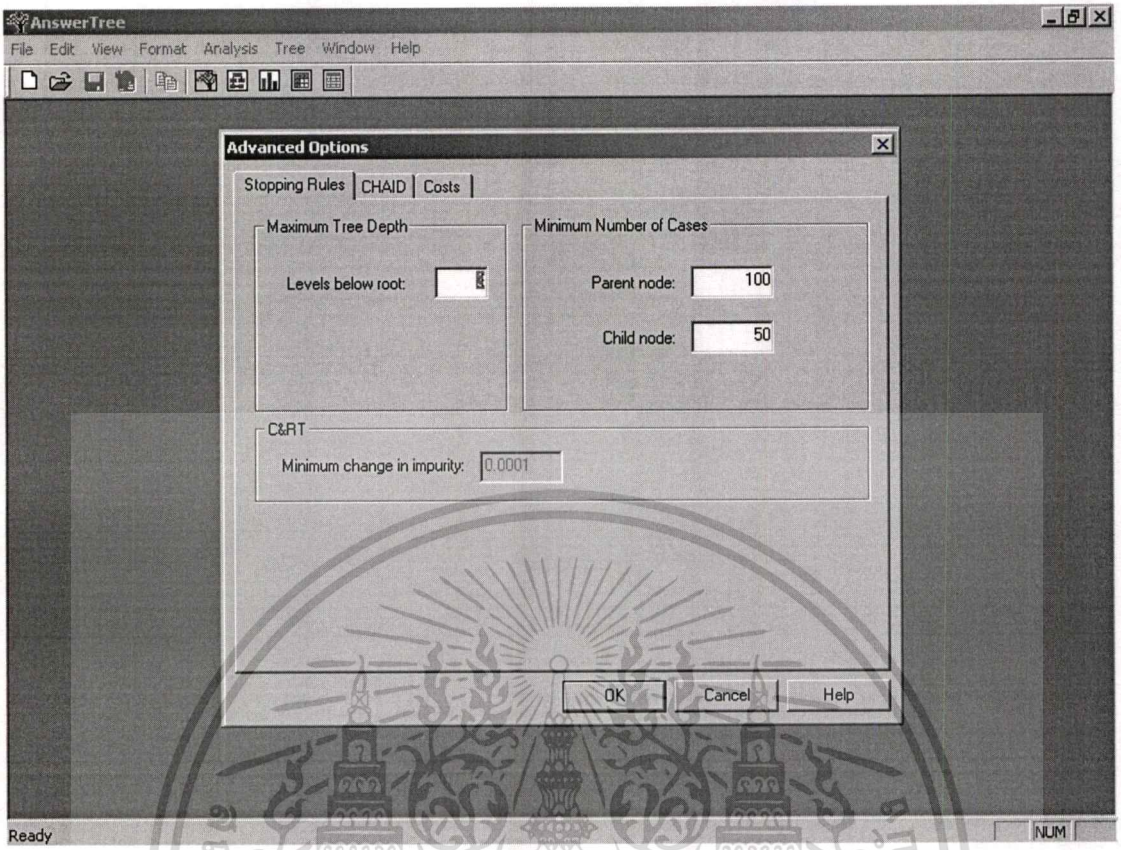
รูปที่ ก.6 แสดงหน้าจอการเลือก Target Attribute และ Variable ที่นำมาใช้



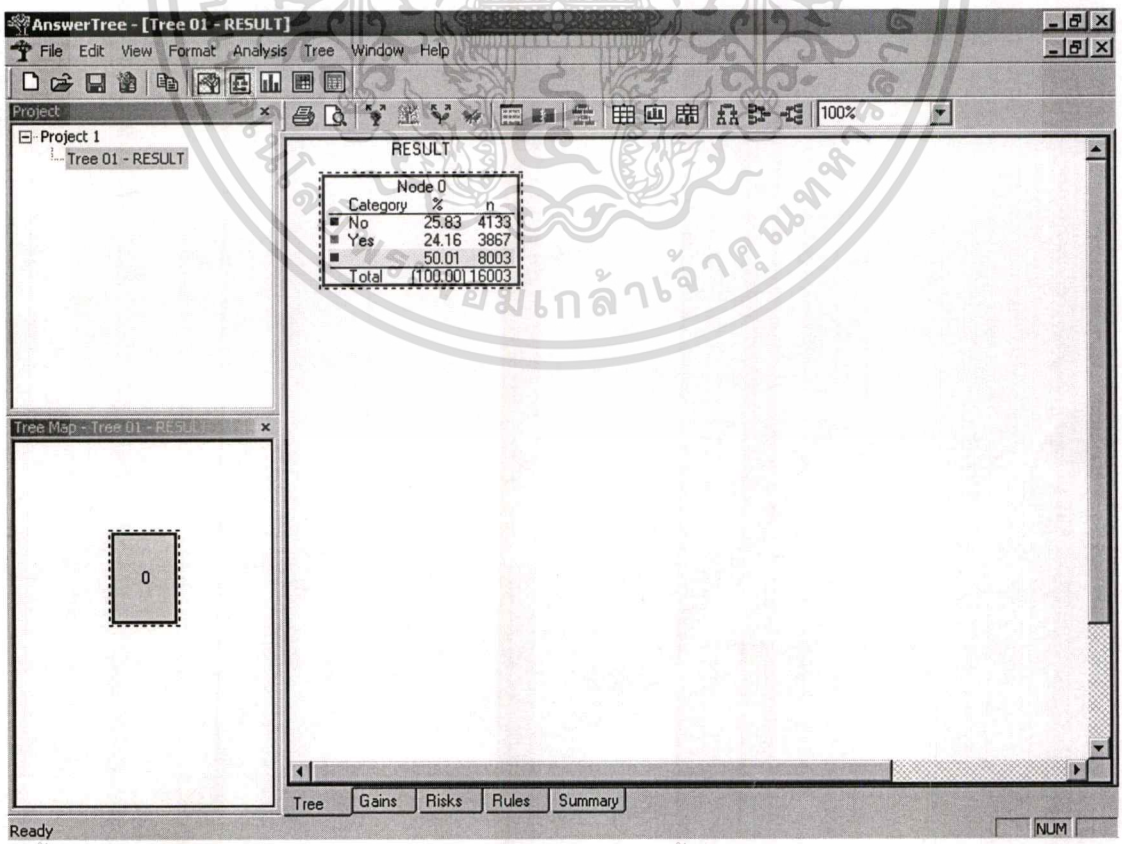
รูปที่ ก.7 แสดงหน้าจอการกำหนด Validation ของข้อมูล



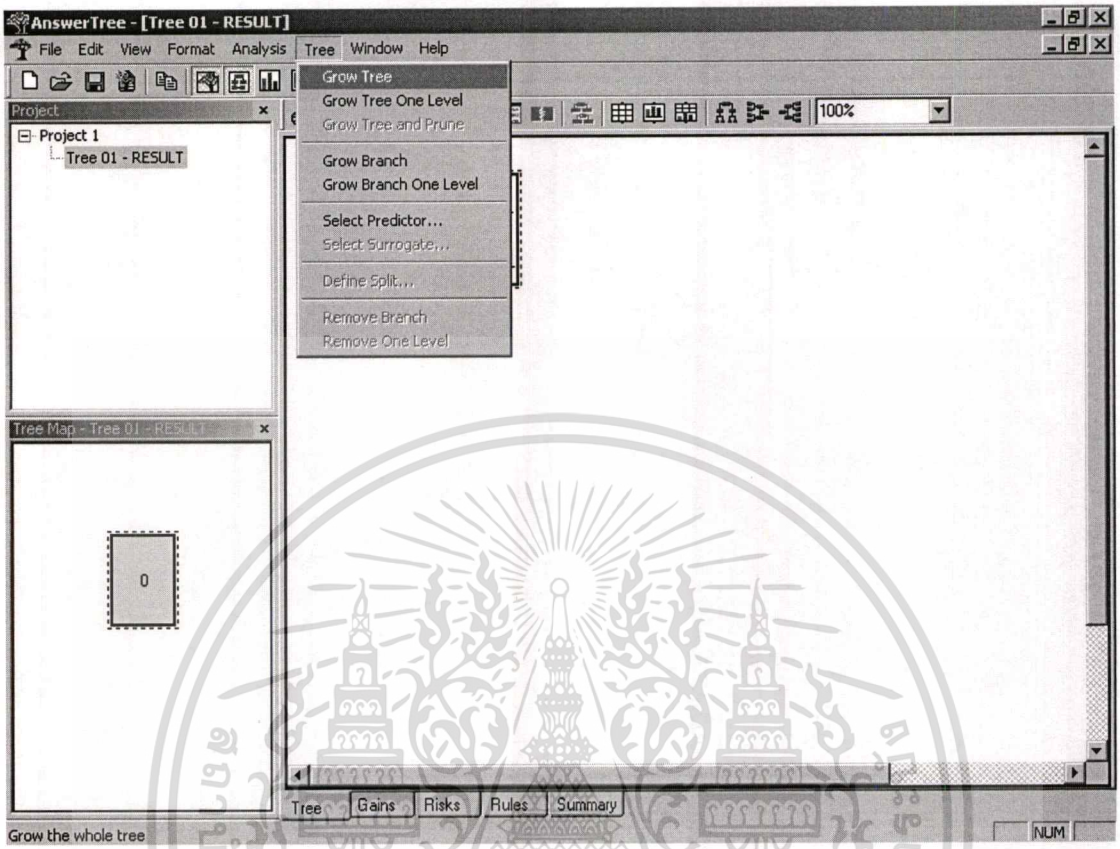
รูปที่ ก.8 แสดงหน้าจอการเลือก Options ในการสร้างโมเดล



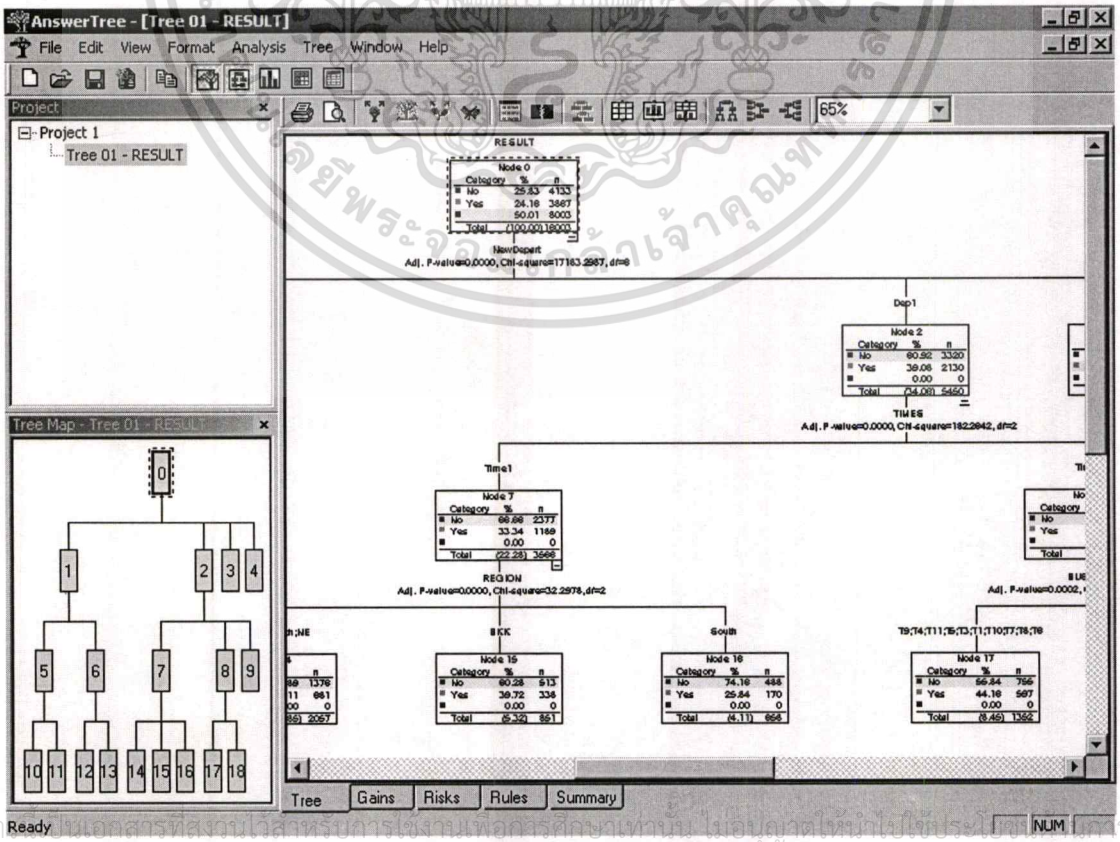
รูปที่ ก.9 แสดงหน้าจอการกำหนด Options ในการสร้างโมเดล



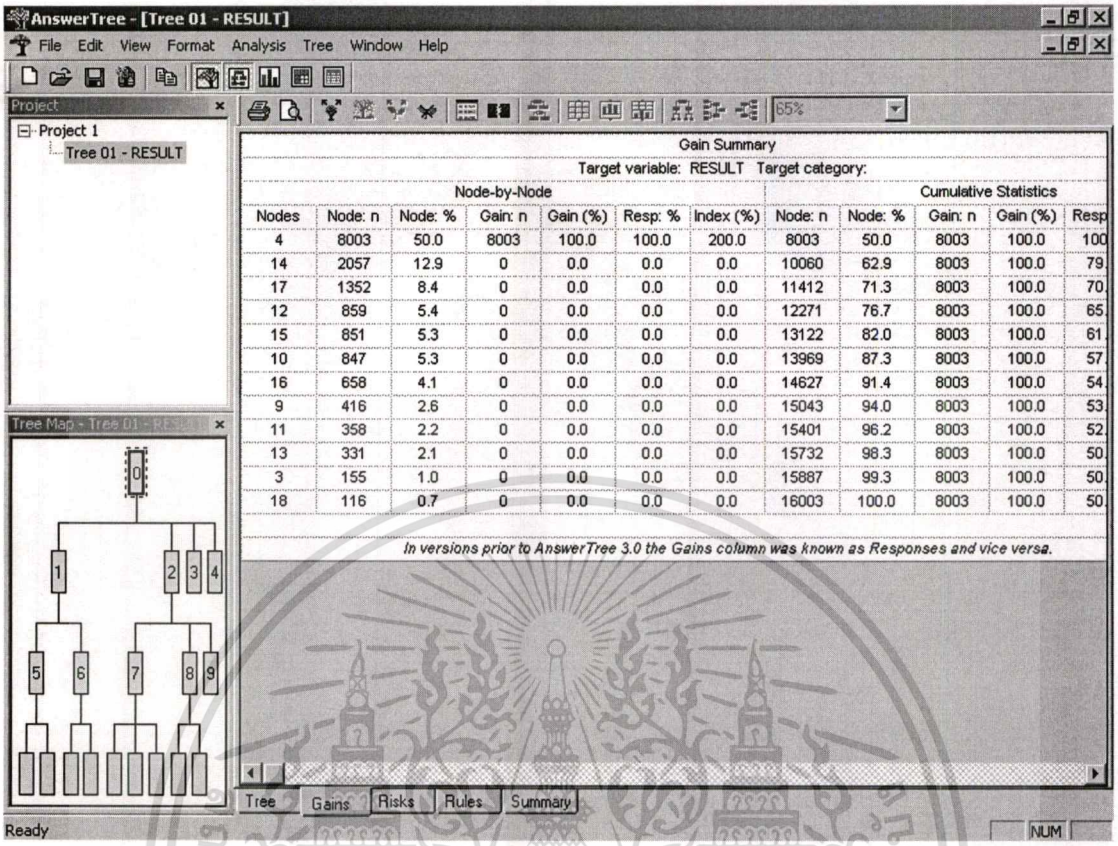
รูปที่ ก.10 แสดงหน้าจอของผล Tree ที่แสดง Root Node



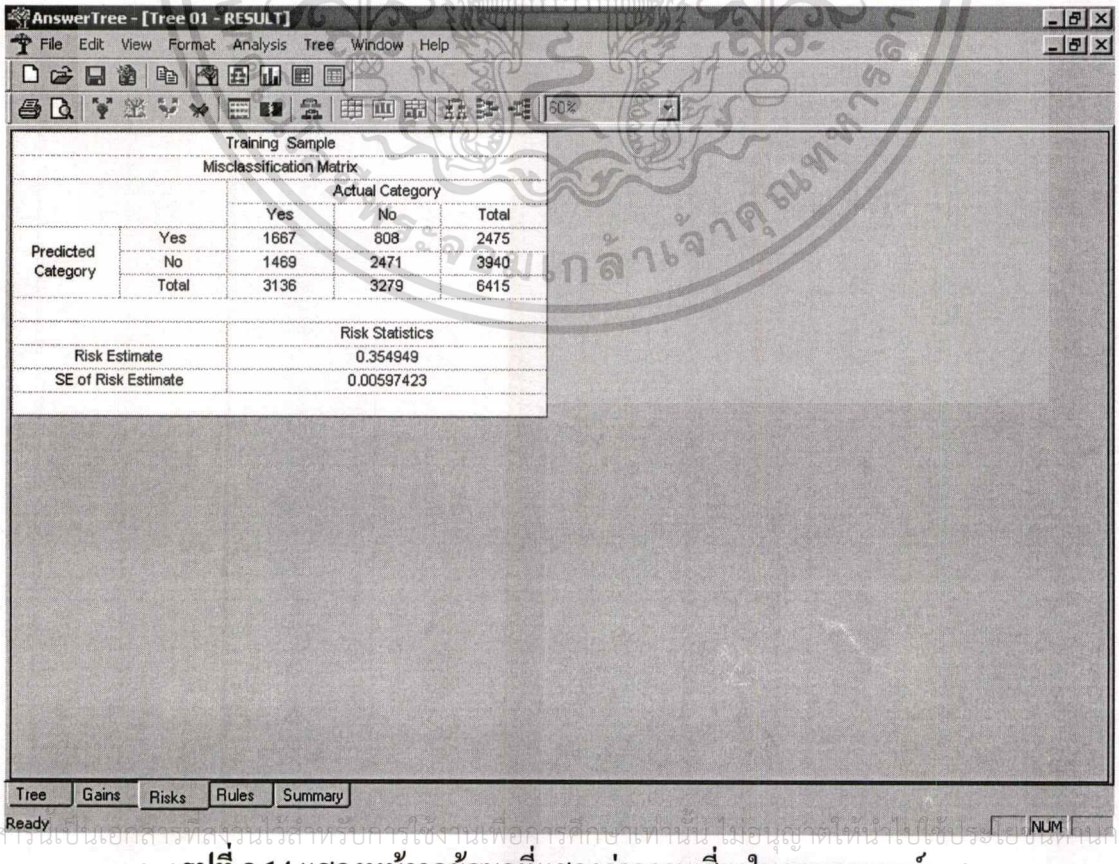
รูปที่ ก.11 แสดงหน้าจอการ Grow Tree



รูปที่ ก.12 แสดงหน้าจอของ Tree ที่เสร็จสมบูรณ์



รูปที่ ก.13 แสดงหน้าจอของข้อมูลที่ได้จากการสร้าง Tree



รูปที่ ก.14 แสดงหน้าจอข้อมูลที่แสดงค่าความเสี่ยงในการพยากรณ์

ประวัติผู้เขียน

นายธันยวีร์ สถิรวรกุล เกิดวันที่ 15 กรกฎาคม พ.ศ.2521 สำเร็จการศึกษาบริหารธุรกิจบัณฑิต จากคณะวิทยาการจัดการ มหาวิทยาลัยสงขลานครินทร์ ในปีการศึกษา 2542 และศึกษาต่อในระดับปริญญามหาบัณฑิต สาขาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในปีการศึกษา 2547 ปัจจุบันทำงานในตำแหน่งเจ้าหน้าที่วางแผน วางแผน กลุ่มจัดการทรัพย์สิน ให้กับ ธนาคารไทยพาณิชย์ จำกัด (มหาชน)