การพัฒนาระบบเพื่อวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย

Cause of Churning in Insurance Business Analysis

โดย

นางสาวธนวันต์ นามวงษ์

รหัส 45066090

อาจารย์ที่ปรึกษา

ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2546
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบเพื่อวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย
นักศึกษา	นางสาวธนวันต์ นามวงษ์
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

ปัจจุบันนี้มีการนำเทคนิคต่าง ๆ ของ Data Mining มาประยุกต์ใช้ในการวิเคราะห์พฤติกรรมของลูกค้า เพื่อให้สามารถรองรับความต้องการของลูกค้าได้อย่างมีประสิทธิภาพ โครงงานนี้เป็นการพัฒนาโปรแกรมสำเร็จรูปเพื่อใช้ในการวิเคราะห์สาเหตุการยกเลิกบริการในธุรกิจประกันภัย โดยใช้อัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมหนึ่งในเทคนิค Classification แบบ Decision tree เป็นพื้นฐานในการพัฒนาโปรแกรม ผลการวิเคราะห์ที่ได้จากการใช้งานโปรแกรมจะอยู่ในรูปของกฎที่จำแนกลักษณะของลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการ เพื่อใช้เป็นแนวทางในการพัฒนากลยุทธ์ทางการตลาด เช่น การวางแผนการโฆษณาและการจัดทำโปรโมชั่นที่น่าสนใจเพื่อรักษากลุ่มลูกค้าดังกล่าวไว้

Title	Cause of Churning in Insurance Business Analysis
Student	Ms. Thanawan Namwong
Advisor	Asst. Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Acadamic Year	2003

ABSTRACT

Presently, the data mining technics has applied to analyse customer behavior to react appropriately and efficiently to customer demand. The purpose of project is to develop an application program to analyse cause of churning in insurance business. C4.5 Algorithm, one of the algorithms in Data Mining classification, was used in this project.

The classification tree and rule are the result of the application, can be used to analyse cause of churning in insurance business and identify customer behaviour who will churn in order to improve the marketing strategy for acquistion and retention customer to continue the service.

กิตติกรรมประกาศ

ในโครงงานฉบับนี้ ที่ผู้จัดทำสามารถพัฒนาจนสำเร็จลุล่วงไปด้วยดีนั้น เนื่องด้วยได้รับคำแนะนำ และความช่วยเหลือ รวมถึงกำลังใจที่ดี ทั้งนี้ผู้จัดทำขอกล่าวคำขอบคุณกับบุคคลและกลุ่มบุคคลต่างๆ ดังนี้

1. บิดามารดา ที่เป็นผู้ให้ทุกอย่าง รวมถึงกำลังใจที่ดีเสมอมา
2. ผศ.ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงาน ขอบคุณสำหรับคำแนะนำ ข้อเสนอแนะ และการสนับสนุนการทำโครงงานฉบับนี้ตั้งแต่เริ่มต้นจนสำเร็จ
3. พี่ๆ เพื่อนๆ น้องๆ ทุกคน ขอขอบคุณในการช่วยเหลือทุกสิ่ง ไม่ว่าจะเป็นด้านฮาร์ดแวร์ ซอฟต์แวร์ ด้านข้อมูล หนังสืออ้างอิง คำแนะนำ ความช่วยเหลือในทุกๆ ด้าน และกำลังใจ รวมถึงการดูแลต่างๆ ที่มีให้ตลอดเวลา

ด้วยความขอบคุณเป็นอย่างสูง

ธนวันต์ นามวงษ์
ผู้จัดทำ

สารบัญ

สารบัญ(ต่อ)

สารบัญตาราง

หน้า

ตารางที่

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมา

การดำเนินธุรกิจประกันภัยในปัจจุบันนี้มีการแข่งขันกันสูงมากขึ้นเรื่อยๆ ผู้ประกอบการจึงต้องพยายามหากลยุทธ์และวิธีการต่าง ๆ มาใช้เพื่อให้สามารถดำเนินกิจการให้ประสบความสำเร็จเหนือคู่แข่ง กลยุทธ์การรักษาลูกค้าถือเป็นแนวทางที่สำคัญแนวทางหนึ่งที่ผู้ประกอบการธุรกิจประกันภัยหันมาให้ความสำคัญ เนื่องจากการที่ลูกค้าเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น (Churn) เป็นสาเหตุทำให้บริษัทสูญเสียรายได้ รวมทั้งค่าใช้จ่ายในการหาลูกค้าใหม่นั้นสูงกว่าค่าใช้จ่ายที่ใช้ในการรักษาลูกค้าเดิมไว้มาก จึงได้มีการนำเอาเทคนิคของดาต้าไมนิ่ง (Data Mining) มาประยุกต์ใช้เพื่อช่วยในการวิเคราะห์ข้อมูลจำนวนมหาศาลที่มีอยู่ในคลังข้อมูล เพื่อระบุถึงสาเหตุและกลุ่มของลูกค้าที่มีแนวโน้มที่จะเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น เพื่อใช้เป็นแนวทางในการวางแผนการโฆษณาและจัดทำโปรโมชั่นที่น่าสนใจเพื่อรักษาลูกค้าไว้ให้ใช้บริการต่อไป

1.2 วัตถุประสงค์ของการศึกษา

วัตถุประสงค์ของการศึกษาและพัฒนาระบบงานนี้เพื่อนำเอาเทคนิคของดาต้าไมนิ่งมาประยุกต์ใช้ในการวิเคราะห์ลักษณะและพฤติกรรมของลูกค้า เพื่อระบุถึงสาเหตุของการยกเลิกการใช้บริการและทำนายว่าลูกค้ากลุ่มใดที่มีแนวโน้มจะยกเลิกการใช้บริการ ทั้งนี้เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้เป็นแนวทางในการวางกลยุทธ์ทางการตลาดและจัดทำโปรโมชั่นที่น่าสนใจเพื่อรักษาลูกค้ากลุ่มดังกล่าวไว้

1.3 ขอบเขตการดำเนินงาน

โครงการนี้เป็นการศึกษาถึงการนำเอาเทคนิคของดาต้าไมนิ่งมาประยุกต์ใช้ โดยอาศัยหลักการของอัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมหนึ่งใน Classification ในการจัดกลุ่มลูกค้า โดยจะนำเสนอผลลัพธ์ในรูปของกฎและ Decision Tree เพื่อนำมาวิเคราะห์ลักษณะของลูกค้าที่ยกเลิกการใช้บริการในธุรกิจประกันภัย

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังต่อไปนี้

1. กำหนดวัตถุประสงค์และเป้าหมายของการดำเนินงาน
2. ศึกษาและรวบรวมข้อมูล โดยจะใช้ข้อมูลในการทำประกันภัยของลูกค้า
3. ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องของการทำดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้
4. ศึกษาอัลกอริทึม C4.5 เพื่อนำมาประยุกต์ใช้กับระบบงาน
5. ออกแบบและพัฒนาโปรแกรมเพื่อใช้ในการแบ่งกลุ่มลูกค้า
6. วิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง
7. สรุปผลการศึกษา และพิจารณาแนวทางการปรับปรุงระบบให้ดีขึ้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาทฤษฎีและเทคนิคของการทำดาต้าไมนิ่งและพัฒนาระบบงานเพื่อวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย คาดว่าจะได้รับประโยชน์ดังนี้

1. เข้าใจหลักการและขั้นตอนของการทำดาต้าไมนิ่ง รวมทั้งอัลกอริทึม C4.5 ที่ใช้ในการจัดกลุ่มข้อมูล
2. เพื่อเป็นแนวทางในการนำดาต้าไมนิ่งมาประยุกต์ใช้กับข้อมูลทางธุรกิจ
3. สามารถนำข้อมูลที่ได้จากการทำดาต้าไมนิ่ง มาใช้เป็นแนวทางในการวางแผนการตลาดเพื่อรักษากลุ่มลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการ
4. เพื่อเป็นแนวทางในการออกแบบและพัฒนาโปรแกรมวิเคราะห์ข้อมูลโดยใช้วิธีการอื่นๆ ต่อไป

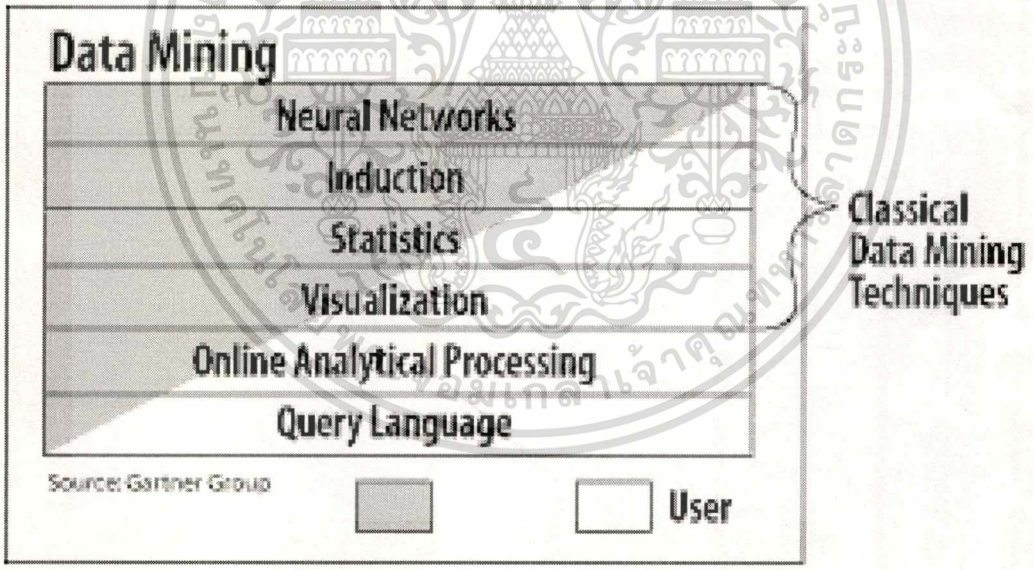
ในบทนี้เป็นการกล่าวถึงวัตถุประสงค์และขอบเขตของการทำงานในเบื้องต้นของระบบ ในบทต่อไปจะกล่าวถึงรายละเอียดของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

2.1 ความหมายของดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่งจัดเป็น Business Intelligent technology ประเภทหนึ่ง ซึ่งเป็นการวิเคราะห์ข้อมูลเพื่อหาแนวโน้มและความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูล เพื่อให้ได้สารสนเทศที่เป็นประโยชน์ เพื่อนำไปเป็นแนวทางในการตัดสินใจใด ๆ ที่ก่อให้เกิดประโยชน์ในทางธุรกิจ ซึ่งถือว่าเป็นจุดประสงค์หลักของการทำดาต้าไมนิ่ง

เทคนิคต่าง ๆ ของดาต้าไมนิ่งนั้นมีมากมาย โดยที่แต่ละเทคนิคจะใช้สำหรับวัตถุประสงค์ที่แตกต่างกันไปขึ้นอยู่กับปริมาณข้อมูลและการมีส่วนร่วมในการแก้ปัญหาของผู้ใช้งาน ดังรูป



รูปที่ 2.1 แสดงสัดส่วนระหว่างเทคนิคต่างๆ ของดาต้าไมนิ่ง
กับการมีส่วนร่วมในการแก้ปัญหาของผู้ใช้

- Neural Network

เป็นเทคนิคสำหรับการวิเคราะห์แบบ Predictive model เป็นระบบที่มีการทำงานซับซ้อนและต้องใช้เวลานานในการสร้างแบบจำลองที่เลียนแบบการทำงานของสมองมนุษย์ ทำให้การแก้ไข

ปัญหาไม่จำเป็นต้องใช้คนช่วยคิดมากนักเทคนิคนี้นิยมนำมาใช้ในการตรวจสอบพฤติกรรมการโกงของลูกค้าบัตรเครดิต หรือลูกค้าสินเชื่อประเภทต่างๆ

- Induction

เป็นเทคนิคที่ใช้ในการอนุมานกฎเกณฑ์และความสัมพันธ์ที่ไม่เคยมีมาก่อนของข้อมูลเพื่อให้สามารถเข้าใจความสัมพันธ์เหล่านั้น เช่น ลูกค้า 50% ที่ซื้อผ้าอ้อมจะต้องซื้อเบียร์ด้วย

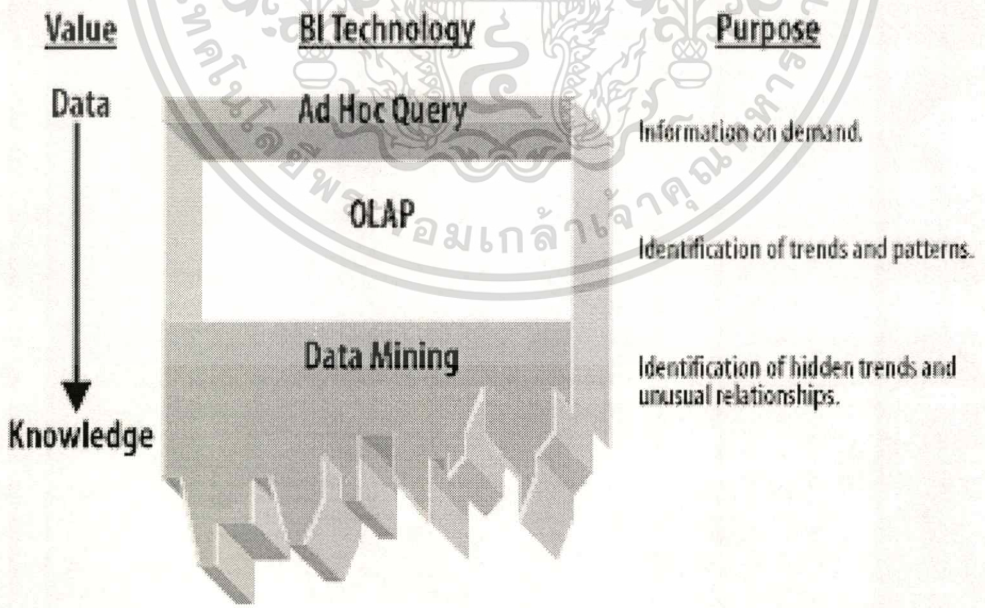
- Statistics

เป็นเทคนิคพื้นฐานของดาต้าไมนิ่ง ซึ่งใช้หลักการทางสถิติเข้ามาช่วยในการสรุปข้อมูล ทำให้ต้องใช้ทักษะส่วนบุคคลสูงในเชิงการคำนวณเพื่อตีความผลลัพธ์ที่ได้

- Visualization

เป็นเทคนิคในการนำเสนอข้อมูลในรูปของ Graphic หรือแผนภาพสามมิติ เพื่อให้ผู้ใช้งานสามารถหาแนวโน้ม รูปแบบและความสัมพันธ์ของตัวแปรต่าง ๆ ได้

ในขณะที่ Online Analytical Processing (OLAP) และภาษาในการสอบถามข้อมูล (Ad Hoc Query) จัดเป็นเทคนิคหนึ่งของดาต้าไมนิ่ง (โดย GartnerGroup) แต่เป็นเทคนิคที่ผู้ใช้งานมีส่วนร่วมอย่างมากในการหาแนวโน้มและความสัมพันธ์ที่ซ่อนอยู่ในข้อมูล จึงเป็นวิธีที่ไม่มีประสิทธิภาพนักในการทำดาต้าไมนิ่ง



รูปที่ 2.2 แสดงวัตถุประสงค์ในการหาผลลัพธ์จากเทคโนโลยีต่างๆ เทียบกับดาต้าไมนิ่ง

จากรูป จะเห็นว่าโปรแกรมประเภท Ad Hoc query, OLAP และ Data Mining ต่างให้คำตอบในมิติที่แตกต่างกัน ดาต้าไมนิ่งจะถูกนำมาใช้ในการหาแนวโน้มที่และความสัมพันธ์ซึ่งซ่อนอยู่ในข้อมูล ซึ่งสิ่งเหล่านี้ก่อให้เกิดคุณค่าในแง่ของการสร้างความรู้ และสารสนเทศที่เป็นประโยชน์เพื่อนำมาช่วยสนับสนุนการตัดสินใจการดำเนินงาน

2.2 กระบวนการทำงานของดาต้าไมนิ่ง

กระบวนการทำงานของดาต้าไมนิ่งประกอบด้วยขั้นตอนดังนี้

2.2.1 กำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

เป็นส่วนที่สำคัญที่สุดของการวิเคราะห์ เนื่องจากถ้าเราวางวัตถุประสงค์ได้ชัดเจนแล้ว ในขั้นตอนต่อๆ มาจะสามารถทำงานได้อย่างตรงประเด็น ทำให้ไม่เสียเวลาในการที่จะต้องเริ่มต้นใหม่ เนื่องจากการวางวัตถุประสงค์ที่ไม่แน่ชัด ผลที่วิเคราะห์ได้ก็อาจจะคลุมเครือ ไม่สามารถนำมาประยุกต์ใช้ต่อไปได้

2.2.2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ค่อนข้างใช้เวลามากกว่าขั้นตอนอื่นๆ อาจใช้เวลานานถึง 60 เปอร์เซ็นต์ ซึ่งประกอบด้วย 3 ขั้นตอนย่อย คือ

1. การเลือกข้อมูล (Data Selection)

   เป็นการคัดข้อมูลที่อยู่ในฐานข้อมูลที่สัมพันธ์กับวัตถุประสงค์ในการวิเคราะห์ ซึ่งข้อมูลนี้จะต้องเป็นข้อมูลที่มีความสมบูรณ์ และถูกต้องที่สุด โดยจะเลือกข้อมูลที่ต้องการและนำข้อมูลที่ไม่ต้องการออกไปซึ่งเป็นการเริ่มต้นของการทำไมนิ่ง เราสามารถแบ่งประเภทของข้อมูลได้เป็น

   - ข้อมูลตัวเลข (Quantitative) คือ ตัวเลขจำนวนเต็ม และ จำนวนจริง เช่น รหัสพนักงาน, อายุ, ค่าจ้าง เป็นต้น
   - ข้อมูลที่ไม่ใช่ตัวเลข (Categorical Data) สามารถแบ่งได้เป็น
     - ข้อมูลที่ไม่มีลำดับความสำคัญ (Nominal) เช่น ชื่อ-นามสกุล, เพศ
     - ข้อมูลที่มีลำดับความสำคัญ (Ordinal) เช่น เกรด (A, B, C, D, F)

2. การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)

   เป็นการกลั่นกรองข้อมูลให้เหมาะสมก่อนที่จะนำไปทำดาต้าไมนิ่ง ซึ่งข้อมูลที่เลือกมานั้นอาจมีข้อมูลที่ผิดพลาดหรือค่าข้อมูลขาดหายไป การกลั่นกรองข้อมูล

ประเภท Categorical มักใช้หลักการทางสถิติ เช่น การจัดการกระจายข้อมูลโดยการนำเอาข้อมูลมาสร้างกราฟ ซึ่งจะช่วยให้เห็นความโน้มเอียงของข้อมูล และสังเกตความผิดปกติของข้อมูลได้ ส่วนข้อมูลประเภท Quantitative การวิเคราะห์ข้อมูลทำได้โดยการหาค่าสูงสุด (Max), ค่าต่ำสุด (Min), ค่าเฉลี่ย (Mean), ค่าที่ปรากฏบ่อย (Mode), ค่ากลาง (Median) เป็นต้น

3. การแปลงข้อมูล (Data Transformation)

ในข้อมูลบางชนิดที่จะต้องนำมาประมวลผลหรือเป็นข้อมูลที่อยู่ในรูปแบบที่ไม่เหมาะกับการนำมาวิเคราะห์ สามารถทำการแปลงรูปแบบของข้อมูลให้อยู่ในรูปแบบที่เหมาะสมได้ เช่น ใช้การกำหนดตัวเลข แทนสถานภาพสมรส โดยกำหนดให้

1 = สถานภาพโสด

2 = สถานภาพแต่งงาน

3 = สถานภาพหย่าร้าง

หรือข้อมูลที่มีค่าที่ผิดไปจากความเป็นจริง เช่น วันที่ ต้องตรวจสอบว่ามีข้อมูลใดที่มีค่าไม่ตรงกับความเป็นจริงหรือไม่ รวมถึงจัดการข้อมูลบางส่วนที่ขาดหายไป ซึ่งอาจจะกำหนดค่าเฉพาะที่ทำให้ทราบว่าเป็นค่าที่ใช้แทนข้อมูลว่าง หรือจะใช้ค่าเฉลี่ยของข้อมูลทั้งหมดมากำหนดแทนก็ได้

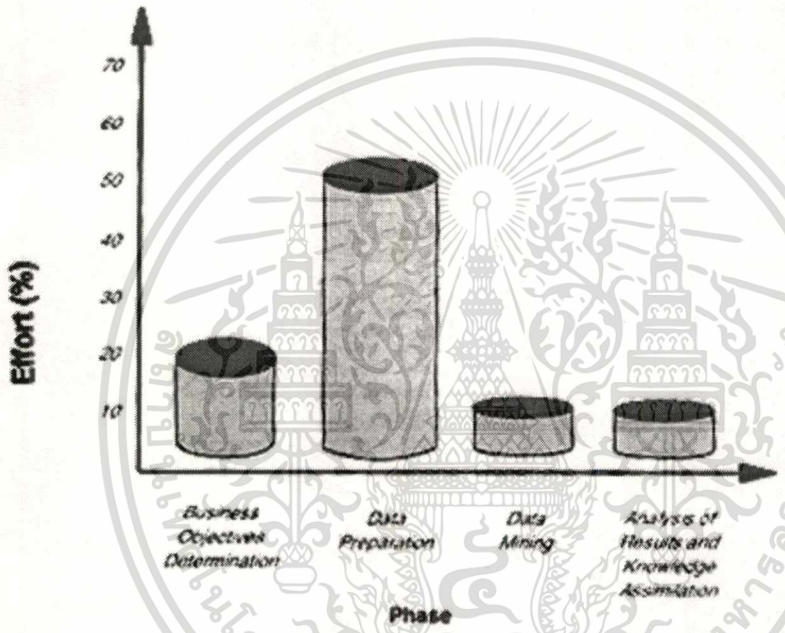
2.2.3 การทำดาต้าไมนิ่ง (Data Mining)

เป็นการนำเอาเทคนิคและกระบวนการมาที่เหมาะสมของดาต้าไมนิ่ง (Data Mining) ช่วยในการวิเคราะห์ข้อมูล ซึ่งการพิจารณาว่าเทคนิคใดจะเหมาะสม ต้องพิจารณาถึงวัตถุประสงค์ที่ตั้งไว้ ข้อมูลที่มีอยู่ และผลที่ต้องการจะได้รับจากการวิเคราะห์

2.2.4 การวิเคราะห์ผลลัพธ์ที่ได้ (Analysis of Result)

ผลที่ได้จากการทำดาต้าไมนิ่ง เป็นเพียงข้อมูลที่จะช่วยให้มองเห็นรูปแบบของข้อมูลที่มีอยู่ว่าเป็นอย่างไร ดังนั้นต้องวิเคราะห์ว่าผลที่ได้ตรงกับวัตถุประสงค์ที่วางไว้หรือไม่ และได้ประโยชน์มากน้อยเพียงใดและเหมาะสมที่จะนำไปเป็นแนวทางในการเพิ่มผลกำไรในธุรกิจนั้นๆ หรือไม่ ถ้าผลที่ได้ไม่เป็นที่น่าพอใจ ก็สามารถที่จะกลับไปเริ่มต้นใหม่ โดยอาจจะพิจารณาหาว่ามีจุดบกพร่องที่ขั้นตอนใด แล้วจึงย้อนไปแก้ไขที่สุดนั้นได้

ทุกขั้นตอนในการวิเคราะห์ข้อมูลที่กล่าวมา จะต้องใช้เวลาในการจัดการเตรียมข้อมูลต่างๆ โดยเราสามารถเปรียบเทียบเวลาที่ใช้ในขั้นตอนต่างๆ ได้ดังรูปที่ 2.3 จะเห็นว่าเวลาส่วนใหญ่ของการวิเคราะห์จะอยู่ในส่วนของการเตรียมข้อมูล ดังนั้น ในการวิเคราะห์ใดๆ ควรให้ความสนใจต่อการเตรียมข้อมูลที่จะใช้ในการวิเคราะห์ พิจารณาเลือกใช้ข้อมูลที่คาดว่าจะเป็นปัจจัยต่อรูปแบบของข้อมูล



รูปที่ 2.3 แผนภาพแสดงเวลาที่ใช้ในแต่ละขั้นตอนของการวิเคราะห์ข้อมูล

2.3 เทคนิคของการทำดาต้าไมนิ่ง

เทคนิคหลักๆ ในการทำดาต้าไมนิ่ง มี 4 ประเภท ดังนี้

1. Predictive Modeling

มีลักษณะคล้ายการเรียนรู้ของมนุษย์ คือจะต้องเข้าใจลักษณะของสิ่งที่ศึกษาอย่างแท้จริง ในดาต้าไมนิ่งเราจะใช้แบบจำลอง (Model) นี้ในการวิเคราะห์ข้อมูลที่มีอยู่ เพื่อกำหนดคุณสมบัติที่สำคัญของข้อมูล ดังนั้นข้อมูลที่มีอยู่จะต้องเป็นข้อมูลที่สมบูรณ์ จึงจะทำให้แบบจำลองให้คำทำนายที่ถูกต้องโดยเริ่มต้นเริ่มต้นจะต้องให้คำตอบที่ถูกต้องกับแบบจำลอง เพื่อแบบจำลองจะได้เห็นถึงข้อสังเกตใหม่ ๆ วิธีนี้เรียกว่า “Supervised Learning” ซึ่งการทำงานจะมีลักษณะคล้ายกับ IF THEN การพัฒนาแบบจำลองพยากรณ์จะนำเอาข้อมูลในอดีตมาสร้างแบบจำลอง โดยแบ่งออกเป็น 2 ขั้นตอน คือ

- Training Phase เพื่อสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต
- Testing Phase เพื่อทดสอบแบบจำลองที่สร้างว่ามีความเหมาะสมหรือไม่ โดยใช้กับข้อมูลที่ถูกแบ่งเอาไว้สำหรับทำการทดสอบ

Predictive Modeling แบ่งได้เป็น 2 ลักษณะคือ

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

2. Database Segmentation

เป็นการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยไม่รู้ล่วงหน้าว่าจะมีทั้งหมดกี่กลุ่ม จึงเรียกเทคนิคนี้ว่า "Unsupervised Learning" โดยการจัดกลุ่มดังกล่าวได้จากการพิจารณาคุณสมบัติในหลาย ๆ มิติของข้อมูล ถ้ารายการในข้อมูลมีลักษณะคล้ายคลึงเป็นกลุ่มเดียวกันได้ก็จะรวมเข้าด้วยกัน เพื่อให้ง่ายต่อการวิเคราะห์ เช่น การแบ่งลูกค้าออกตามอายุ, เพศ, รายได้ เป็นต้น

3. Link Analysis

เป็นกระบวนการที่ศึกษาว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันหรือไม่ อย่างไร ความสัมพันธ์นี้เรียกว่า "Association" เช่น การค้นหาความสัมพันธ์ระหว่างสินค้าหรือบริการที่ลูกค้ามักจะซื้อด้วยกัน เป็นต้น เพื่อนำผลการทำนายมาเป็นแนวทางในการส่งเสริมการขายและการจัดวางสินค้าให้เหมาะสม

4. Deviation Detection

เป็นวิธีการที่หาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าแตกต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) เช่น การตรวจสอบลายเซ็นปลอมหรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

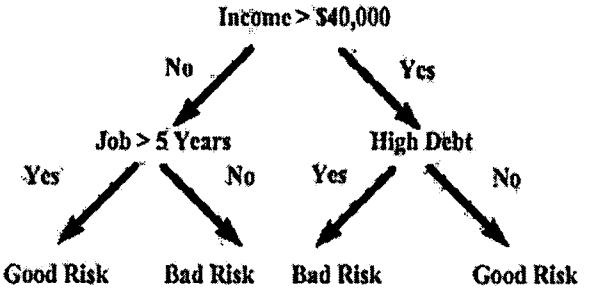
นอกจากนี้แต่ละเทคนิคของดาต้าไมนิ่งยังสามารถเลือกที่จะนำเอาอัลกอริทึมต่าง ๆ มาใช้ได้ ซึ่งการเลือกใช้อัลกอริทึมนั้นขึ้นอยู่กับปัจจัยหลายอย่าง เช่น จุดมุ่งหมายของการวิเคราะห์ ลักษณะของของข้อมูล ชนิดของข้อมูล และจำนวนข้อมูลที่มีอยู่ ในการทำดาต้าไมนิ่งบางครั้งอาจต้องมีการเปลี่ยนแปลงเทคนิคกรทำ ถ้าเทคนิคที่ใช้นั้นไม่เหมาะสม

ขั้นตอนที่สำคัญของกระบวนการนำมาใช้อยู่ที่การกำหนดถึงกลุ่มข้อมูลที่จะนำมา และการสร้างแบบจำลองที่เหมาะสมกับข้อมูล ซึ่งจะทำให้ผลของการทำดาต้าไมนิ่งได้ผลอย่างถูกต้องและรวดเร็ว

2.4 การจัดกลุ่ม (Classification)

Classification เป็นเทคนิคหนึ่งในดาต้าไมนิ่งที่ใช้สำหรับสร้างแบบจำลองพยากรณ์ (Predictive Model) โดยจะทำการสร้างแบบจำลองจากกลุ่มข้อมูลตัวอย่างที่เลือกมาจากฐานข้อมูลขนาดใหญ่ และแบบจำลองนั้นสามารถพยากรณ์ผลลัพธ์ของข้อมูลที่ไม่เคยพบเห็นมาก่อน บนพื้นฐานความสัมพันธ์ของกลุ่มข้อมูลที่มีอยู่เดิม แบบจำลองที่สามารถทำงานได้ตามลักษณะนี้เรียกว่า Supervised learning สำหรับเทคนิคที่ใช้ใน Classification นั้นยังแบ่งได้เป็น 2 แบบ คือ Tree Induction และ Neural Induction โดยในที่นี้จะนำเสนอเทคนิคของ Tree Induction ในการประยุกต์ใช้ในการวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย รวมทั้งทำนายว่าลูกค้ากลุ่มใดมีแนวโน้มที่จะยกเลิกการใช้บริการ เพื่อเป็นแนวทางให้ผู้ประกอบการสามารถหากลยุทธ์ทางการตลาดมาช่วยในการรักษาลูกค้ากลุ่มดังกล่าวไว้ โดยเทคนิค Tree Induction เป็นการนำข้อมูลที่มีอยู่มาสร้างแบบจำลองพยากรณ์ในรูปแบบของ Decision Tree

เทคนิค Classification แบบ Decision Tree เป็นเทคนิคการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของ Tree ซึ่งประกอบด้วย Node แรกสุดที่เรียกว่า Root node จาก Root node ก็จะแตกออกเป็น Node และ Node ระดับสุดท้ายเรียกว่า Leaf node โดยที่แต่ละ Node ที่ไม่ใช่ Leaf node จะแทนจุดที่มีการตรวจสอบเงื่อนไขหรือ เป็นการตัดสินใจเกี่ยวกับลักษณะของข้อมูลที่นำมาพิจารณา และจากผลลัพธ์ที่ได้ในแต่ละครั้งก็จะมีการเลือกเส้นทางที่แน่นอน โดยในการจัดแบ่งประเภทของข้อมูลนั้นจะ เริ่มต้นพิจารณาจาก Root node และอ้างลงมาตามโครงสร้างของ Tree เรื่อย ๆ จนกระทั่งถึง Leaf node ซึ่งแสดงว่ากระบวนการวิเคราะห์ข้อมูลนั้นเสร็จสิ้นลง และสามารถบอกลักษณะหรือประเภทของข้อมูลนั้นได้ตาม Node ที่ข้อมูลผ่านการตรวจสอบ ดังรูป



รูปที่ 2.4 Decision Tree

การนำข้อมูลมาสร้าง Tree นั้นมีขั้นตอนพื้นฐาน คือ

1. หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูล โดย Attribute นี้จะถูกนำมาสร้างเป็น Root node โดยจะมี Target
   Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf node ถูกกำหนดไว้ก่อน
2. นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกเป็นกลุ่ม
3. แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root node
4. วนกลับไปทำแบบขั้นตอนแรก คือ หา Attribute ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหา Attribute ที่ใช้แบ่งข้อมูลต่อไป

สำหรับอัลกอริทึมที่ใช้สร้าง Decision Tree มีหลายอัลกอริทึม เช่น CHAID, CART, SLIQ, ID3, C4.5 และ C5.0 เป็นต้น ซึ่งในแต่ละอัลกอริทึมจะมีวิธีการที่แตกต่างกันในการหา Attribute ที่จะ นำมาใช้แบ่งข้อมูล ซึ่งในที่นี้ได้เลือกใช้อัลกอริทึม C4.5 มาประยุกต์ใช้ในการพัฒนาระบบเพื่อวิเคราะห์หาสาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย และเนื่องจากอัลกอริทึม C4.5 เป็นอัลกอริทึมที่พัฒนาจากอัลกอริทึม ID3 จึงขอกล่าวถึงการทำงานของ

อัลกอริทึม ID3 และ C4.5 ในส่วนที่เพิ่มเติมละปรับปรุงการทำงานของ ID3 ออกเป็นหัวข้อดังนี้

- อัลกอริทึม ID3
- อัลกอริทึม C4.5

2.5 อัลกอริทึม ID3 (ID3 Algorithm)

การเลือก Attribute ที่มีความสำคัญที่สุดเพื่อใช้แบ่งข้อมูล จะใช้หลักการของ Gain Criterion ในการวัด โดยกำหนดให้

T แทน Training Set

S แทน set ของข้อมูลใด ๆ

Freq(C_j,S) แทน จำนวนของข้อมูลใน S ซึ่งอยู่ใน Class C_j

|S| แทน จำนวนของข้อมูลใน S

info(S) หรือ entropy ของ set S เป็นการวัดค่าของ information

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

และเมื่อนำสูตรนี้ไปประยุกต์ใช้กับ Training Set จะได้ info(T)

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

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

gain (X) เป็นการวัดค่าของ Information ที่ได้รับถ้า เลือก Attribute X

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

ต่อไปจะอธิบายการทำงานของ ID3 โดยใช้ข้อมูลจากตาราง ดังนี้

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

ตารางที่ 2.1 Training set

จากตารางที่ 2.1 จะเห็นว่าประกอบด้วย class 2 class คือ Play และ Don't Play โดยข้อมูลอยู่ใน class Play จำนวน 9 record และ class Don't Play จำนวน 5 record จะได้ว่า

$$Info(T) = -9/14 \times log_2(9/14) - 5/14 \times log_2(5/14)$$
$$= 0.940 \; bits$$

พิจารณา Attribute ต่าง ๆ โดยหาค่า Gain ของแต่ละ Attribute ออกมา แล้วเลือก Attribute ที่มีค่า Gain สูงสุดมา เป็นตัวแบ่งข้อมูล หรือ Root node

พิจารณาที่ attribute outlook ซึ่งสามารถแบ่งข้อมูลได้เป็น 3 subset จะได้ว่า

$$Info_X(T) = 5/14 \times (-2/5 \times log_2(2/5) - 3/5 \times log_2(3/5))$$
$$+ 4/14 \times (-4/4 \times log_2(4/4) - 0/4 \times log_2(0/4))$$
$$+ 5/14 \times (-3/5 \times log_2(3/5) - 2/5 \times log_2(2/5))$$
$$= 0.694 \; bits$$

ดังนั้น ค่า Gain ของ Attribute outlook หรือ gain(x) = 0.940 – 0.694 = 0.246 bits

ในทำนองเดียวกัน ถ้าพิจารณาที่ Attribute windy จะแบ่ง ข้อมูลได้ 2 subset โดย set แรกมีข้อมูลจำนวน 6 record ซึ่งอยู่ ใน class Play 3 record และอยู่ใน Class Don't Play 3 record ส่วน set ที่ 2 มีข้อมูลทั้งหมด 8 record อยู่ใน Class Play 6 record และอยู่ใน class Don't Play 2 record ดังนั้น

$$Info_X(T) = 6/14 \times (-3/6 \times log_2(3/6) - 3/6 \times log_2(3/6))$$
$$+ 8/14 \times (-6/8 \times log_2(6/8) - 2/8 \times log_2(2/8))$$
$$= 0.892 \; bits$$

$$gain(X) = 0.940 - 0.892 = 0.048 \; bits$$

จะพบว่าค่า Gain ที่ได้จากการแบ่ง Training Set โดยใช้ Attribute outlook > Windy ดังนั้นควรใช้ Attribute outlook ในการแบ่ง Training Set

ตามหลักการของ ID3 ต้องคำนวณหาค่า Gain ของทุก Attribute แล้วเลือก Attribute ที่มีค่า Gain สูงสุด แต่จากข้อมูล ตัวอย่างพบว่า ค่าใน Attribute Temp และ Humidity เป็นค่าชนิด continuous ซึ่งในกรณีนี้ ID3 ไม่สามารถจัดการได้ ต้องใช้ C4.5 ซึ่งจะกล่าวถึงในหัวข้อถัดไป

2.6 อัลกอริทึม C4.5 (C4.5 Algorithm)

C4.5 เป็นอัลกอริทึมที่พัฒนามาจาก ID3 โดยเพิ่ม feature ต่าง ๆ ขึ้นมา ดังนี้

- **Gain ratio criterion** พัฒนาขึ้นมาเพื่อแก้ปัญหาของ Gain Criterion กรณีที่ Attribute มีค่าที่ Unique การแบ่งข้อมูลโดยใช้ Attribute นี้จะทำให้เกิด subset จำนวนมาก ซึ่งใน แต่ละ subset จะประกอบด้วยข้อมูลเพียง 1 record เท่านั้น ทำให้ $info_X(T) = 0$ ซึ่งจะมีผลให้ค่า Information Gain ของ Attribute นี้มีค่าสูงมาก และการแบ่งข้อมูลโดย ใช้ Attribute นี้ไม่ก่อให้เกิดประโยชน์ใดๆ ต่อการทำนาย อัลกอริทึม C4.5 แก้ไขปัญหานี้โดยใช้ค่า Gain ratio ซึ่ง คำนวณโดยใช้ split info(X) และ gain ratio(X) โดย split info(X) เป็นค่า Information ที่ได้จากการแบ่ง T ออกเป็น n subset

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

Gain ratio(X) เป็นการวัดว่าการแบ่งข้อมูลโดยใช้ Attribute นั้นๆ ก่อให้เกิดประโยชน์ต่อการทำนาย หรือไม่

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

จากตัวอย่างที่แล้วพบว่าการแบ่งข้อมูลโดยใช้ Attribute outlook ทำให้เกิด subset ทั้งหมด 3 subset ซึ่งประกอบ ด้วยจำนวน record เท่ากับ 5, 4 และ 5 record ตามลำดับ Split Information คำนวณได้ดังนี้

$$\begin{aligned} split\ info\ (X) = & -5/14 \times log_2(5/14) \\ & -4/14 \times log_2(4/14) \\ & -5/14 \times log_2(5/14) = 1.577\ bits. \end{aligned}$$

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

ซึ่งพบว่าการใช้ Gain ratio criterion ทำให้ Tree ที่ได้มี ขนาดเล็กกว่าการใช้ Gain criterion

- **Unknown attribute values**
  - การหา Attribute เพื่อใช้แบ่งข้อมูลทำได้โดย

    1. หาค่า info(T) และ $info_X(T)$ โดยพิจารณาเฉพาะ ข้อมูลที่รู้ค่าของ A

    2. หาค่า gain(X) โดย

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

    3. หาค่า Split info(X) โดยพิจารณากลุ่มของข้อมูลที่ไม่ รู้ค่าของ A เป็นอีก 1 subset เช่น ถ้า Attribute ที่จะนำมา ทดสอบมีค่าที่เป็นไปได้ n ค่า split info(X) จะถูก คำนวณ โดยแบ่งข้อมูลออกเป็น n+1 subset

  - การแบ่ง Training set สมมติ Attribute ที่เลือกจาก ขั้นตอนแรกมีค่าที่เป็นไปได้ คือ $O_1, O_2, \ldots O_n$ เมื่อข้อมูล 1 record ใน T ซึ่งมีค่า O_i ถูกกำหนดให้ subset T_i ค่าความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset T_i เท่ากับ 1 และความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset อื่น ๆ เท่ากับ 0 แต่ถ้าค่า ใน Attribute ไม่ทราบค่า ความน่าจะเป็นจะมีค่า น้อยลง สำหรับข้อมูลแต่ละ record ในแต่ละ subset T_i weight จะเท่ากับความน่าจะเป็นของ O_i ที่จุดนั้น ๆ ทำให้ $|T_i|$ เป็นผลรวมของค่า weight w ซึ่งค่าใน Attribute ไม่ ทราบค่าจะถูกกำหนดให้แต่ละ subset T_i ด้วย weight

$$w \times Probability\ of\ outcome\ O_i$$

    โดยความน่าจะเป็นคือ ผลรวมของ Weight ของข้อมูล ทั้งหมดใน T ซึ่งมีค่า O_i หารด้วยผลรวมของ weight ของ ข้อมูลทั้งหมดใน T ซึ่งค่าใน Attribute เป็นค่าที่ทราบค่า

  - การใช้ decision tree ที่ได้มาทำนายกลุ่มของข้อมูล ใน กรณีที่ค่าใน attribute ที่จะทดสอบที่ decision node เป็น ค่าที่ไม่ทราบค่า ทำให้ไม่สามารถแบ่งข้อมูลได้ กรณีนี้ ระบบจะสำรวจทุกเส้นทางที่เป็นไปได้ และรวมผลที่ได้ จากการ classification ด้วยวิธีการทางคณิตศาสตร์ โดย ผลที่ได้จะเกิดได้หลายเส้นทางจาก root ของ tree หรือ subtree ไปยัง leaf node และ class ที่ได้จากการทำนาย จะเป็น class ที่มีความน่าเป็นสูงสุด

ต่อไปจะนำเสนอตัวอย่างเพื่อให้เกิดความเข้าใจยิ่งขึ้น โดยใช้ตัวอย่างจากตารางที่ 2.1 โดยแบ่งเป็น 3 ขั้นตอนดังนี้

1. การหา Attribute เพื่อให้แบ่งข้อมูล สมมติว่าค่าใน attribute Outlook ใน record ที่ 6 เป็นค่าที่ไม่ทราบค่า ซึ่งแทนด้วย "?" ซึ่งเราจะพิจารณาเฉพาะข้อมูล 13 record ที่เหลือ จะได้ความถี่ดังนี้

	Play	Don't Play	Total
Outlook = Sunny	2	3	5
Overcast	3	0	3
Rain	3	2	5
Total	8	5	15

ตารางที่ 2.2 แสดงความถี่ของข้อมูล

ทำการคำนวณค่าต่างๆ โดยพิจารณา Attribute Outlook ดังนี้

$$\begin{aligned} Info(T) &= -8/13 \times log_2(8/13) - 5/13 \times log_2(5/13) \\ &= 0.961 \ bits \end{aligned}$$

$$\begin{aligned} Info_X(T) &= 5/13 \times (-2/5 \times log_2(2/5) - 3/5 \times log_2(3/5)) \\ &\quad + 3/13 \times (-3/3 \times log_2(3/3) - 0/3 \times log_2(0/3)) \\ &\quad + 5/13 \times (-3/5 \times log_2(3/5) - 2/5 \times log_2(2/5)) \\ &= 0.747 \ bits \end{aligned}$$

$$\begin{aligned} gain(T) &= 13/14 \times (0.961 - 0.747) \\ &= 0.199 \ bits \end{aligned}$$

จะพบว่าค่า gain ที่ได้ลดลงเล็กน้อย จากเดิม 0.246 เป็น 0.199 bits ส่วนค่า split information จะพิจารณาจากข้อมูลใน training set ทั้งหมด จึงทำให้ค่าที่ได้มีเพิ่มขึ้นจาก 1.577 เป็น 1.809 bits ดังนี้

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

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

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

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

และค่า gain ratio ลดลงจาก 0.156 เป็น 0.110

2. การแบ่ง Training set เมื่อข้อมูลใน training set ทั้ง 14 record ถูกแบ่งออกโดยใช้ค่าใน Attribute outlook record ที่มีค่าใน Attribute outlook เป็นค่าที่ไม่ทราบค่า จะถูกกำหนดให้ในทุก subset คือ Sunny, Overcast และ Rain ด้วยค่า weight เท่ากับ 5/13, 3/13 และ 5/13 ตามลำดับ พิจารณาที่ subset แรก ดังนี้

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

ตารางที่ 2.3 แสดง subset ของ outlook = Sunny

ถ้า subset นี้ถูกแบ่งต่อไปโดยใช้ Attribute humidity การกระจายของ class ใน subset จะเป็นดังนี้

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

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

decision tree ที่ได้จะมีโครงสร้างดังนี้

outlook = sunny :

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

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

outlook = overcast : Play (3.2)

outlook = rain :

| windy = true : Don't Play(2.4/0.4)

| windy = false : Play(3.0)

โดยค่าของตัวเลขที่ leaf node จะอยู่ในรูป (N) หรือ (N/E) จะมีความสำคัญ

N เป็นจำนวนข้อมูลทั้งหมดที่มาถึง leaf node นั้น ๆ

E เป็นจำนวนข้อมูลที่ไม่อยู่ใน class ที่ระบุไว้ เช่น Don't Play (3.4/0.4) หมายความว่า จำนวนข้อมูลที่มาถึงที่ leaf node นี้เท่ากับ 3.4 และ 0.4 ในจำนวนนี้ไม่อยู่ใน class Don't Play

3. การใช้ decision tree ที่ได้มาทำนายกลุ่มของข้อมูล สมมติ ข้อมูล คือ

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

จากค่าใน outlook พบว่าต้อง move ไปยัง subset แรก แต่เนื่องจากไม่สามารถตรวจสอบค่าที่ humidity ได้ จึงทำการพิจารณา ดังนี้

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

จะพบว่าการกระจายของ class สุดท้ายสำหรับข้อมูลนี้เท่ากับ

Play : $2.0/5.4 \times 100\% + 3.4/5.4 \times 12\% = 44\%$

Don't Play : $3.4/5.4 \times 88\% = 56\%$

- **Continuous attribute values** สมมติว่า A เป็น Attribute ชนิด continuous numeric value การทดสอบค่า ที่ Attribute นี้ จะแบ่งเป็น $A \leq Z$ และ $A > Z$ โดยทำการเปรียบเทียบค่า Threshold value Z โดยการหาค่า Threshold ที่เหมาะสมมีขั้นตอนดังนี้

1. เรียงลำดับ Training Set ด้วยค่าใน Attribute A จากน้อยไปมาก และเลือกพิจารณาเฉพาะค่าที่ไม่ซ้ำกันมาพิจารณาจะได้ $\{v_1, v_2, \ldots, v_n\}$
2. หาค่า Threshold ใด ๆ ซึ่งค่า Threshold ใด ๆ จะอยู่ระหว่าง v_i และ v_{i+1} โดยคำนวณจาก Midpoint ของแต่ละช่วงดังนี้ $v_i + v_{i+1} / 2$ โดย C4.5 จะเลือกค่าที่มากที่สุดใน Attribute A แต่ต้องไม่เกินค่า Midpoint นั้นๆ จาก Training Set เป็นค่า Threshold ของแต่ละช่วง เพื่อที่ว่าค่า Threshold ทั้งหมดที่ปรากฏอยู่ใน Tree หรือ Rule จะเป็นค่าที่เกิดขึ้นจริงในข้อมูล
3. หาค่า Threshold ที่เหมาะสม โดยพิจารณาจากค่า Threshold ที่มีค่า information Gain สูงสุด

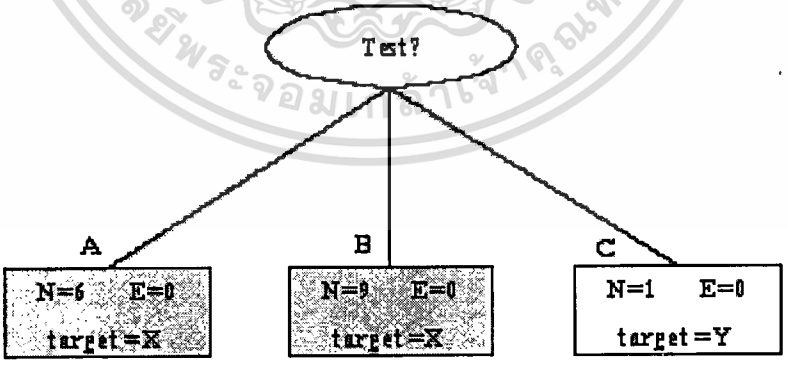
- **Pruning decision trees** การแบ่งข้อมูลใน Training set เพื่อสร้าง Decision trees จะกระทำไปจนกระทั่งข้อมูลในแต่ละ subset อยู่ใน class เดียวกัน ซึ่งผลลัพธ์ที่ได้อาจทำให้ tree มีความซับซ้อนมากเกินไปที่เรียกว่า "Overfits the data" ซึ่งปัญหานี้สามารถแก้ไขได้โดยทำการ Pruning ซึ่งการ Pruning จะทำให้แต่ละ leaf node ของ tree ที่ได้ไม่จำเป็นที่จะต้องประกอบด้วยข้อมูลที่อยู่ใน class เดียวกันทั้งหมด โดยแต่ละ leaf node จะมีการระบุการกระจายของข้อมูลแต่ละ class ไว้ ซึ่งจะบอกถึงความน่าจะเป็นที่ข้อมูลจะอยู่ใน class นั้นๆ อัลกอริทึมของ C4.5 จะทำการ Pruning โดยการตัด subtree ที่ทำให้เกิดการผิดพลาดในการทำนายออกไป แล้วทำการแทนที่ Subtree นั้นด้วย leaf node โดยเทคนิคนี้จะใช้เพียงข้อมูลใน Training set ที่ใช้ในการสร้าง tree เท่านั้น และการคำนวณความผิดพลาดที่เกิดจากการทำนายของแต่ละ leaf node และ subtree จะทำได้โดยสมมติว่าจะทำการแบ่งกลุ่ม set ของข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Training set โดยการคำนวณจะใช้ function ทางสถิติ ซึ่งอยู่บนพื้นฐานของการกระจายแบบ Binomial จำนวน error ที่เกิดขึ้นเมื่อข้อมูลมีขนาดเท่ากับ $N = N \times U_{CF}(E,N)$

  โดย N แทน ขนาดของข้อมูลที่ leaf node ใดๆ

  E แทน จำนวนของ error ที่เกิดขึ้นใน set ของข้อมูลที่ leaf node ใดๆ

  $U_{CF}(E,N)$ แทน ความน่าจะเป็นสูงสุดที่จะเกิด error และ C4.5 จะใช้ Confidence level ที่ 25%

ต่อไปจะอธิบายการ Pruning โดยพิจารณา Subtree ดังรูป



รูปที่ 2.5 Subtree ก่อนทำการ Pruning

จากรูปจะพบว่าค่าที่เป็นไปได้ที่เกิดจากการทดสอบมี 3 ค่าคือ A, B และ C และ Target attribute มี 2 ค่าคือ X และ Y ซึ่งในกรณีนี้ไม่พบ error ที่เกิดขึ้นใน Training set ใน leaf node ที่ 1 พบว่า N = 6 และ E = 0 ดังนั้น

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

ดังนั้น ถ้าเราใช้ Leaf node นี้ในการแบ่งกลุ่มข้อมูลจำนวน 6 record จำนวน error ที่เกิดขึ้นในการทำนายจะเท่ากับ 6 × 0.206 สำหรับ leaf nodeที่ 2 และ 3 จะได้ $U_{25\%}(0,9) = 0.143$ และ $U_{25\%}(0,1) = 0.750$ ตามลำดับ ดังนั้นจำนวน error ที่เกิดจากการทำนายของ subtree นี้เท่ากับ

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

และจะพบว่าถ้าทำการแทนที่ subtree นี้ด้วย leaf node ที่มี Target = X เมื่อ X เป็นค่าที่มีความถี่มากที่สุดของ Target attribute ของ Training subset จำนวน 16 record จะเกิด error 1 record และจำนวน error ที่เกิดจากการทำนายเท่ากับ

$$16 \times U_{25\%}(1,16) = 16 \times 0.157 = 2.52$$

จะพบว่า subtree นี้มีจำนวนของ error ที่เกิดจากการทำนายสูงกว่า ดังนั้นจึงทำการแทนที่ด้วย leaf node

บทที่ 3
วิธีดำเนินการศึกษาเพื่อทำการวิเคราะห์สาเหตุการยกเลิก การใช้บริการในธุรกิจประกันภัย

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงต้องมีการกำหนดวัตถุประสงค์ การเตรียมข้อมูล และกระบวนการต่างๆ ตามขั้นตอนของการทำดาต้าไมนิ่ง ดังนี้

3.1 กำหนดวัตถุประสงค์

การที่ลูกค้าเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่นหรือยกเลิกการใช้บริการ(Churn) เป็นปัญหาที่สำคัญปัญหาหนึ่งที่พบในธุรกิจการให้บริการประกันกันภัย เนื่องจากจำนวนผู้ให้บริการประกันภัยที่มีอยู่จำนวนมาก ทำให้ลูกค้ามีทางเลือกมากขึ้นในการเลือกใช้บริการ ซึ่งเป็นสาเหตุทำให้ผู้ให้บริการสูญเสียรายได้ รวมทั้งค่าใช้จ่ายที่ใช้ในการหาลูกค้าใหม่นั้นสูงกว่าค่าใช้จ่ายที่ใช้ในการรักษาลูกค้าเดิมไว้มาก ผู้ประกอบการจึงหันมาให้ความสำคัญกับการรักษาลูกค้าให้ใช้บริการกับบริษัทของตนต่อไป

วัตถุประสงค์ของการประยุกต์ใช้ดาต้าไมนิ่งในการจัดการกับการเปลี่ยนใจในการใช้บริการของลูกค้าในธุรกิจประกันภัยเพื่อวิเคราะห์หา

- สาเหตุของการยกเลิกการใช้บริการ
- กลุ่มลูกค้าที่มีแนวโน้มที่จะเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น

ทั้งนี้ เพื่อเป็นแนวทางให้ผู้ประกอบการธุรกิจประกันภัยทราบถึงสาเหตุของการยกเลิกการใช้บริการ เพื่อให้สามารถหากลยุทธ์ทางการตลาดมาใช้ในรักษาลูกค้ากลุ่มดังกล่าวไว้

3.2 การเตรียมข้อมูล (Prepare Data)

ในธุรกิจประกันภัยข้อมูลที่นำมาใช้ในการทำดาต้าไมนิ่ง เป็นข้อมูลเกี่ยวกับการทำประกันภัยรถยนต์ของลูกค้า สามารถแบ่งได้เป็นประเภทหลัก ๆ ดังนี้

- ข้อมูลสัญญาการทำประกันภัย ประกอบด้วยรายละเอียดเกี่ยวกับ วันที่เริ่มต้นของกรมธรรม์ วันที่สิ้นสุดของกรมธรรม์ เลขที่กรมธรรม์ ประเภทของกรมธรรม์ ทุนประกันภัย เบี้ยประกันภัย ค่าอากรแสตมป์ ค่าภาษีส่วนลดค่าเบี้ย ประกัน สถานะกรมธรรม์ เป็นต้น

- ข้อมูลความคุ้มครองของกรมธรรม์ ประกอบด้วย ประเภทความคุ้มครอง รายละเอียด และจำนวนเงินเอาประกันภัยสำหรับแต่ละประเภทความคุ้มครอง เป็นต้น
- ข้อมูลรถประกัน ประกอบด้วย รุ่นของรถยนต์ ปีที่ผลิต เลขทะเบียนรถ เลขที่ตัวถัง เลขเครื่องยนต์ ขนาดเครื่องยนต์ เป็นต้น
- ข้อมูลการแจ้งอุบัติเหตุ ประกอบด้วย เลขที่กรมธรรม์ วันที่แจ้งเหตุ วันเวลาที่เกิดเหตุ สถานที่เกิดเหตุ เป็นต้น
- ข้อมูลการเรียกค่าสินไหมทดแทน ประกอบด้วย เลขที่เคลม วันที่เคลม ค่าสินไหมตั้งจ่าย ค่าสินไหมทดแทนจ่าย เป็นต้น
- ข้อมูลลูกค้า ประกอบด้วย ชื่อ ที่อยู่ อายุ เพศ อาชีพ และรายละเอียดต่างๆ ของลูกค้า เป็นต้น

โดยข้อมูลที่เลือกมาใช้ในการทำดาต้าไมนิ่งมีรายละเอียดดังต่อไปนี้

ชื่อข้อมูล	ประเภทของข้อมูล	ความหมาย
Age	Number	อายุของลูกค้า
Sex	Varchar2 (1)	เพศของลูกค้า
Year	Number	จำนวนปีที่ใช้บริการ
Sum_Insured	Number	ทุนประกันภัย
Premium	Number	เบี้ยประกันภัย
Disc_Perc	Number	ส่วนลดค่าเบี้ยประกันภัย (%)
Car_Brand	Varchar2 (15)	ยี่ห้อรถ
Car_year	Number	อายุรถ
No_of_claim	Number	จำนวนครั้งในการเรียกร้องค่าสินไหม
Claim_Amount	Number	จำนวนเงินค่าสินไหมทดแทน
Status	Text	สถานะการยกเลิกกรมธรรม์ (Active/Cancel)

ตารางที่ 3.1 ตารางข้อมูล

รหัส	ความหมาย
F	เพศหญิง
M	เพศชาย

ตารางที่ 3.2 รายการเพศของลูกค้า

รายการข้อมูล
AUDI
BENZ
BMW
FORD
HONDA
ISUZU
MAZDA
MITSUBISHI
NISSAN
TOYOTA
VOLKSWAKEN
VOLVO

ตารางที่ 3.3 รายการยี่ห้อรถ

โดย Target Attribute ที่ใช้ในการสร้างแบบจำลองพยากรณ์คือ สถานะการยกเลิกกรมธรรม์

ซึ่ง Cancel หมายถึง ลูกค้ายกเลิกกรมธรรม์

Active หมายถึง ลูกค้าไม่ได้ยกเลิกกรมธรรม์

จากนั้นต้องนำข้อมูลที่คัดเลือกมาทำความสะอาดเพื่อจัดการกับ Noisy data และ Missing values แต่เนื่องจากข้อมูลที่เลือกมามีความสมบูรณ์อยู่แล้ว ไม่พบทั้ง Noisy data และ Missing values จึงไม่มีการกระทำใด ๆ กับข้อมูล

จากนั้นทำการแบ่งข้อมูลเป็น 2 ชุด โดยชุดแรกใช้สำหรับทำการ Test เพื่อสร้างแบบจำลองพยากรณ์จำนวน 800 รายการ และชุดที่ 2 ใช้สำหรับทดสอบความถูกต้องของแบบจำลองที่ได้จำนวน 200 รายการ

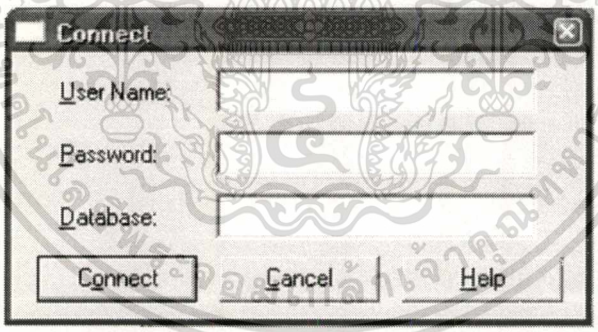
เมื่อได้ทำการจัดเตรียมข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดไปเป็นการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น ได้เลือกใช้โปรแกรม Oracle Form 6iในการพัฒนาโปรแกรม

3.3 การจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่ทำการพัฒนาขึ้น

ในส่วนของการจัดกลุ่มข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้นประกอบด้วย 4 ขั้นตอน ดังนี้

3.3.1 ติดต่อฐานข้อมูล

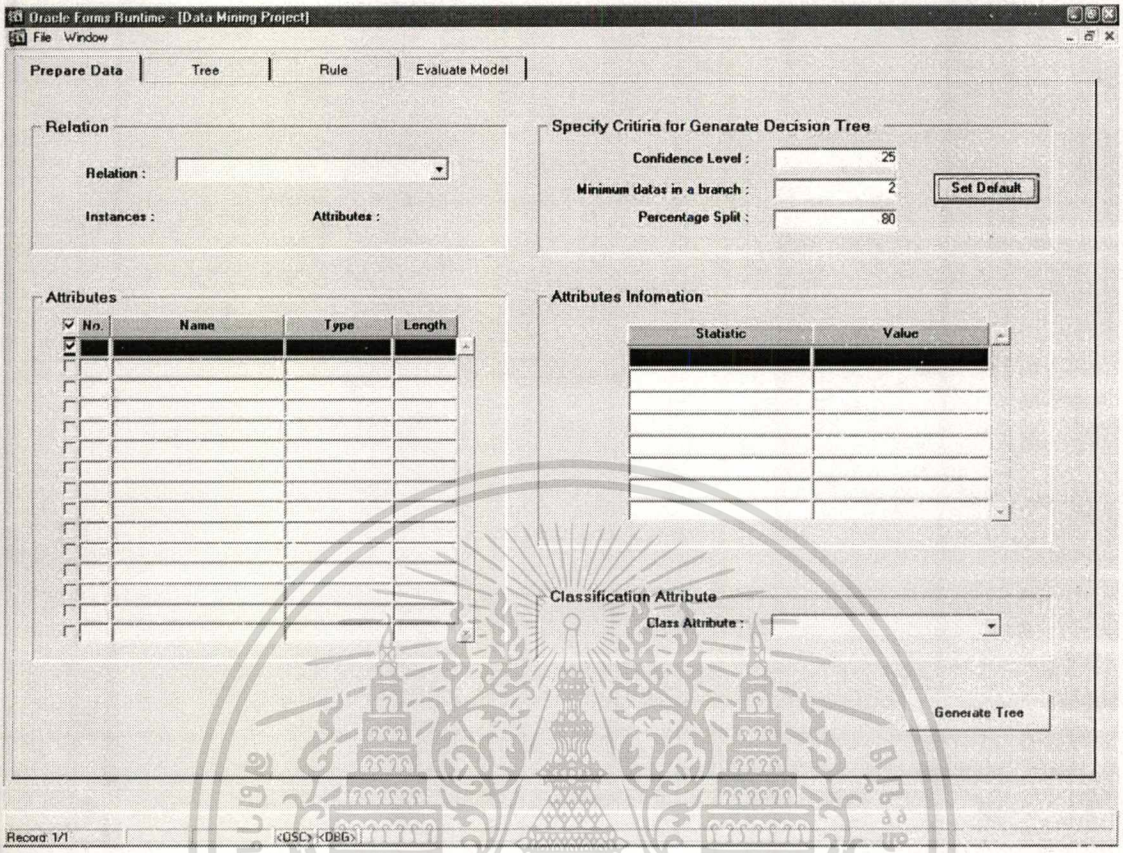
ในการทำการวิเคราะห์ข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น จะต้องเริ่มจากการติดต่อฐานข้อมูลก่อน โดยการ log on เข้าสู่ระบบที่ต้องการทำการวิเคราะห์ โดยในโครงงานนี้ข้อมูลที่ใช้จะอยู่ในรูปแบบของ Table ที่อยู่ในฐานข้อมูล Oracle ซึ่งโปรแกรมนี้สามารถรองรับการติดต่อฐานข้อมูลของ Oracle โดยตรง



รูปที่ 3.1 หน้าจอ log on เพื่อเข้าสู่ระบบ

3.3.2 คัดเลือกข้อมูล

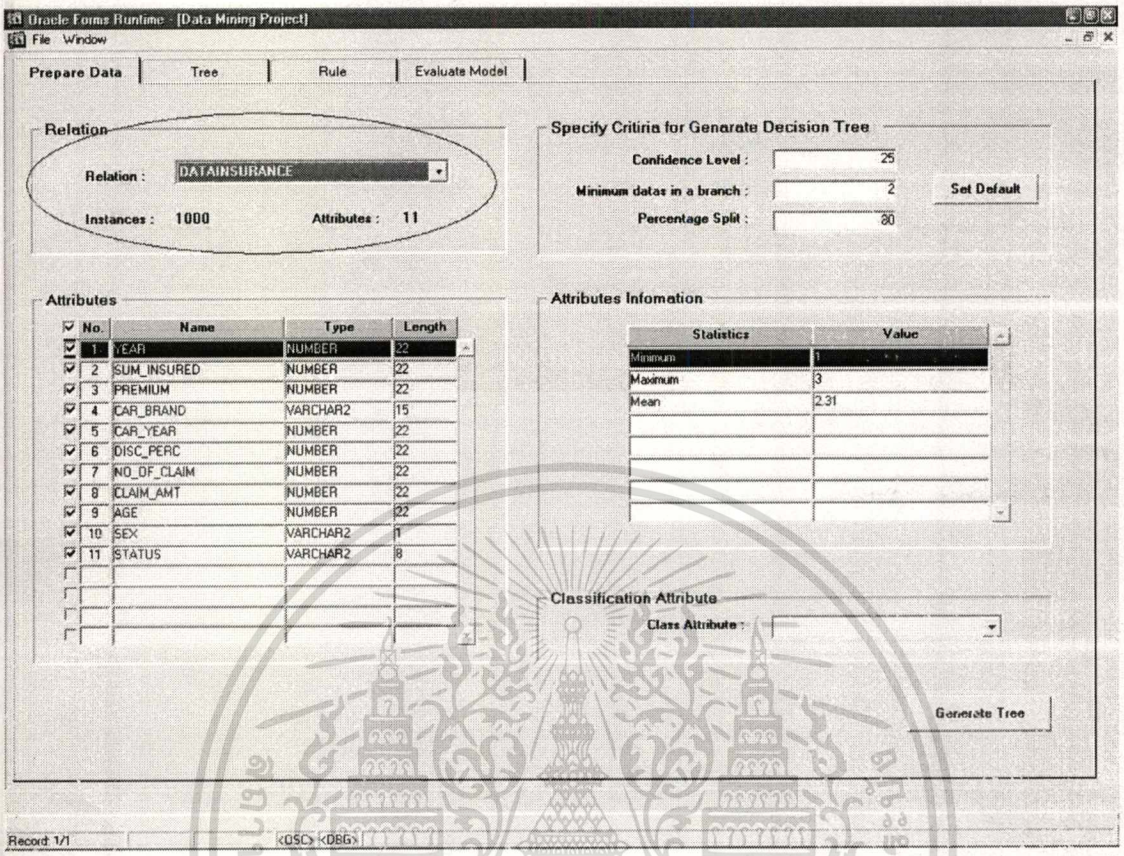
หลังจากเข้าสู่ระบบแล้ว จะปรากฏหน้าจอดังรูปที่ 3.2 เพื่อเลือกตารางข้อมูลที่ต้องการทำการวิเคราะห์จากฐานข้อมูล



รูปที่ 3.2 หน้าจอหลักเพื่อเลือกข้อมูล

หน้าจอหลักเพื่อเลือกข้อมูลประกอบด้วย 5 ส่วนหลัก ๆ ดังนี้

- Relation : ใช้ในการเลือกตารางข้อมูลที่ต้องการวิเคราะห์
- Specify Criteria for Generate Decision Tree : กำหนดเงื่อนไขในการสร้าง Tree
- Attributes : ใช้ในการเลือก Attributeที่จะนำมาวิเคราะห์
- Attributes Information : ใช้แสดงคุณสมบัติต่าง ๆ ของ Attribute
- ClassificationAttributes : เพื่อระบุ Attribute เป้าหมายในการวิเคราะห์

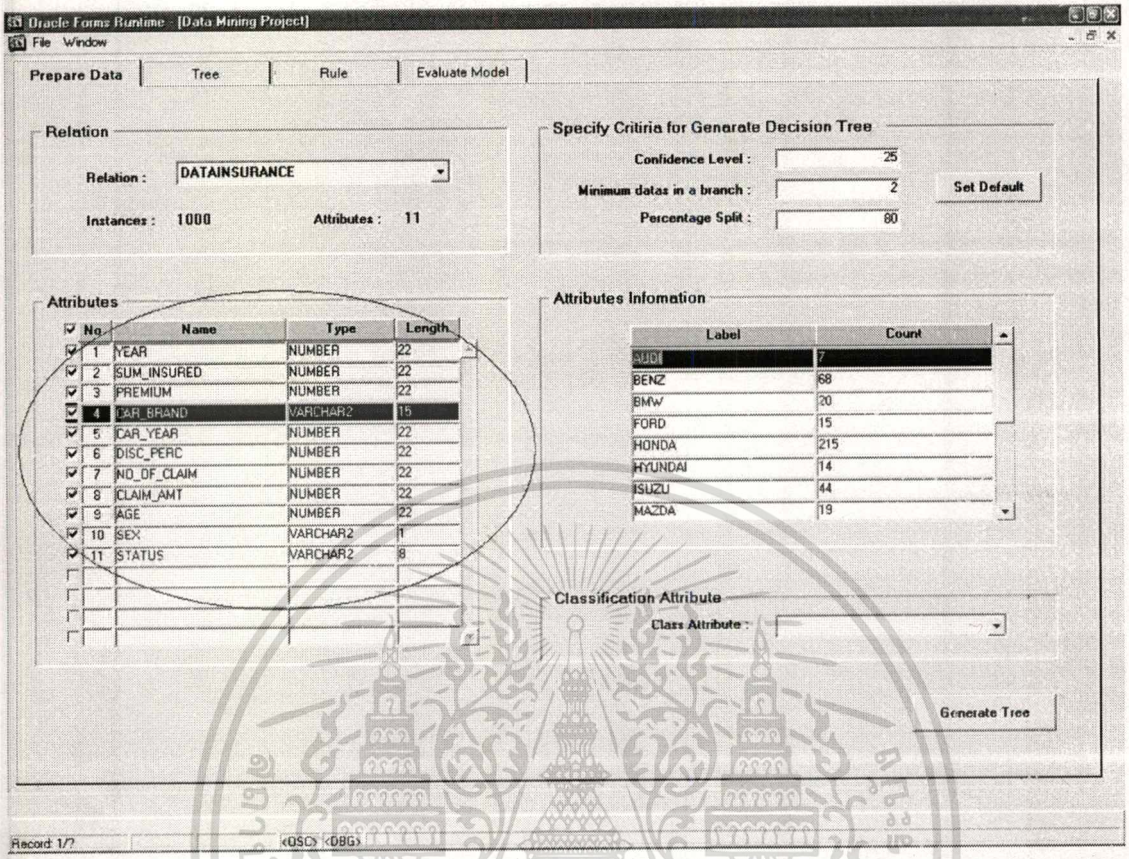


รูปที่ 3.3 หน้าจอแสดงการเลือกตารางข้อมูล

เมื่อเลือกตารางข้อมูลที่ต้องการวิเคราะห์แล้ว ระบบจะแสดงรายละเอียดของตารางที่เลือก เช่น Instance : 1000 หมายถึง จำนวนข้อมูลทั้งหมดของตารางที่เลือก

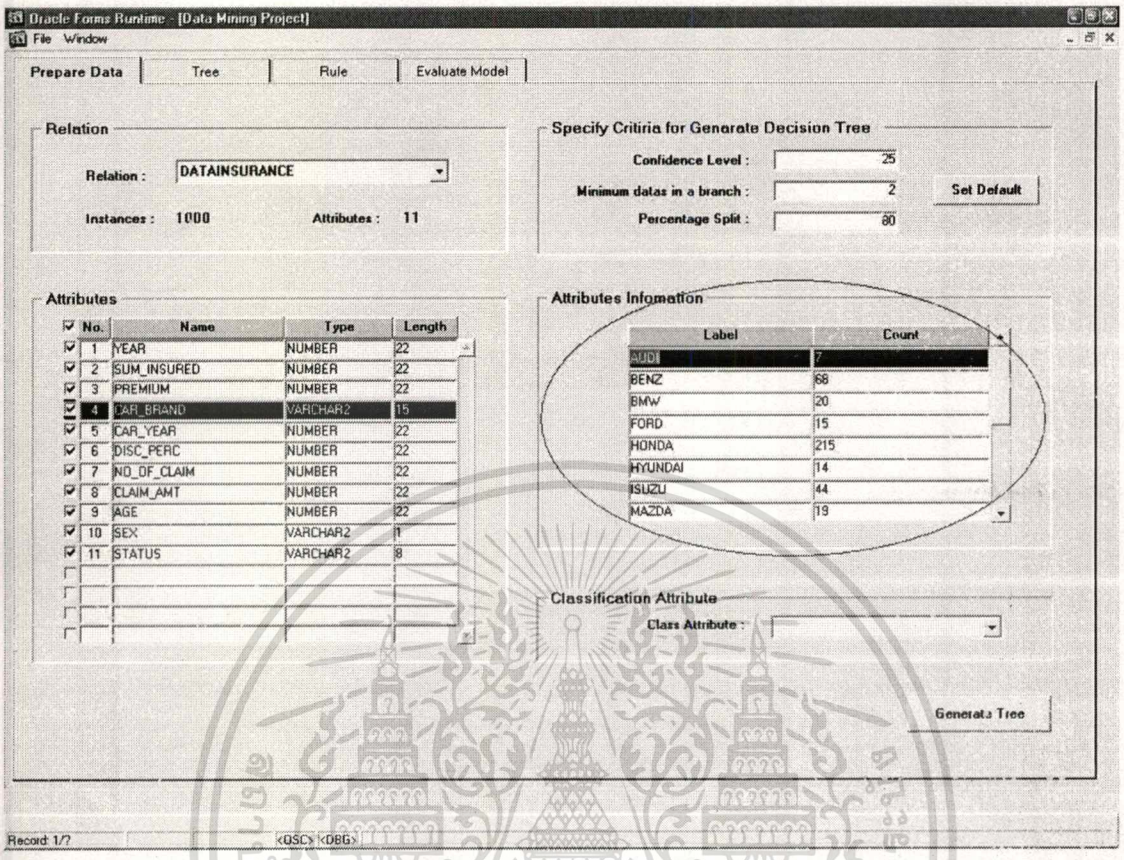
Attributes : 11 หมายถึง จำนวน Attribute ทั้งหมดของตารางที่เลือก

นอกจากนี้ ระบบจะแสดงรายการ Attribute ทั้งหมดของตารางข้อมูลที่เลือกนั้น รวมแสดงทั้งประเภทและความยาวของ Attribute ต่างๆ เพื่อให้ผู้ใช้สามารถเลือกได้ว่าต้องการเลือก Attribute ใดบ้างมาวิเคราะห์ โดย Check ที่ Check box ที่อยู่ด้านหน้าของแต่ละรายการ Attribute ดังรูปที่ 3.4

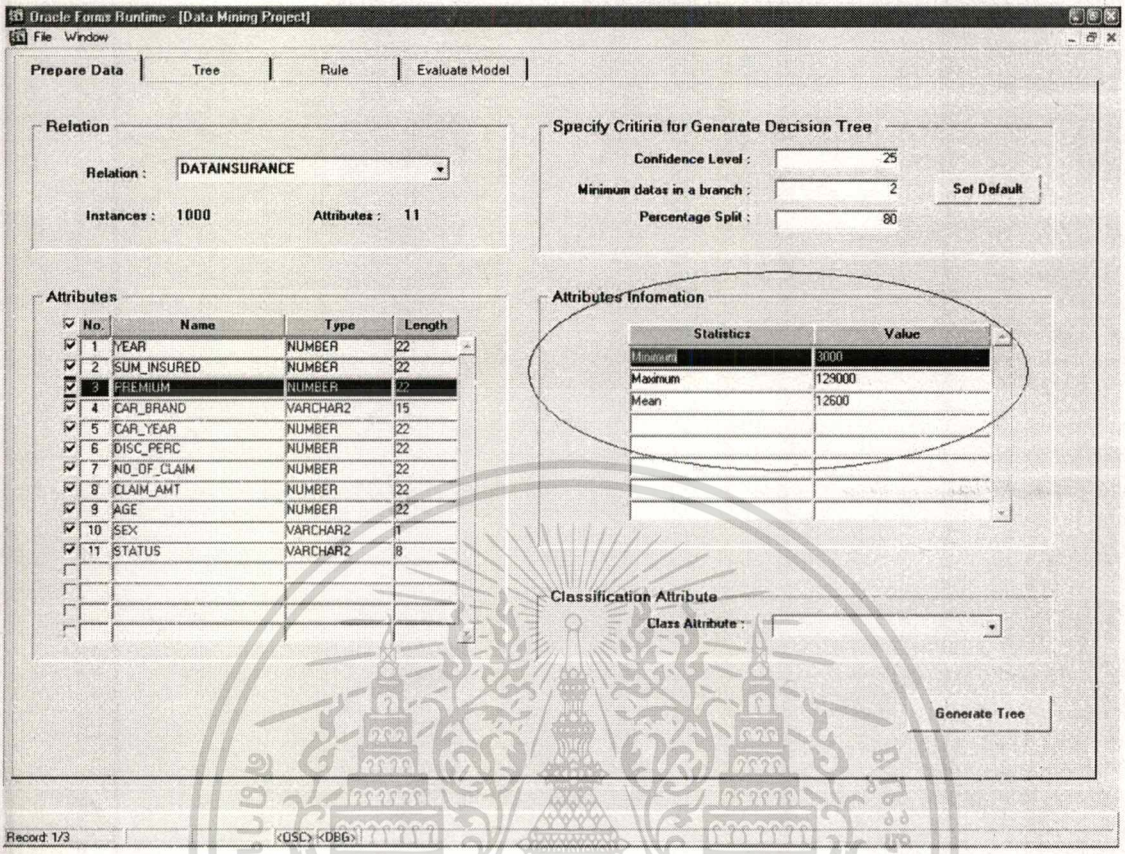


รูปที่ 3.4 หน้าจอแสดงรายละเอียดของ Attribute

นอกจากนี้ ระบบจะแสดงรายละเอียดของข้อมูลในแต่ละ Attribute โดยถ้าเป็น Attribute ที่มีประเภทของข้อมูลเป็นข้อความ (Varchar2) จะแสดงค่าของข้อมูลและค่าความถี่ที่เกิดขึ้นของข้อมูลใน Attribute นั้น ๆ ส่วน Attribute ที่มีประเภทของข้อมูลเป็นตัวเลข (Number) จะแสดงค่าต่ำสุด ค่าสูงสุด และค่าเฉลี่ยของข้อมูล แสดงตัวอย่างหน้าจอดังรูปที่ 3.5 และ 3.6



รูปที่ 3.5 หน้าจอแสดงรายละเอียดข้อมูลของ Attribute ประเภทข้อความ



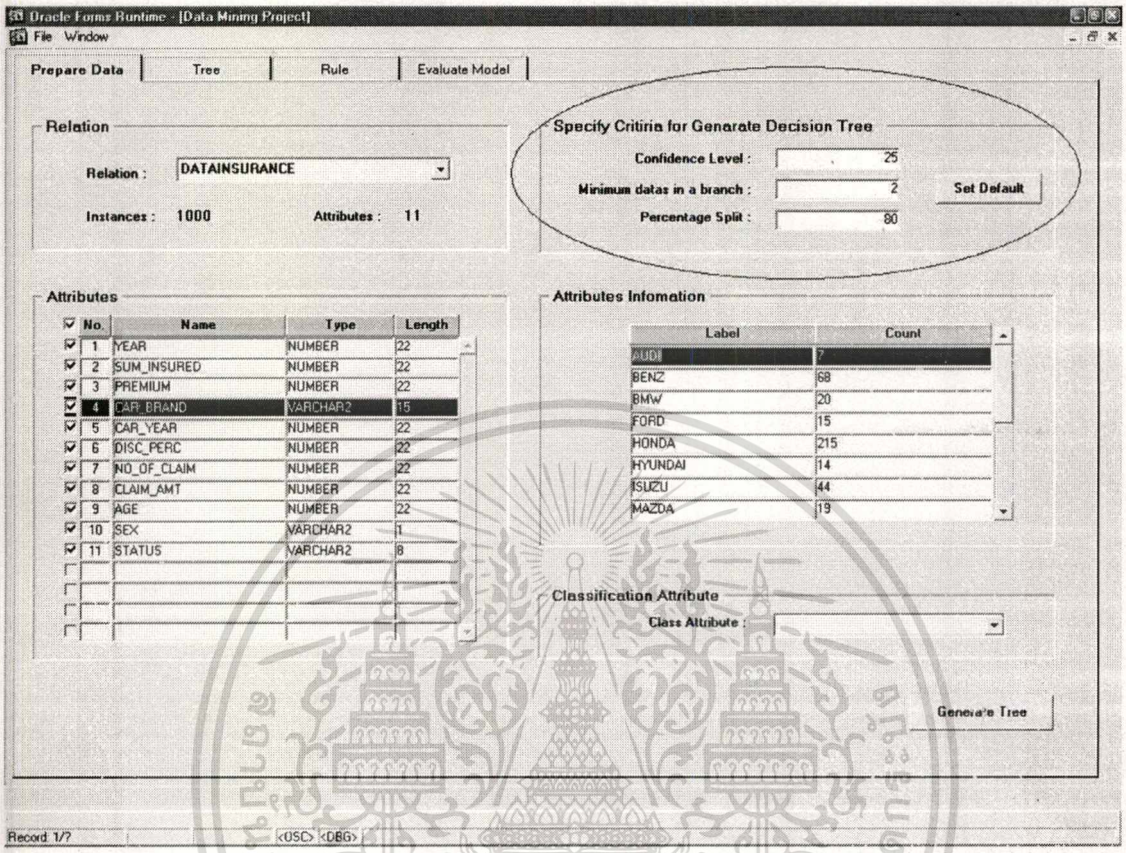
รูปที่ 3.6 หน้าจอแสดงรายละเอียดข้อมูลของ Attribute ประเภทตัวเลข

หลังจากเลือก Attribute ที่ต้องการวิเคราะห์แล้ว จะต้องกำหนด Classification Attribute ด้วยเพื่อระบุว่า Attribute ใดเป็น Attribute ที่เป็นเป้าหมายของการวิเคราะห์ หลังจากระบุข้อมูลครบถ้วนแล้ว สามารถคลิกที่ปุ่ม Generate Tree เพื่อทำการสร้าง Tree สำหรับการวิเคราะห์ข้อมูลดังกล่าว

3.3.3 การกำหนดเงื่อนไขให้กับโปรแกรม

เมื่อทำการคัดเลือกข้อมูลที่จะให้โปรแกรม ก่อนที่จะสร้าง Tree ที่ใช้ในการวิเคราะห์ข้อมูล ผู้ใช้สามารถกำหนดค่า Confidence Level, Minimum record และ Percentage Split เพื่อกำหนดเงื่อนไขในการสร้าง Decision Tree และกฎได้ ดังรูปที่ 3.7 โดยระบบจะกำหนดค่า default ต่าง ๆ ดังนี้

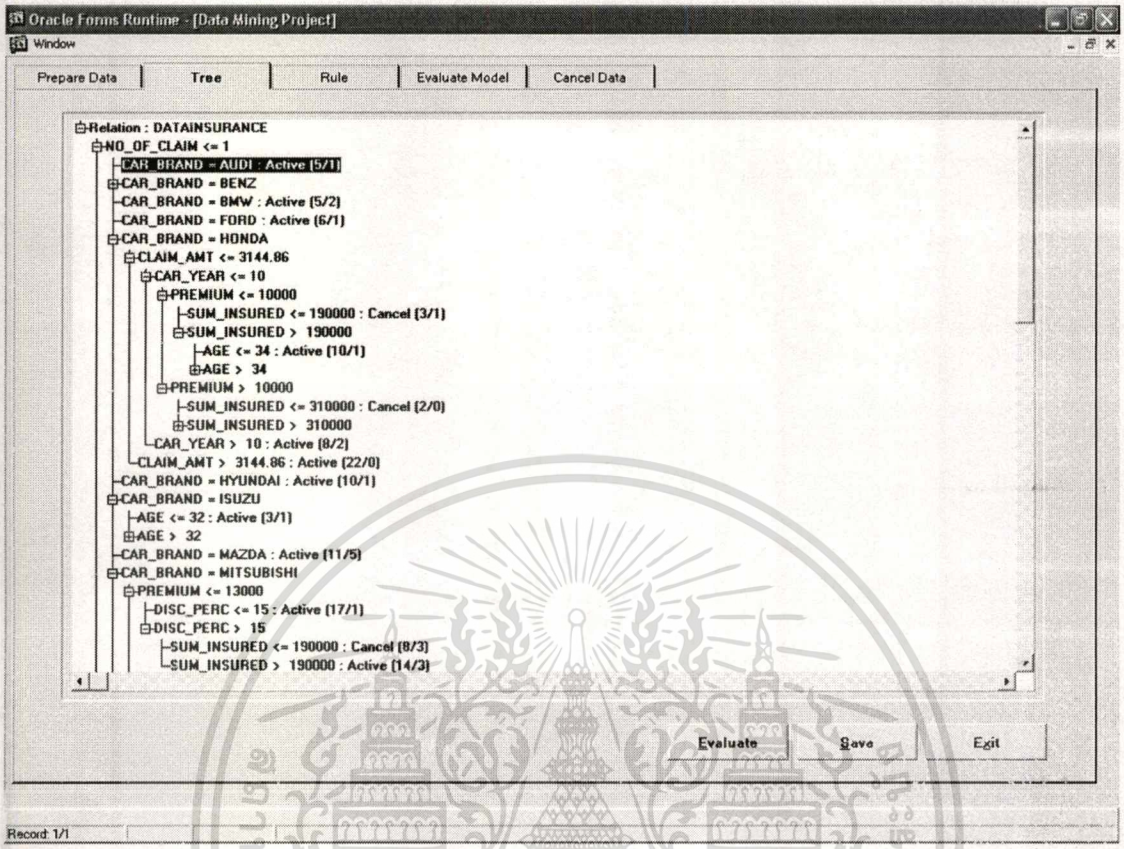
Confidence Level	:	25%
Minimum datas in a bracnch	:	2
Percentage Split	:	80%



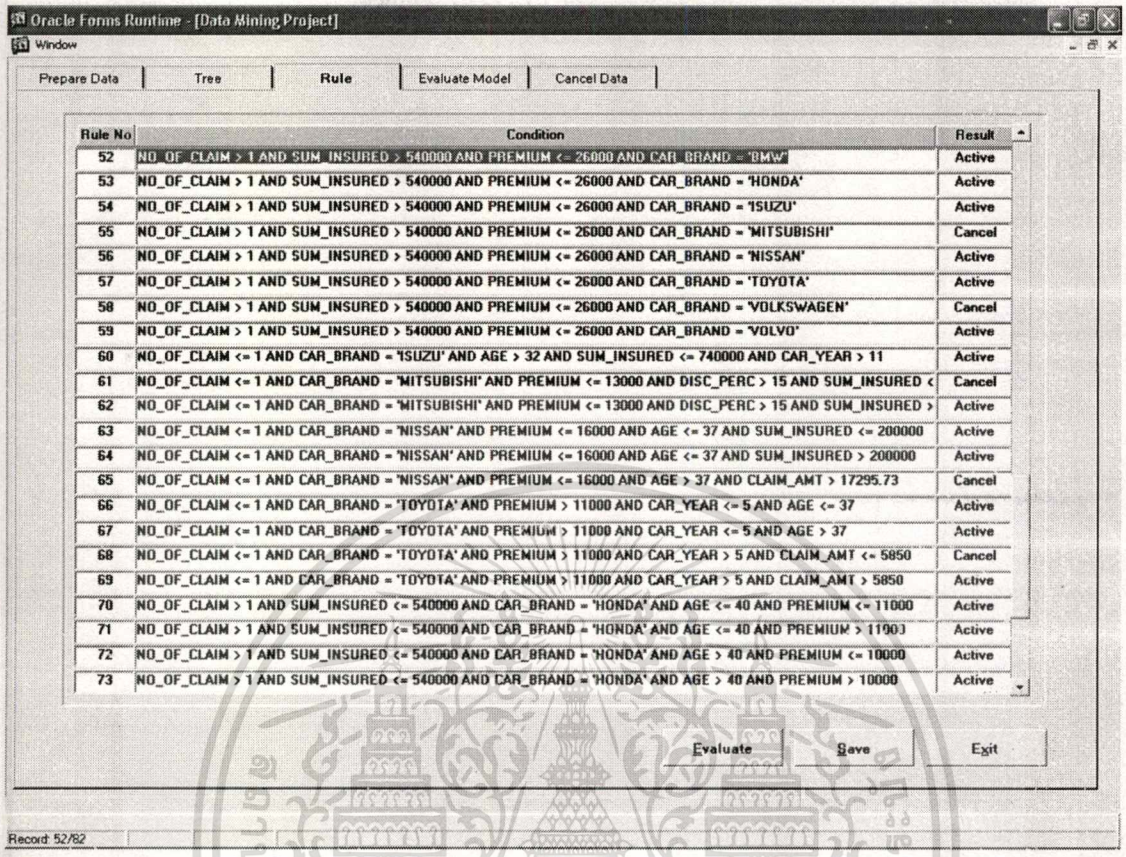
รูปที่ 3.7 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

3.3.4 การแสดงผล

เมื่อโปรแกรมทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงในรูปที่ 3.8 และ 3.9 ตามลำดับ โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใด และข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภทที่ (Class) ที่ข้อมูลส่วนใหญ่ใน Node นั้นตกอยู่ โดยสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้ โดยการเลือกปุ่ม Save



รูปที่ 3.8 หน้าจอแสดงผลลัพธ์ในรูป Decision Tree



รูปที่ 3.9 หน้าจอแสดงผลลัพธ์ในรูปของกฎ

3.3.5 การทดสอบความถูกต้องของแบบจำลองพยากรณ์

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูล Train แล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากน้อยเพียงใด โดยการนำข้อมูลที่แบ่งไว้สำหรับการ Test กับแบบจำลองพยากรณ์ที่ได้ โดยเลือกปุ่ม Evaluate เพื่อให้ระบบแสดงความถูกต้องของแบบจำลองโดยใช้ข้อมูล Train และเลือกปุ่ม Test เพื่อทำการทดสอบแบบจำลองโดยใช้ข้อมูล Test โดยความถูกต้องของระบบสามารถคำนวณได้จาก

Accuracy = Sensitive x (pos/(pos + neg)) + Specificity x (neg/(pos + neg))

โดย Sensitive = t_pos/pos

Specificity = t_neg/neg

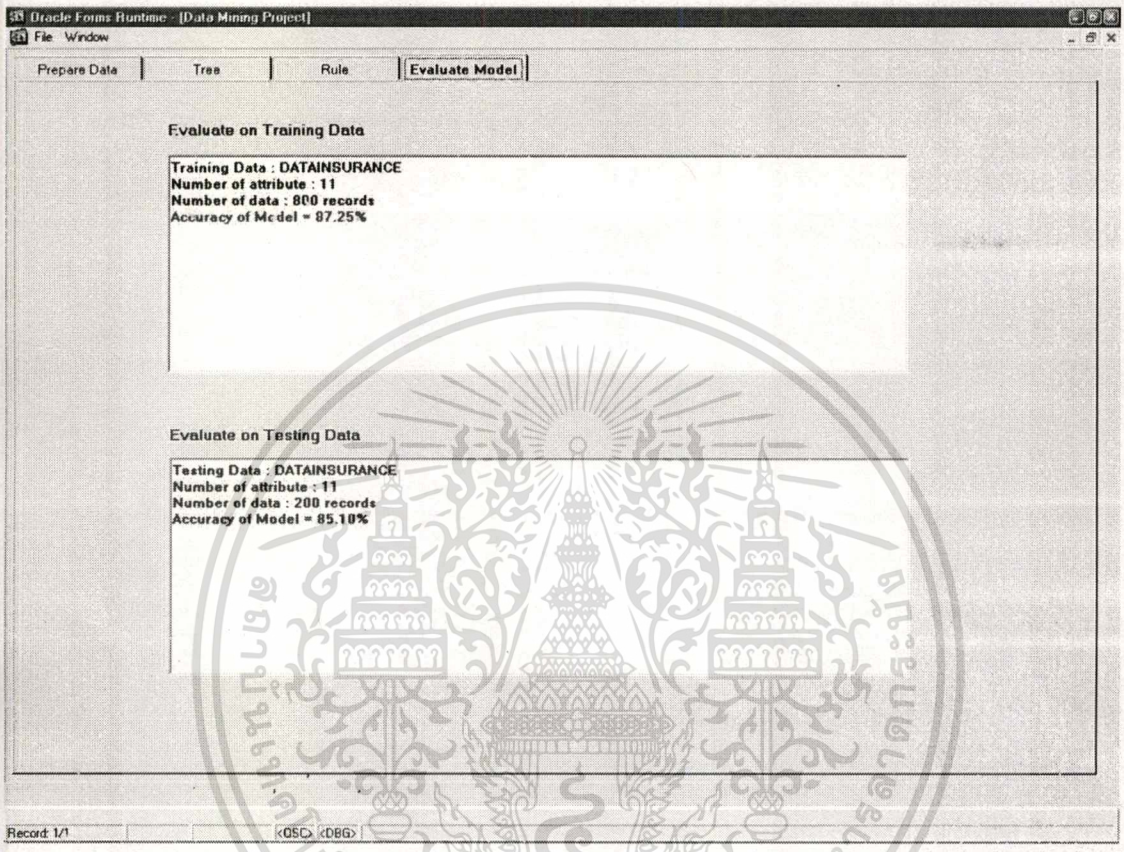
T_pos แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive ที่สามารถทำนายกลุ่มได้ถูกต้อง

Pos แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Positive

t_neg แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative ที่สามารถทำนายกลุ่มได้ถูกต้อง

neg แทนจำนวนข้อมูลตัวอย่างที่อยู่ในกลุ่ม Negative

โดยหลังจากสั่งให้ระบบทำการตรวจสอบความถูกต้องของแบบจำลองพยากรณ์ ระบบจะแสดงผลการทดสอบดังรูปที่ 3.10



รูปที่ 3.10 หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

3.4 การวิเคราะห์ผลดำเนินงาน

ผลจากการทดสอบแบบจำลองพยากรณ์ที่ได้โดยใช้ข้อมูลทดสอบจำนวน 200 รายการ สามารถวัดความถูกต้องของแบบจำลองการจัดหมวดหมู่สำหรับการทำนายข้อมูลในแต่ละกลุ่ม โดยสรุปรายการที่แบบจำลองสามารถทำนายได้ถูกต้องคิดเป็น 85.1 เปอร์เซ็นต์ของข้อมูลทั้งหมด และจากการวิเคราะห์พบว่าความผิดพลาดจำนวน 14.9 เปอร์เซ็นต์นี้ อาจเกิดจากข้อมูลมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ ซึ่งอาจส่งผลให้ข้อมูลไม่สามารถเป็นตัวแทนทางสถิติได้ทั้งหมด

บทที่ 4
สรุปผลการศึกษาและข้อเสนอแนะ

4.1 สรุปผลการศึกษา

โครงงานพัฒนาระบบฉบับนี้มีวัตถุประสงค์หลักเพื่อที่จะนำเสอนและประยุกต์ใช้ดาต้าไมนิ่งในธุรกิจประกันภัย ซึ่งดาต้าไมนิ่งนั้นเป็นกระบวนการท่ใช้ค้นหาข้อมูลที่มีประโยชน์จากฐานข้อมูลขนาดใหญ่ เพื่อนำมาช่วยในการตัดสินใจ ซึ่งวิธีการปัญหาด้วยดาต้าไมนิ่งนั้นมีอยู่ด้วยกันหลายรูปแบบขึ้นอยู่กับวัตถุประสงค์ของการทำงาน โดยในโครงงานนี้ได้เสนอเทคนิคการสร้างแบบจำลองพยากรณ์ (Predictive Modeling) เพื่อวิเคราะห์ลักษณะและพฤติกรรมของลูกค้า และระบุสาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย โดยมีวัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้เป็นแนวทางในการวางกลยุทธ์ทางการตลาดและจัดทำโปรโมชั่นที่น่าสนใจเพื่อรักษาลูกค้ากลุ่มดังกล่าวไว้ โดยใช้อัลกอริทึม C 4.5 ซึ่งเป็นอัลกอริทึมของ Classification Tree ที่มีการใช้งานกันอย่างกว้างขวาง อันเนื่องมาจากความมีประสิทธิภาพในการแก้ปัญหา นอกจากนี้ยังมีความยืดหยุ่นและเข้าใจได้ง่าย

ผลจากการศึกษาทำให้ได้ระบบที่ใช้สำหรับจัดกลุ่มของข้อมูล ซึ่งสามารถนำไปประยุกต์ใช้ได้กับทุกธุรกิจ และจากการนำข้อมูลเข้าไปสร้างแบบจำลองพยากรณ์และทำการทดสอบพบว่าแบบจำลองพยากรณ์ที่ได้มีความถูกต้องคิดเป็น 85.1 เปอร์เซ็นต์ของข้อมูลที่ใช้ทดสอบ และจากการวิเคราะห์พบว่าความผิดพลาดจำนวน 14.9 เปอร์เซ็นต์นี้ อาจเกิดจากข้อมูลมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ ซึ่งอาจส่งผลให้ข้อมูลไม่สามารถเป็นตัวแทนทางสถิติได้ทั้งหมด

4.2 ข้อเสนอแนะ

ระบบที่พัฒนาขึ้นนี้สามารถที่จะนำไปใช้กับข้อมูลในธุรกิจอื่นได้ เนื่องจากไม่ได้จำกัดขอบเขตเฉพาะกับธุรกิจประกันภัยเท่านั้น

ข้อมูลที่นำมาทดสอบนี้อาจมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้ ทั้งนี้ผู้ใช้ที่มีความเข้าใจในธุรกิจนั้น ๆ สามารถเตรียมข้อมูลที่ถูกต้องเหมาะสมมาวิเคราะห์ เพื่อให้

ระบบสามารถสร้างแบบจำลองพยากรณ์เพื่อทำนายสาเหตุการยกเลิกการใช้บริการได้ถูกต้องและแม่นยำมากขึ้น

จากการนำอัลกอริทึม C 4.5 ซึ่งเป็นอัลกอริทึมของ Classification Tree พบว่ากฎที่ได้สามารถเข้าใจได้ง่าย และสามารถบอกได้ว่าปัจจัยใดมีผลต่อการทำนายมากที่สุด แต่อย่างไรก็ตามวิธีนี้ก็มีข้อเสีย คือ ถ้าข้อมูลมีจำนวนน้อย ความผิดพลาดในการทำนายก็จะสูงขึ้น เนื่องจากเมื่อทำการแตก Tree ไปเรื่อยๆ จำนวนของข้อมูลจะลดลง ซึ่งข้อมูลจำนวนน้อยจะเป็นตัวแทนทางสถิติได้ยาก โดยยิ่งทำให้มี level มากขึ้นความน่าเชื่อถือจะยิ่งน้อยลง

บรรณานุกรม

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

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

Jiawei Han et al. 2001. **Data Mining : Concepts and Techniques**. Morgan Kaufann.

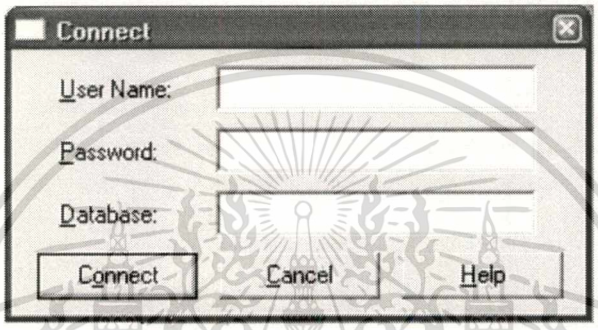
Two Crows Corporation. 1999. **Introduction to Data Mining and Knowledge Discovery.** [Online] Available : http://www.twocrows.com .

ภาคผนวก ก

คู่มือการใช้ระบบงานวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย

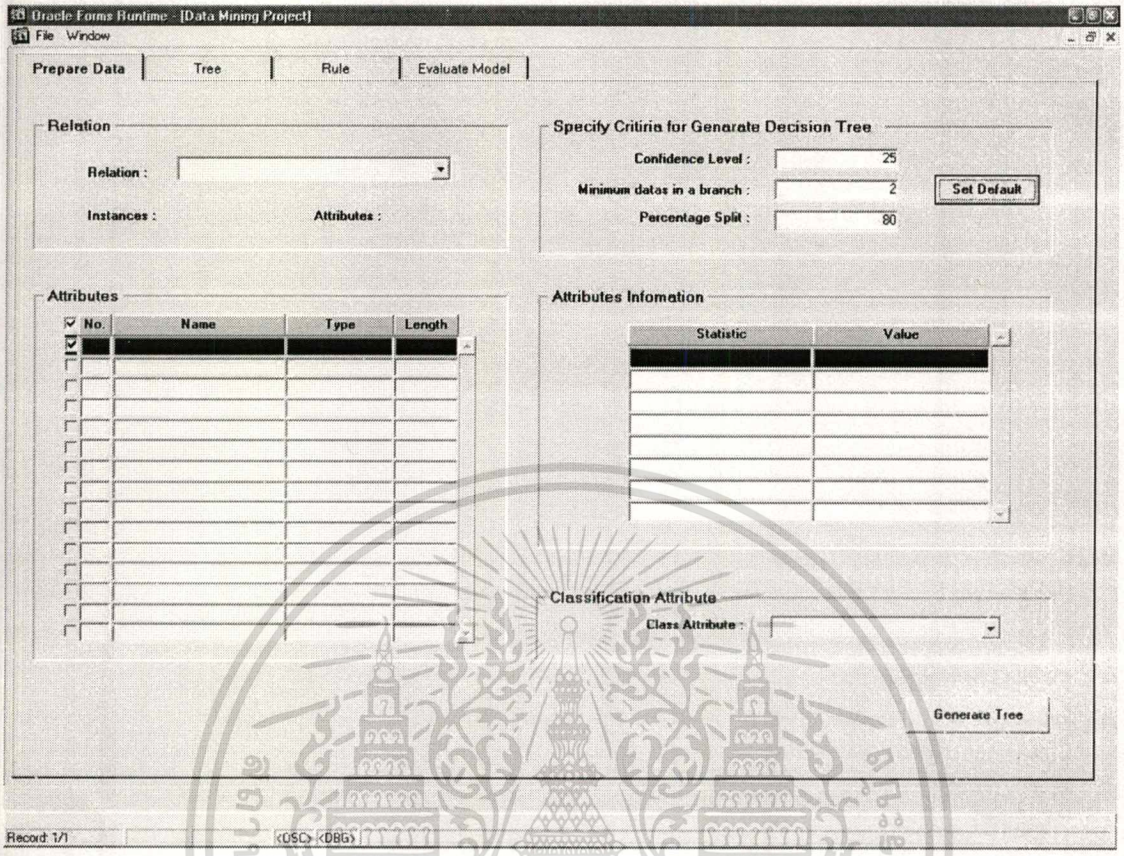
การทำงานของระบบงานวิเคราะห์สาเหตุการยกเลิกการใช้บริการในธุรกิจประกันภัย

ในการทำการวิเคราะห์ข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น จะต้องเริ่มจากการติดต่อฐานข้อมูลก่อน โดยการ log on เข้าสู่ระบบที่ต้องการทำการวิเคราะห์ โดยในโครงงานนี้ข้อมูลที่ใช้จะอยู่ในรูปแบบของ Table ที่อยู่ในฐานข้อมูล Oracle ซึ่งโปรแกรมนี้สามารถรองรับการติดต่อฐานข้อมูลของ Oracle โดยตรง



รูปที่ ก.1 หน้าจอ log on เพื่อเข้าสู่ระบบ

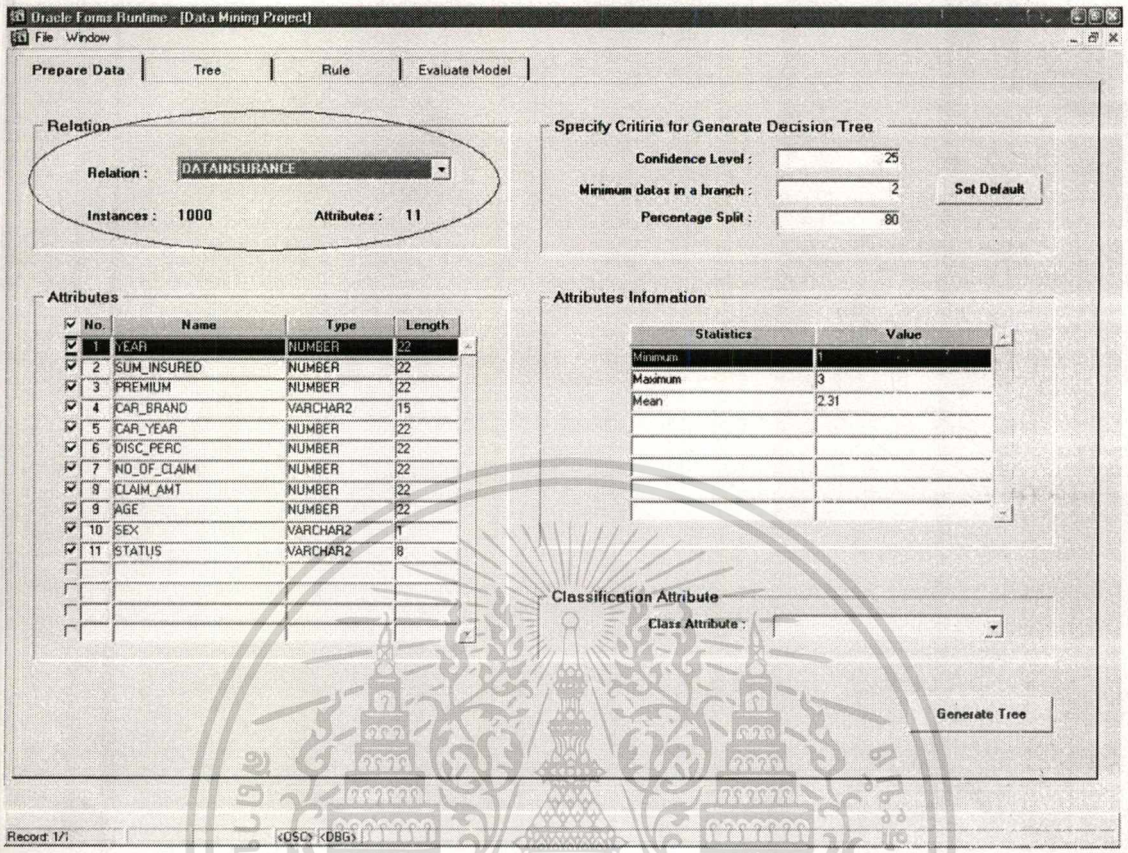
หลังจากเข้าสู่ระบบแล้ว จะปรากฏหน้าจอดังรูปที่ ก.2 เพื่อเลือกตารางข้อมูลที่ต้องการทำการวิเคราะห์จากฐานข้อมูล



รูปที่ ก.2 หน้าจอหลักเพื่อเลือกข้อมูล

หน้าจอหลักเพื่อเลือกข้อมูลประกอบด้วย 5 ส่วนหลัก ๆ ดังนี้

- Relation : ใช้ในการเลือกตารางข้อมูลที่ต้องการวิเคราะห์
- Specify Criteria for Generate Decision Tree : กำหนดเงื่อนไขในการสร้าง Tree
- Attributes : ใช้ในการเลือก Attributeที่จะนำมาวิเคราะห์
- Attributes Information : ใช้แสดงคุณสมบัติต่าง ๆ ของ Attribute
- ClassificationAttributes : เพื่อระบุ Attribute เป้าหมายในการวิเคราะห์



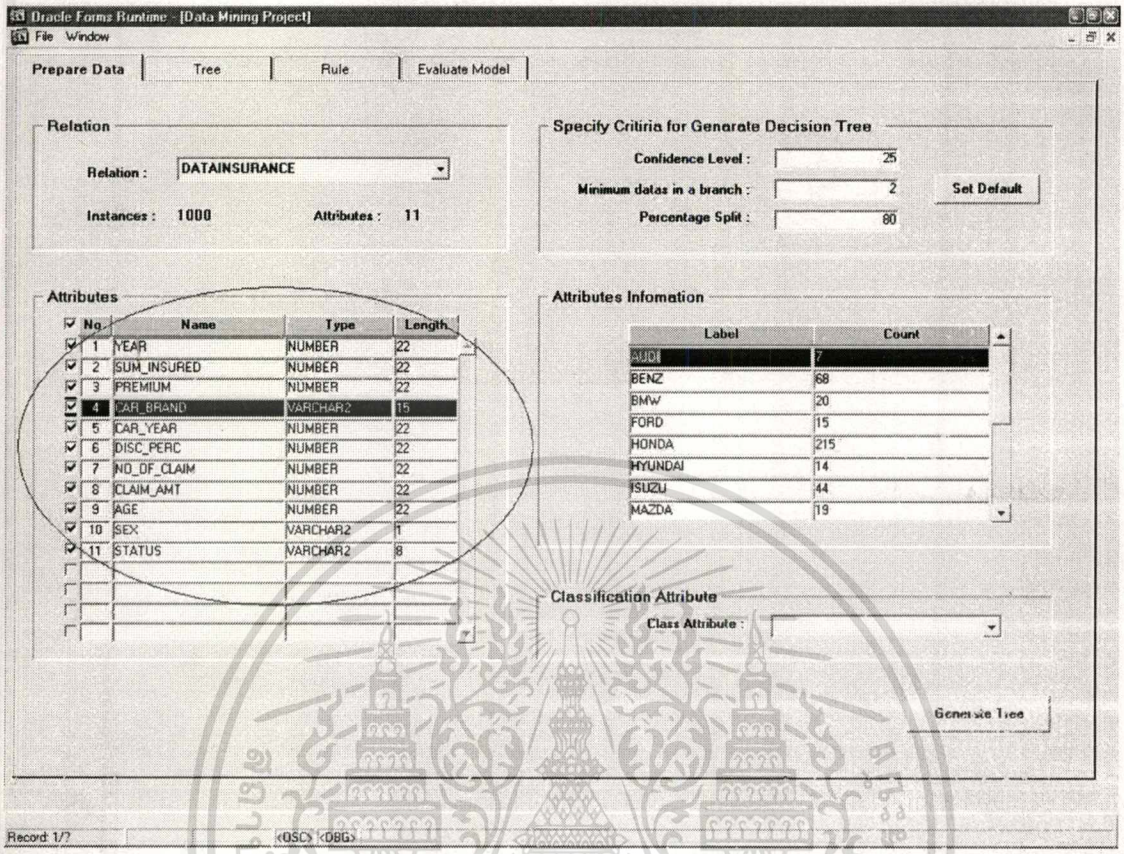
รูปที่ ก.3 หน้าจอแสดงการเลือกตารางข้อมูล

หลังจากเลือกตารางข้อมูลที่ต้องการวิเคราะห์แล้ว ระบบจะแสดงรายละเอียดของตารางที่เลือกนั้น ซึ่งประกอบด้วย จำนวนข้อมูล และ จำนวน Attribute ของตาราง

เช่น Instance : 1000 หมายถึง จำนวนข้อมูลทั้งหมดของตารางที่เลือก

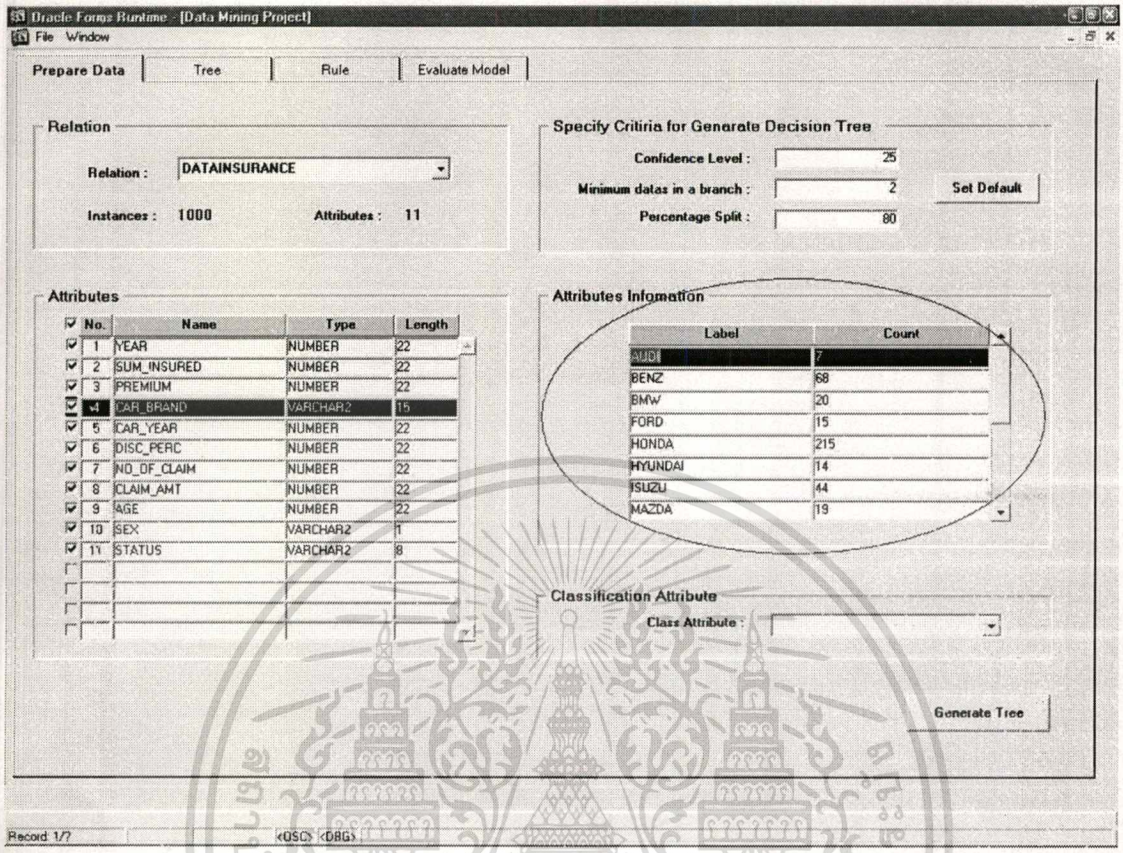
Attributes : 11 หมายถึง จำนวน Attribute ทั้งหมดของตารางที่เลือก

พร้อมทั้งแสดงรายการ Attribute ทั้งหมดของตารางข้อมูลที่เลือกนั้น รวมแสดงทั้งประเภทและความยาวของ Attribute ต่างๆ เพื่อให้ผู้ใช้สามารถเลือกได้ว่าต้องการเลือก Attribute ใดบ้างมาวิเคราะห์ โดย Check ที่ Check box ที่อยู่ด้านหน้าของแต่ละรายการ Attribute ดังรูปที่ ก.4

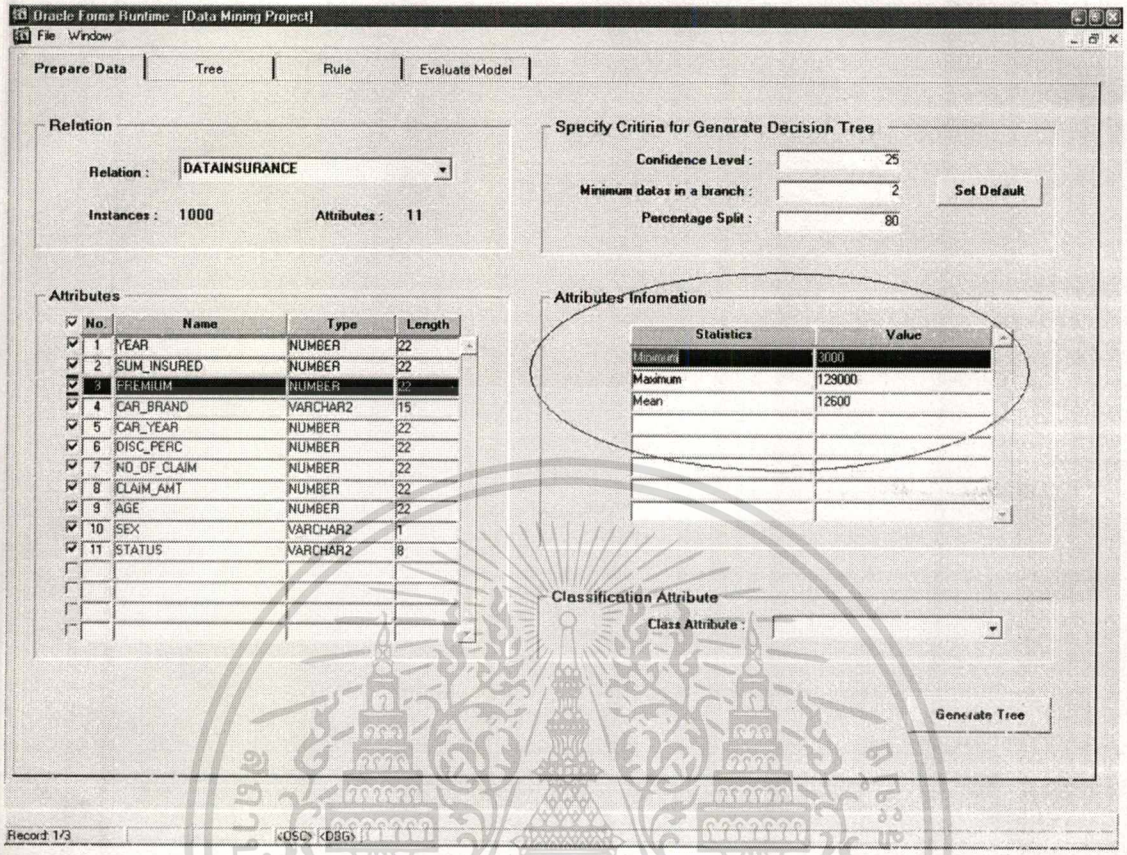


รูปที่ ก.4 หน้าจอแสดงรายละเอียดของ Attribute

นอกจากนี้ ระบบจะแสดงรายละเอียดของข้อมูลในแต่ละ Attribute โดยถ้าเป็น Attribute ที่มีประเภทของข้อมูลเป็นข้อความ (Varchar2) จะแสดงค่าของข้อมูลและค่าความถี่ที่เกิดขึ้นของข้อมูลใน Attribute นั้น ๆ ส่วน Attribute ที่มีประเภทของข้อมูลเป็นตัวเลข (Number) จะแสดงค่าต่ำสุด ค่าสูงสุด และค่าเฉลี่ยของข้อมูล แสดงตัวอย่างหน้าจอดังรูปที่ ก.5 และ ก.6



รูปที่ ก.5 หน้าจอแสดงรายละเอียดข้อมูลของ Attribute ประเภทข้อความ



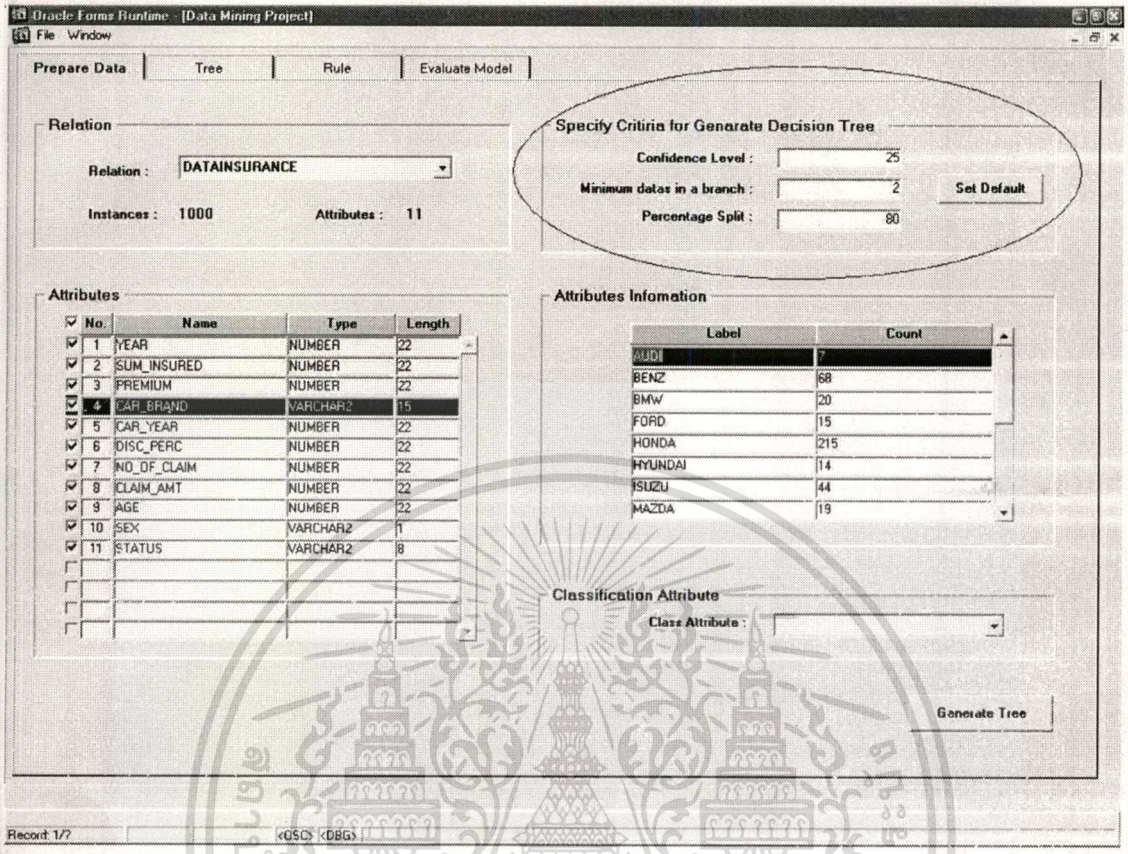
รูปที่ ก.6 หน้าจอแสดงรายละเอียดข้อมูลของ Attribute ประเภทตัวเลข

หลังจากเลือก Attribute ที่ต้องการวิเคราะห์แล้ว จะต้องกำหนด Classification Attribute ด้วยเพื่อระบุว่า Attribute ใดเป็น Attribute ที่เป็นเป้าหมายของการวิเคราะห์ หลังจากระบุข้อมูลครบถ้วนแล้ว สามารถคลิกที่ปุ่ม Generate Tree เพื่อทำการสร้าง Tree สำหรับการวิเคราะห์ข้อมูลดังกล่าว

เมื่อทำการคัดเลือกข้อมูลที่จะให้โปรแกรม ก่อนที่จะสร้าง Tree ที่ใช้ในการวิเคราะห์ข้อมูล ผู้ใช้สามารถกำหนดค่า Confidence Level, Minimum record และ Percentage Split เพื่อกำหนดเงื่อนไขในการสร้าง Decision Tree และกฎได้ ดังรูปที่ ก.7 โดยระบบจะกำหนดค่า default ต่าง ๆ ดังนี้

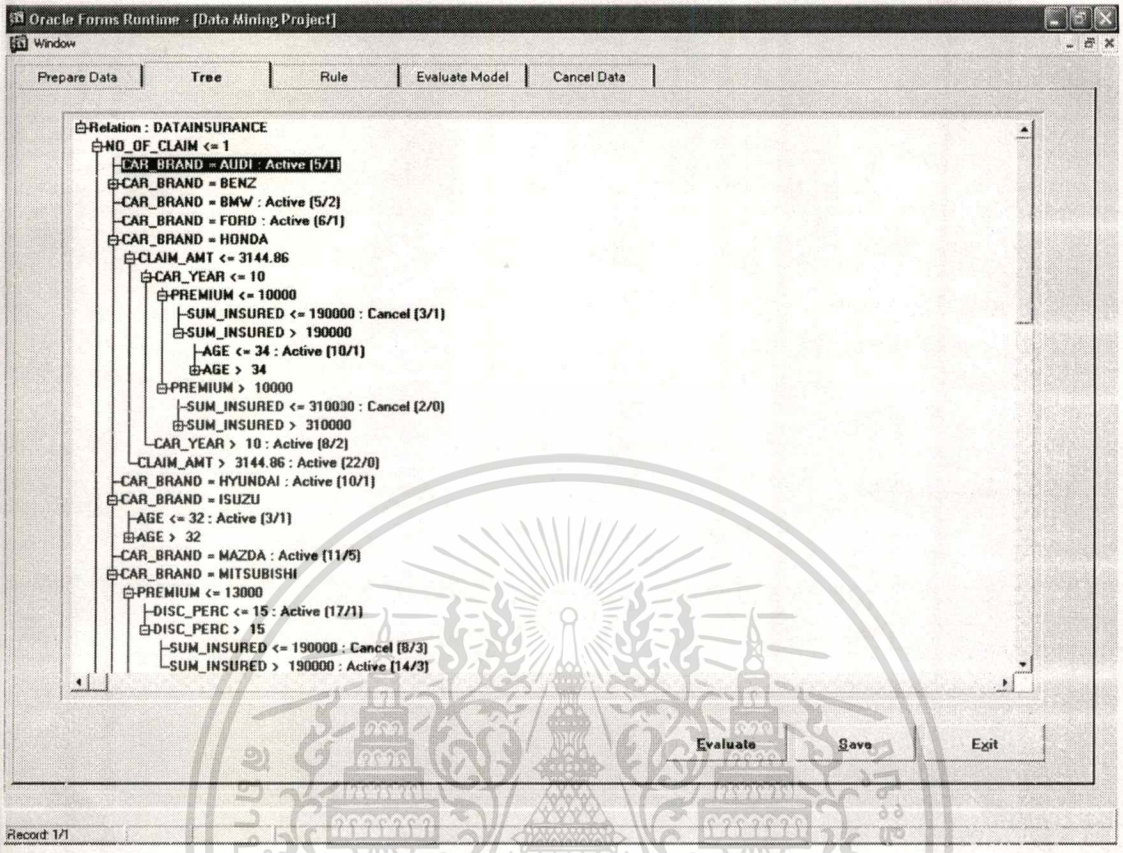
Confidence Level	:	25%
Minimum datas in a bracnch	:	2
Percentage Split	:	80%

หากต้องการใช้ค่า Default ของระบบ ให้เลือกปุ่ม Set default

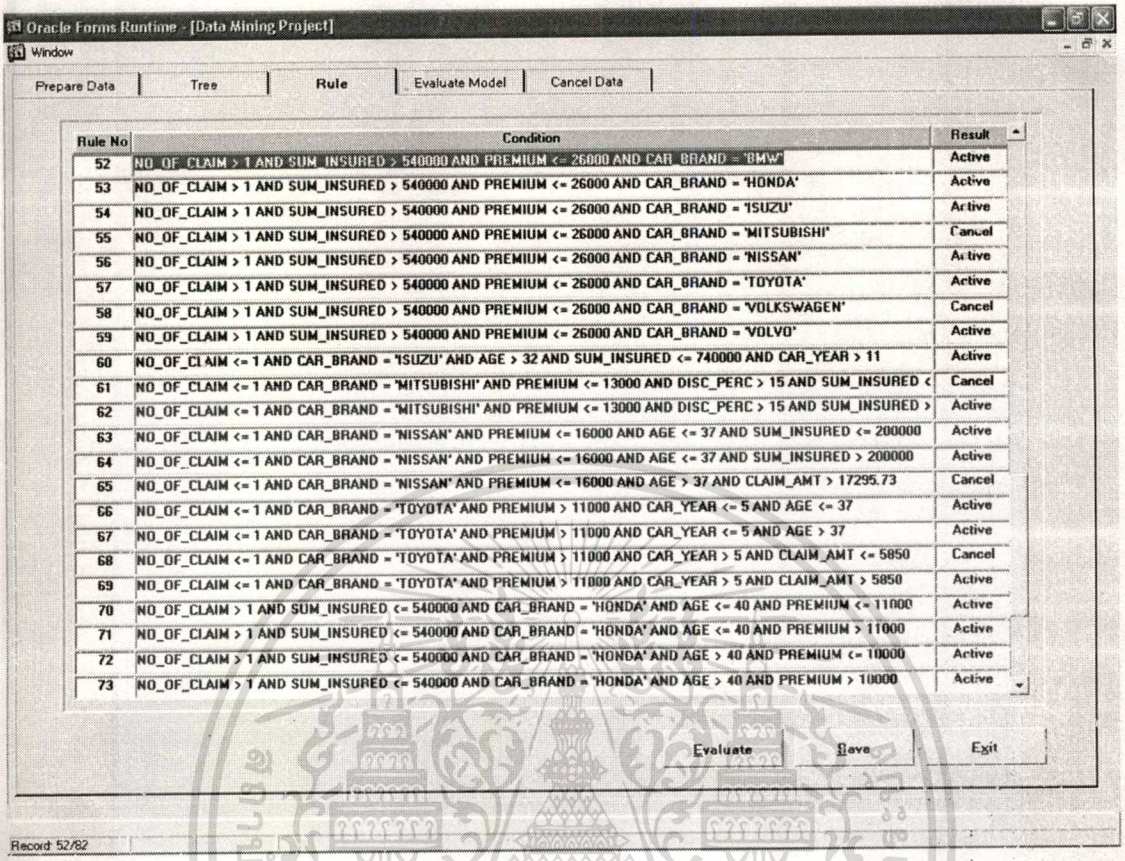


รูปที่ ก.7 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

หลังจากเลือก Generate Tree เพื่อให้โปรแกรมทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว จะแสดงผลลัพธ์เป็นโครงสร้างต้นไม้และกฎ ดังแสดงในรูปที่ ก.8 และ ก.9 ตามลำดับ โดยจะบอกว่าข้อมูลตกอยู่ในกิ่งใดเป็นจำนวนเท่าใด และข้อมูลที่ผิดพลาดเป็นจำนวนเท่าใด รวมทั้งบอกประเภทที่ (Class) ที่ข้อมูลส่วนใหญ่ใน Node นั้นตกอยู่ โดยสามารถบันทึกโครงสร้างต้นไม้นี้เก็บไว้เพื่อเรียกดูภายหลังได้ โดยการเลือกปุ่ม Save



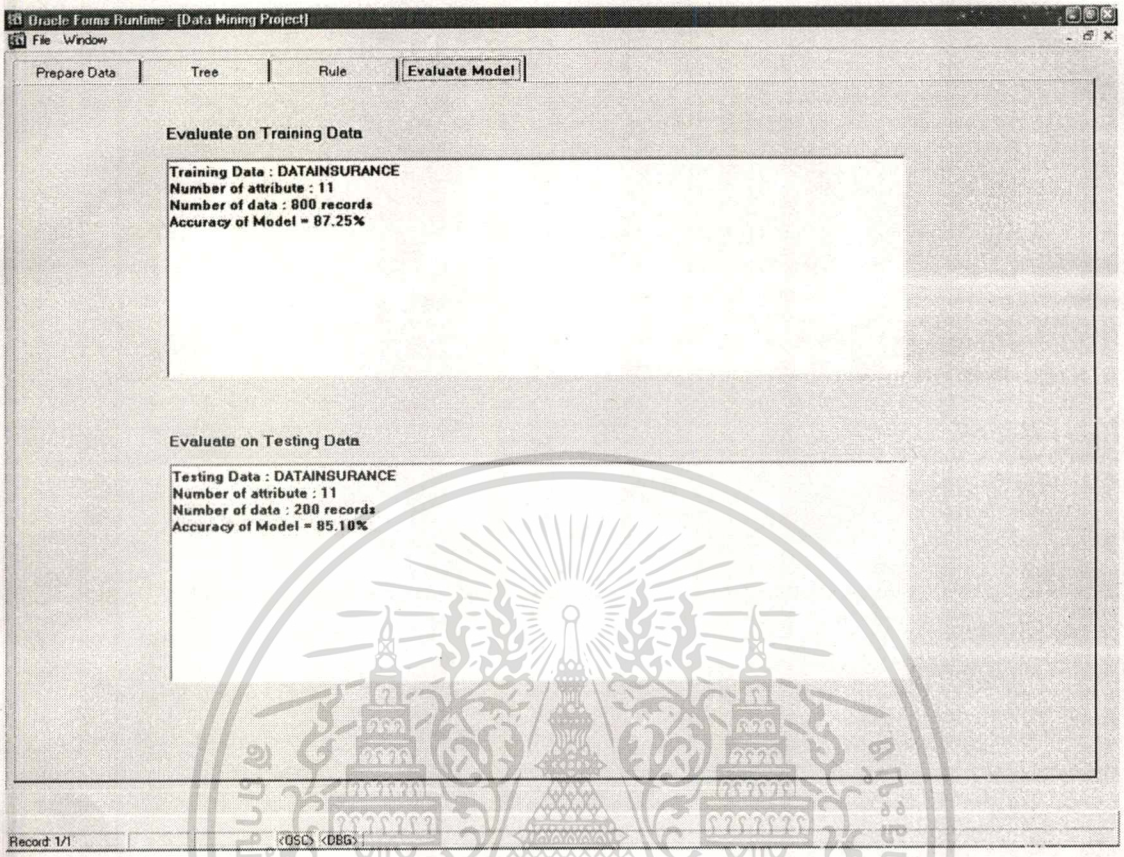
รูปที่ ก.8 หน้าจอแสดงผลลัพธ์ในรูป Decision Tree



รูปที่ ก.9 หน้าจอแสดงผลลัพธ์ในรูปของกฎ

เมื่อสร้างแบบจำลองพยากรณ์จากข้อมูล Train แล้ว ขั้นตอนต่อไปจะเป็นการนำแบบจำลองพยากรณ์ที่ได้มาตรวจสอบว่ามีความน่าเชื่อถือมากน้อยเพียงใด โดยการนำข้อมูลที่แบ่งไว้สำหรับการ Test กับแบบจำลองพยากรณ์ที่ได้ โดยเลือกปุ่ม Evaluate เพื่อให้ระบบแสดงความถูกต้องของแบบจำลองโดยใช้ข้อมูล Train และเลือกปุ่ม Test เพื่อทำการทดสอบแบบจำลองโดยใช้ข้อมูล Test ที่แบ่งไว้ในตอนแรกจากการกำหนด Percentage Split

หลังจากสั่งให้ระบบทำการตรวจสอบความถูกต้องของแบบจำลองพยากรณ์ ระบบจะแสดงผลการทดสอบดังรูปที่ ก.10



รูปที่ ก.10 หน้าจอแสดงผลลัพธ์จากการทดสอบแบบจำลอง

ประวัติผู้เขียน

นางสาวธนวันต์ นามวงษ์ เกิดวันที่ 26 กันยายน พ.ศ 2520 สำเร็จการศึกษา วิทยาศาสตร์บัณฑิต จากภาควิชาคณิตศาสตร์ประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ปีการศึกษา 2540 ปัจจุบันทำงานในตำแหน่งนักวิเคราะห์ระบบ ให้กับบริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด มหาชน