การพัฒนาระบบวิเคราะห์กลุ่มลูกค้าเป้าหมายที่สร้างความคุ้มค่าให้กับธุรกิจการประกันภัย

PROSPECT ANALYSIS SYSTEM IMPLEMENTATION FOR INSURANCE BUSINESS

โดย

นุสรา จิรเจริญจิตต์

NUSSARA CHIRACHAROENJIT

อาจารย์ที่ปรึกษา

รศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 1 ปีการศึกษา 2549

PROSPECT ANNALYSIS SYSTEM IMPLEMENTATION FOR INSURANCE BUSINESS

NUSSARA CHIRACHAROENJIT

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

1/ 2006

ชื่อหัวข้อ	การพัฒนาระบบวิเคราะห์กลุ่มลูกค้าเป้าหมายที่สร้างความคุ้มค่าให้กับธุรกิจการประกันภัย
นักศึกษา	นางสาวนุสรา จิรเจริญจิตต์
รหัสนักศึกษา	47066619
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

ในปัจจุบันกลุ่มลูกค้าในธุรกิจการประกันภัยได้มีการเพิ่มจำนวนขึ้นอย่างรวดเร็ว จึงทำให้ธุรกิจการประกันภัยมีกลุ่มลูกค้าที่หลากหลายและมีจำนวนเพิ่มมากขึ้นตามไปด้วย ดังนั้นการวิเคราะห์ข้อมูลเพื่อทำการแยกแยะกลุ่มลูกค้าที่คาดว่าจะสามารถสร้างความคุ้มค่าให้กับองค์กร จึงเป็นกุญแจที่สำคัญอีกอย่างหนึ่งที่ทำให้ผู้บริหารสามารถนำมาใช้เป็นเครื่องมือในการสร้างความสำเร็จในการดำเนินธุรกิจประเภทนี้ได้เป็นอย่างดี

โครงการพัฒนาระบบงานนี้จึงมีวัตถุประสงค์เพื่อพัฒนาระบบสำหรับวิเคราะห์กลุ่มลูกค้าเป้าหมายที่สร้างความคุ้มค่าให้กับธุรกิจการประกันภัยเมื่อเทียบกับค่าสินไหมทดแทนโดยจะนำเอาปัจจัยต่างๆที่เกี่ยวข้องกับกลุ่มลูกค้าเป้าหมาย รวมทั้งเงื่อนไขที่เกี่ยวข้องกับการทำประกันภัยที่มีส่วนในการวิเคราะห์และออกแบบระบบ เพื่อให้ผู้บริหารสามารถมองเห็นภาพรวมของธุรกิจประเภทนี้ได้อย่างเป็นรูปธรรมมากยิ่งขึ้น

Title	Prospect Analysis System Implementation for General Insurance Business
Student	Miss. Nussara Chiracharoenjit
Student ID.	47066619
Degree	Master of Science
Programme	Information Science
Academic Year	2006
Advisor	Asst. Prof. Dr. Warapoj Kreesuradej

ABSTRACT

The rapid increase of customers in insurance business also made the variety of customer types.If we can analyze for separate the valuable customers that we can gain the benefit from them into another group for marketing purpose will be the essential key for management person in insurance company to succeed in insurance business.

This project has an objective to study and develop the information analysis system for the valuable insurance customers in the term of loss from claim procedure included with the major factors that relate the loss ratio for the prospect customers.This project will enhance the management capability by develop the system that they can gain the analyzed information visually.

กิตติกรรมประกาศ

โครงการพัฒนาระบบฉบับนี้สำเร็จได้อย่างดี ได้รับความช่วยเหลือทั้งด้านความรู้ แนวทางปฏิบัติ และกำลังใจจากหลายท่าน ต้องขอขอบคุณบุคคลดังต่อไปนี้ คุณพ่อผู้ล่วงลับ และคุณแม่ ที่ให้การอบรมเลี้ยงดูและเป็นกำลังใจอย่างดียิ่งเสมอมา แม้ว่าคุณพ่อจะไม่ได้อยู่แล้วแต่คำสอนและข้อคิดของคุณพ่อก็ยังคงเป็นสิ่งที่มีค่ายิ่งที่ทำให้ข้าพเจ้าได้ใช้เป็นแนวทางในการดำเนินชีวิตในทุกๆด้านตลอดมา, รศ.ดร.วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ที่ปรึกษา ที่ได้ให้คำแนะนำและคำปรึกษาที่ดี เป็นแรงสนับสนุนที่ทำให้ข้าพเจ้าสามารถดำเนินโครงการนี้จนแล้วเสร็จ โดยเฉพาะความใส่ใจและความเข้าใจที่อาจารย์มีต่อศิษย์อย่างข้าพเจ้า ทำให้ข้าพเจ้ารู้สึกทราบซึ้งในความอนุเคราะห์จากอาจารย์เป็นอย่างมาก และขอขอบพระคุณเป็นอย่างสูง, คุณสิทธิโชค พรไพศาลสุข ผู้ดูแลระบบ บริษัท อลิอันซ์ซี.พี. ประกันภัย ที่ได้ใช้ความเชี่ยวชาญ ให้ความช่วยเหลือ ให้คำปรึกษา โดยเฉพาะอย่างยิ่งความช่วยเหลือในด้านซอฟต์แวร์เป็นอย่างดี และสุดท้าย พี่ๆ เพื่อนๆ น้องๆ ทุกคน ที่ช่วยเหลือและเป็นกำลังใจให้

นุสรา จิรเจริญจิตต์

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมา

ธุรกิจในโลกปัจจุบันต่างเผชิญภาวะการแข่งขันสูง การแข่งขันทางการตลาดเป็นไปในระดับสากลมากขึ้น ทำให้เกิดคู่แข่งจำนวนมาก ธุรกิจประกันภัยก็เป็นอีกหนึ่งธุรกิจที่ประสบปัญหาเช่นเดียวกัน ประกอบกับการเปิดการค้าเสรีของธุรกิจประกันภัยในประเทศไทยทำให้เกิดการลงทุนของบริษัทประกันภัยต่างชาติในประเทศไทยมากขึ้น ดังนั้นผู้ดำเนินธุรกิจประกันภัยทั้งหลายต่างพยายามมองหาช่องทาง หรือ กลยุทธ์ที่จะทำให้สามารถแข่งขันกับตลาดโลกได้ โดยหนึ่งในกลยุทธ์ที่เป็นที่นิยมคือ การให้ความสำคัญกับลูกค้า และเนื่องจากข้อมูลของธุรกิจประกันภัยมีเป็นจำนวนมหาศาลจึงทำให้การสืบค้นหรือนำข้อมูลมาวิเคราะห์โดยปกติทำได้ค่อนข้างลำบาก และโอกาสเกิดความผิดพลาดสูง ดังนั้นโครงการจึงได้นำเสนอระบบที่พัฒนาโดยอาศัยเทคนิคของดาต้าไมนิ่ง (Data mining) มาประยุกต์ใช้ โดยอาศัยการวิเคราะห์เชิงปริมาณในการค้นหากลุ่มลูกค้าที่จะสร้างความคุ้มค่าให้กับองค์กร เพื่อช่วยลดข้อจำกัดดังกล่าว รวมทั้งเพื่อใช้เป็นพื้นฐานหรือแนวทางในการพัฒนาระบบลูกค้าสัมพันธ์ต่อไปในอนาคต

1.2 วัตถุประสงค์ของการศึกษา

วัตถุประสงค์ของการศึกษาและพัฒนาระบบงานนี้ เพื่อพัฒนาระบบที่จะช่วยในการวิเคราะห์ข้อมูลที่มีปริมาณมากๆ เช่น ข้อมูลกลุ่มลูกค้าที่จะสร้างความคุ้มค่าให้องค์กร และเพิ่มความแม่นยำในการจำกัดความผิดพลาดที่เกิดจากการทำงานด้วยมือ โดยการใช้โปรแกรมที่ทำงานโดยอัตโนมัติ และช่วยให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้เป็นแนวทางในการพัฒนาระบบลูกค้าสัมพันธ์ต่อไปในอนาคตได้

1.3 ขอบเขตการดำเนินงาน

ระบบงานที่ทำการศึกษานี้ จะเป็นการศึกษาและพัฒนาระบบงานโดยใช้ เทคนิคของดาต้าไมนิ่งด้วยแบบจำลองดิซิชั่นทรี (Decisions Tree Model) ซึ่งเป็นอัลกอริทึ่มประเภท Classification เพื่อช่วยในการจัดกลุ่มลูกค้า โดยอาศัยเครื่องมือหลักที่ช่วยให้การทำงานในการนำข้อมูลที่มีอยู่มาวิเคราะห์หาผลลัพธ์พฤติกรรมของลูกค้าที่จะสร้างความคุ้มค่าให้กับองค์กร ดังนี้

1. ฐานข้อมูล Oracle 9i Database
2. Developper2000 Builder Tool

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังต่อไปนี้

1. กำหนดวัตถุประสงค์และเป้าหมายของการดำเนินงาน
2. ศึกษาและรวบรวมข้อมูลที่จะนำมาใช้พัฒนาระบบ
3. ศึกษาแนวคิดและทฤษฎีที่เกี่ยวกับการทำดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้
4. ศึกษาอัลกอริทึม C4.5 เพื่อนำมาประยุกต์ใช้กับระบบงาน
5. ออกแบบและพัฒนาระบบงาน
6. วิเคราะห์ผลลัพธ์ที่ได้
7. สรุปผลการศึกษาและพัฒนาปรับปรุงระบบให้ดีขึ้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

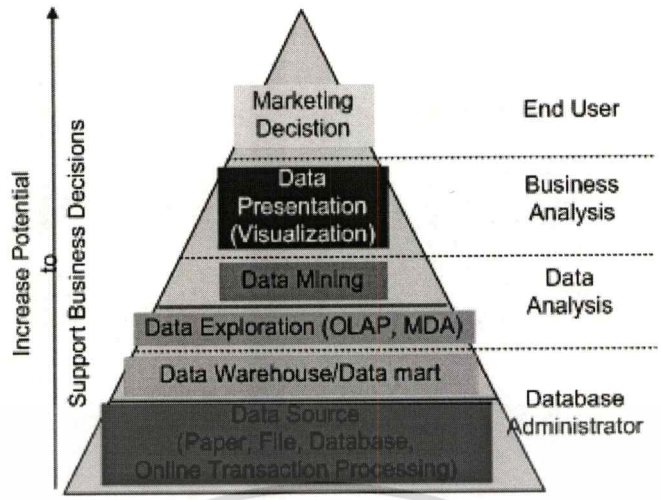
1. ความรู้ ความเข้าใจ ในหลักการและขั้นตอนของการทำดาต้าไมนิ่ง
2. เพื่อสร้างตัวแบบพฤติกรรมลูกค้า ซึ่งจะใช้ในการพยากรณ์ว่าลูกค้ารายใดน่าจะทำธุรกรรมกับองค์กร
3. สามารถนำข้อมูลที่ได้จากการทำดาต้าไมนิ่งมาใช้เป็นแนวทางในการพัฒนาระบบลูกค้าสัมพันธ์ต่อไป
4. เพื่อเพิ่มความแม่นยำในการจำกัดความผิดพลาดที่เกิดจากการทำงานด้วยมือ โดยการใช้โปรแกรมที่ทำงานโดยอัตโนมัติ

ในบทนี้เป็นการกล่าวถึงวัตถุประสงค์และขอบเขตของการทำงานในเบื้องต้นของระบบที่จะพัฒนา โดยในบทต่อไปกล่าวถึงรายละเอียดของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

บทที่ 2
หลักการและทฤษฎีที่เกี่ยวข้องกับดาต้าไมนิ่ง

2.1 ความหมายของดาต้าไมนิ่ง (Data Mining)

การทำเหมืองข้อมูล หรือ ดาต้าไมนิ่ง เป็นกระบวนการค้นหาสารสนเทศหรือข้อมูลความรู้ที่อยู่ในฐานข้อมูลขนาดใหญ่ที่ซับซ้อน เพื่อนำข้อความรู้ที่ได้ไปใช้ประโยชน์ในการตัดสินใจ สารสนเทศที่ได้อาจนำมาสร้างการพยากรณ์หรือสร้างตัวแบบสำหรับการจำแนกหน่วยหรือกลุ่ม หรือแสดงความสัมพันธ์ระหว่างหน่วยต่างๆ หรือให้ข้อสรุปของสาระในฐานข้อมูล การทำเหมืองข้อมูลประกอบขึ้นด้วยการนำกระบวนการทางสถิติและการเรียนรู้ผ่านระบบคอมพิวเตอร์เพื่อสร้างตัวแบบ กฎเกณฑ์ รูปแบบ การพยากรณ์และข้อความรู้ จากฐานข้อมูลขนาดใหญ่ โดยการทำเหมืองข้อมูลนั้นมีขั้นตอนการดำเนินงานหลายขั้นตอนซึ่งต้องอาศัยเทคนิคหรือวิธีการต่างๆ เช่น วิธีการจัดกลุ่ม การค้นหาความสัมพันธ์ การพยากรณ์ เป็นต้น การดำเนินงานมักอยู่ในลักษณะของการสร้างตัวแบบ(Modeling) ที่อธิบายความเป็นไปหรือสภาพการณ์หนึ่งที่เกิดขึ้นแล้ว หรือที่ทราบคำตอบ แล้วนำตัวแบบนี้มาใช้อธิบายสถานการณ์ที่ยังไม่เกิดขึ้น หรือที่ไม่ทราบคำตอบ ตัวแบบเหล่านี้อาจเป็นตัวแบบที่เรียบง่ายไปจนถึงตัวแบบที่ยุ่งยากซับซ้อน และอาจใช้การผสมผสานแนวคิดหรือเครื่องมือต่างๆ เข้าด้วยกัน เพื่อที่จะสามารถสกัดข้อความรู้ที่อยู่ในข้อมูลขนาดใหญ่ได้ โดยใช้เทคโนโลยีคลังข้อมูล (Data Warehouse) เข้ามาช่วยในการจัดการข้อมูลเพื่อเพิ่มประสิทธิภาพของการทำเหมืองข้อมูล ดังนั้นถ้ามีฐานข้อมูลขนาดใหญ่ที่มีข้อมูลคุณภาพดี เทคโนโลยีการทำเหมืองข้อมูลจะช่วยในการค้นหรือแสวงหาโอกาสทางธุรกิจใหม่ โดยการทำเหมืองข้อมูลซึ่งจะก่อให้เกิดกระบวนการอัตโนมัติในการค้นพบสารสนเทศ หรือข้อความรู้ในฐานข้อมูลขนาดใหญ่ ด้วยการใช้วิธีการเช่นการพยากรณ์แนวโน้มและพฤติกรรมการบริโภคแบบอัตโนมัติ หรือเกิดกระบวนการอัตโนมัติในการค้นพบรูปแบบที่ไม่เคยรู้จักมาก่อน ด้วยการใช้วิธีการค้นหาเข้าไปในรายละเอียดของฐานข้อมูลเพื่อหารูปแบบที่ซ่อนอยู่ในฐานข้อมูล ดังรูปที่ 2.1 เริ่มต้นตั้งแต่ ตารางข้อมูลธรรมดาไปจนถึงการตัดสินใจระดับสูง ดังจะเห็นได้ว่า Data Mining เป็นส่วนประกอบอันใหม่ที่มีความสำคัญของเครื่องมือทางธุรกิจอย่างหนึ่งคุณค่าของข้อมูลที่ใช้สำหรับสนับสนุนการตัดสินใจซึ่งจะเพิ่มขึ้นจากล่างไปบนสุดของรูปปิรามิด จำนวนของข้อมูลและขนาดและระดับการตัดสินใจในข้อมูลที่ลักษณะที่ต่างๆ กัน จึงมีระดับของผู้ตัดสินใจต่างกัน Database administrator จะตัดสินใจบนระดับของ Data Warehouse และแหล่งข้อมูลเท่านั้น ส่วนนักวิเคราะห์ธุรกิจและผู้บริหารจะตัดสินใจบนเหนือของปิรามิด

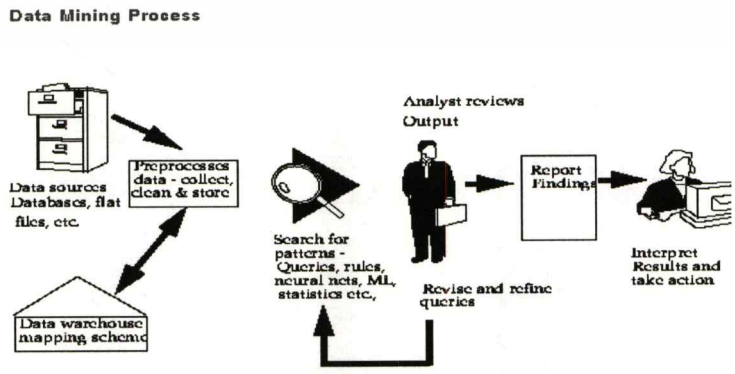


รูปที่ 2.1 แสดง คาต้าไมนิ่ง และเครื่องมือทางธุรกิจต่าง ๆ (Cabena et al., 1997)

การนำข้อมูลของ Data Warehouse ที่รวบรวมข้อมูลจากหลายๆ ที่และดึงข้อมูลเหล่านั้นเข้าไปในฐานข้อมูล ที่มีขนาดใหญ่ โดยคาต้าไมนิ่งจะนำข้อมูลมาสร้างแบบจำลองทางสถิติ ในการหารูปแบบความสัมพันธ์ของฐานข้อมูลที่มีอยู่ ในการช่วยวิเคราะห์การตัดสินใจในธุรกิจหรือกิจการอื่นๆ ตามต้องการ

2.2 กระบวนการทำงานของดาต้าไมนิ่ง

กระบวนการทำงานของคาต้าไมนิ่งจะเริ่มตั้งแต่การเตรียมข้อมูลจากแหล่งต่างๆ โดยอาศัยนักวิเคราะห์ซึ่งจะทำการตรวจสอบและสืบค้นข้อมูลเหล่านั้นโดยการกำหนดกฎเกณฑ์ หรือใช้เทคนิคต่างๆ เช่น นิวรอนเน็ตเวิร์ก หรือ ML รวมไปถึงวิธีการทางสถิติต่างช่วย เพื่อให้ได้สารสนเทศที่สามารถนำไปตีความ หรือ นำไปปฏิบัติให้บรรลุตามเป้าหมายในการพยากรณ์ต่างๆได้ ดังรูปที่ 2.2



รูปที่ 2.2 แสดงกระบวนการทำงานของดาต้าไมนิ่ง

2.2.1 ขั้นตอนหลัก ในการทำเหมืองข้อมูล

2.2.1.1 การวิเคราะห์ความต้องการของธุรกิจ

การวิเคราะห์ถึงความต้องการธุรกิจ จะช่วยให้เข้าใจถึงประเด็นที่ผู้บริหารต้องตัดสินใจ ซึ่งเกี่ยวกับการสร้างความสำเร็จให้กับธุรกิจ อาศัยการวิเคราะห์ความต้องการของธุรกิจซึ่งจะทำให้เกิดความเข้าใจถึงสถานะในปัจจุบันของธุรกิจ และทำให้ผู้บริหารสามารถกำหนดเรื่องที่ต้องตัดสินใจได้ดียิ่งขึ้น โดยการกำหนดวัตถุประสงค์ทางธุรกิจ คือจะต้องเข้าใจปัญหาและความต้องการทางธุรกิจ การกำหนดวัตถุประสงค์ทาง ธุรกิจนั้นจะเป็นส่วนที่กำหนดว่าเมื่อไหร่ที่จะใช้ดาต้าไมนิ่งในการแก้ปัญหา ซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ ทางธุรกิจ และการวิเคราะห์เบื้องต้นว่า ข้อมูลที่มีอยู่เป็นอย่างไรบ้าง มีข้อมูลอย่างไร และต้องการอะไรจากข้อมูล ซึ่งขั้นตอนนี้จะสามารถมองถึง อัลกอริทึม และฐานข้อมูลที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้ การใช้งานดาต้าไมนิ่งให้ได้ประโยชน์สูงสุดจำเป็นต้องมีการกำหนดวัตถุประสงค์ที่ชัดเจน เช่น ต้องการ เพิ่มยอดการตอบรับการขายทางจดหมาย ขึ้นอยู่กับการระบุเป้าหมายว่า จะเพิ่มอัตราการตอบรับหรือเพิ่มมูลค่าการตอบรับซึ่ง จำเป็นที่จะต้องสร้างแบบจำลองที่แตกต่างกัน วัตถุประสงค์ที่กำหนดขึ้นมาจะต้องมีการระบุวิธีการในการวัดผลลัพธ์ที่ได้จาก โครงการ รวมถึงต้นทุนที่สมเหตุสมผลด้วย

2.2.1.2 การวิเคราะห์ความต้องการข้อมูล

เนื่องจากคลังข้อมูลขนาดใหญ่ที่มีอยู่นั้น มีข้อมูลที่หลากหลาย ซึ่งมีทั้งข้อมูลที่จำเป็นต้องใช้ในการทำเหมืองข้อมูลกับข้อมูลอื่นซึ่งไม่เป็นที่ต้องการในขณะนี้ จึงต้องมีขั้นตอนการกำหนดรายการและประเภทของข้อมูลที่จะนำมาใช้ โดยมีการตรวจสอบในด้านของคุณภาพของข้อมูล จำนวน ปริมาณเนื้อหาและการเข้าถึงข้อมูล เพื่อกำหนดเป็นข้อมูลที่ต้องการทำเหมือง ในกรณีที่มีข้อมูลจำนวนมาก อาจใช้การเลือกตัวอย่างข้อมูลมาทำเหมืองก่อนได้เพื่อลดค่าใช้จ่าย ซึ่งเมื่อกำหนดข้อมูลที่ต้องการได้แล้วก็จะต้องทำการเตรียมข้อมูล ซึ่งมีขั้นตอนการเตรียมข้อมูลดังนี้

1) การเตรียมข้อมูล (Data Preparation) ขั้นตอนนี้เป็นหัวใจของขั้นตอนในการทำทั้งหมด เป็นช่วงที่ใช้เวลามากที่สุดในขั้นตอน โดยปกติแล้วต้องการเวลาประมาณ 60% ของเวลาทั้งหมดในการเตรียมข้อมูล

2) การเลือกข้อมูล (Data Selection) จุดประสงค์ คือการระบุแหล่งของข้อมูลที่มี และทำการดึงเอาข้อมูลออกมาใช้สำหรับการวิเคราะห์เบื้องต้นในการ เตรียมตัวสำหรับการที่จะทำการทำเหมืองข้อมูล ในขั้นต่อ ๆ ไป การเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจ ที่ได้กำหนดไว้ตั้งแต่ต้น และการเลือกข้อมูลก็ยังถูกกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องถูกทำความเข้าใจว่าตัวแปรแต่ละตัว

แปรหมายความว่าอะไร ประกอบด้วยอะไร ไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้น แต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล, ค่าที่เป็นไปได้, แหล่งกำเนิดของข้อมูล, รูปแบบของข้อมูล และลักษณะอื่น ๆ จะมีตัวแปร 2 ชนิดคือ

- ตัวแปรแบบ Categorical
- Nominal Variable กล่าวถึงชนิดนี้ของ Object ที่มันอ้างถึงแต่ไม่มีลำดับ ในค่าที่เป็นไปได้ (Possible Value) ตัวอย่างเช่น สถานะการแต่งงาน (โสด, แต่งงาน, หย่า, ไม่ทราบ), เพศ (ชาย, หญิง), ระดับการศึกษา (ปริญญาโท, ปริญญาตรี, ม.ปลาย)
- Ordinal Variable มีลำดับสำหรับค่าที่เป็นไปได้ ตัวอย่างเช่น ลำดับของ ลูกค้า (ดี, ปานกลาง, ไม่ดี)
- ตัวแปรแบบ Quantitative ซึ่งมีการวัดความแตกต่างระหว่างค่าที่เป็นไปได้
- Continuous (ค่าที่ต่อเนื่อง) เช่นรายได้, เฉลี่ยจำนวนครั้งที่ซื้อ, รายได้
- Discrete (ค่าเป็นจำนวนเต็ม) เช่นจำนวนพนักงาน, เวลาปี (เดือน, ฤดู, ไตรมาส)
- ตัวแปรของข้อมูลมีหลายตัวมากแต่ตัวแปรที่ถูกเลือกสำหรับทำคาต้าไมนิ่งนั้นถูกเรียกว่า "Active Variable" เพราะว่ามันจะถูกใช้สร้างความแตกต่างของกลุ่มย่อยต่างๆ และสามารถถูกนำมาพยากรณ์ผลลัพท์ได้ เมื่อทำการเลือกข้อมูลจะต้อง พิจารณาอายุของข้อมูลด้วย เพราะว่าสถานการณ์ภายนอกเปลี่ยนแปลงตลอดเวลา ซึ่งจะทำให้ประสิทธิภาพของการทำเหมืองข้อมูลลดลง ตัวอย่าง รสนิยมการใช้ชีวิต การเปลี่ยนงาน

3) การกลั่นกรองข้อมูล (Data Preprocessing) จุดประสงค์หลัก เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นเหมาะสม ข้อมูลที่สมบูรณ์เป็นเครื่องประกัน ว่าการทำเหมืองข้อมูลจะสำเร็จ ในขั้นตอนนี้เป็นขั้นตอนที่มีปัญหามากกว่า ในขั้นตอนของการเตรียมข้อมูล เพราะข้อมูลส่วนใหญ่ที่มีในองค์กร ไม่ได้ถูกเตรียมมาเพื่องานเหมืองข้อมูลโดยเฉพาะ ข้อมูลจะถูกนำมาจากแหล่งต่าง ๆ ถูกจัดเก็บไม่ดี ข้อมูลที่ถูกนำมาจาก ภายนอก แล้วนำมาเพื่อให้เข้ากับข้อมูลภายในที่มีอยู่ ปัญหาหลักของ Data คือ คุณภาพและ Data Integrity ในขั้นตอนนี้ก่อนอื่นจะต้องทำการทบทวนโครงสร้างของข้อมูลใหม่ และวัดคุณภาพของมัน โดยวิธีทางสถิติ หรือสุ่มตัวอย่างเครื่องมือที่ใช้ในการทำการกลั่นกรองข้อมูลมีดังต่อไปนี้

- ค่าตัวแปรเป็นแบบ Categorical การแบ่งความถี่ของค่าจะเป็นวิธีที่ทำให้เกิดความเข้าใจใน Data Content เครื่องมือทางด้านกราฟฟิคจะเป็นตัวช่วยให้เห็นและกำหนดค่าที่หายไป
- ตัวแปรแบบ Quantitative ตัวแปรประเภทนี้มักมีการใช้การวัด ตัวอย่างเช่น ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย ค่ากลาง ค่ามัธยฐาน และค่าอื่น ๆ ทางสถิติ เมื่อนำค่าพวกนี้มาเข้าสูตรคำนวณก็จะบอกถึงค่าที่ ไม่สมบูรณ์ หรือค่าที่มีปัญหา

เครื่องมือทางกราฟฟิคอื่น ๆ เช่น Scatterplots คือรูป 2 มิติซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปรขึ้นไป ระหว่างการทำขั้นตอนการกลั่นกรองข้อมูลจะมีปัญหาบ่อย ๆ ที่มักพบได้ ได้แก่

○ Noisy Data คือตัวแปรตัวหนึ่งหรือมากกว่ามีค่าซึ่งเกินกว่าค่าที่คาดไว้ ซึ่งอาจจะหมายถึงแง่ดีหรือแง่ร้ายก็ได้ ในแง่ดีก็คือ มันจะแสดงอย่างชัดเจนถึงโอกาสที่กำลังมองหาอยู่ ในแง่ร้าย คือ มันอาจจะเป็นข้อมูลที่ไม่สมบูรณ์ สาเหตุ ที่เกิดขึ้นได้อาจจะมาจากความเลินเล่อของมนุษย์ ตัวอย่างเช่น Operator ใส่อายุให้คนเป็น 300 ปี หรือใส่ค่าของรายได้ เป็นติดลบ ค่าเหล่า นี้ควรจะถูกแก้ไข หรือเอาออกจากการวิเคราะห์ ควรมีขั้นตอนการเช็คข้อมูลก่อนนำมาใช้

○ ค่าที่หายไป Missing Value คือค่าที่ไม่ได้แสดงในข้อมูลที่ได้เลือกแล้ว หรือค่าที่ไม่สมบูรณ์ที่ลบออกไป ระหว่างการทำ Noise Detection ค่าอาจจะหายไปเพราะเกิดจากความเลินเล่อของมนุษย์ เพราะว่าไม่มีข้อมูลนั้นระหว่างการทำ Input ข้อมูล การจัดการกับค่าที่หายไป นั้นสามารถจัดการได้ด้วยเทคนิคที่ต่าง ๆ กัน

เมื่อทำการเก็บข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปที่ควรกระทำก็ คือการตรวจสอบข้อมูล เหตุที่ต้องทำการตรวจสอบข้อมูลมี 2 ข้อ ข้อแรก นักวิเคราะห์ควรมีความคุ้นเคยกับตัวข้อมูล ไม่ใช่รู้แต่ชื่อของ attribute และความหมายของมันเท่านั้น แต่ต้องรู้ถึงเนื้อหา (content) หรือความมุ่งหมายที่แท้จริงของข้อมูลด้วย ข้อสอง อาจมีความผิดพลาดของการเก็บสะสมข้อมูลเกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากฐานข้อมูลหลาย ๆ แหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้ในการวิเคราะห์ ซึ่งนักวิเคราะห์ ที่ดีจะต้องทำการตรวจสอบข้อมูลเหล่านี้ให้ถูกต้อง ตัวอย่างของความผิดพลาดที่เกิดขึ้น ได้แก่ ความผิดพลาดในการเก็บข้อมูล จาก attribute ที่ไม่ต้องการ ซึ่งเกิดจากความสับสนในการตั้งชื่อ attribute นั้น (mislabeling of field) เช่น หากต้องการเก็บค่าของระดับการศึกษาของผู้สมัครเข้าศึกษาต่อ ซึ่งในความเป็นจริงถูกเก็บไว้ใน attribute ที่ชื่อ “LEVEL_EDU” แต่ในฐานข้อมูลนั้นบังเอิญมี attribute อีกตัวหนึ่งชื่อ “EDUCATION” ซึ่งเก็บระดับการศึกษาที่ผู้สมัครต้องการเข้าศึกษา ซึ่งหากไม่ได้ตรวจสอบความสัมพันธ์และความมุ่งหมายที่แท้จริงของแต่ละ attribute แล้ว ก็อาจเกิดการสับสน โดยเก็บข้อมูลของ attribute “EDUCATION” ไปแทนก็ได้ ซึ่งเมื่อนำข้อมูลที่ได้ไปทำ Data Mining ผลลัพธ์ที่ได้ ก็จะผิดพลาดด้วย

4) การแปลงข้อมูล (Data Transformation) ระหว่างขั้นตอนของการแปลงข้อมูล โดยที่ข้อมูลที่ได้กลั่นกรองแล้วจะถูกแปลงให้เป็นรูปแบบของข้อมูลที่พร้อมจะถูก วิเคราะห์ รูปแบบของข้อมูลที่พร้อมจะถูกวิเคราะห์ คือรูปแบบของข้อมูลที่ไม่มีความขัดแย้ง ถูกจัดระเบียบมาอย่างเรียบร้อย กลั่นกรองมาจากแหล่งข้อมูลภายนอก และภายใน ซึ่งขั้นตอนนี้เป็นขั้นตอนที่สำคัญมากเนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่า นักวิเคราะห์ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของ Input อย่างไร ตัวอย่างเช่น

หลักการรูปแบบของข้อมูลถูกกำหนด แล้ว ข้อมูลที่ถูกกลั่นกรองจะเหมาะสมกับรูปแบบเฉพาะสำหรับแต่ละ กรรมวิธีของดาต้าไมนิ่งที่จะถูกใช้ การแปลงข้อมูลยัง รวมไปถึงการทำ Data Recording และ Data Format Conversion เช่นการแปลงวันที่ เป็นต้น

ทางสถิติการทำการแปลงข้อมูลยังมีเทคนิคของ Data Reduction จุดประสงค์เพื่อที่จะลดตัวแปรสำหรับการทำการ Process โดยการนำเอาตัวแปรตั้งแต่ 2 ตัวขึ้นไปมารวมกันแล้ว ทำการ Process ข้อดีก็คือลดจำนวนของตัวแปรลง และยัง สามารถจัดการได้ง่ายขึ้น อีกเทคนิคเรียกว่า Discretization โดยการแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยการแบ่ง ค่าของตัวแปรที่จะเป็น Input ให้เป็นช่วง ๆ เช่นการแปลงเงินเดือน อายุ

อีกเทคนิคเรียกว่า One of N โดยการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric ตัวอย่างเช่น ชนิดของรถ Ford, Lincoln, Nissan ให้เป็น 100, 010, 001 ปกติแบบนี้มักจะเป็น Input ของพวก Neural Network หลังจากนั้นก็จะต้องทำการปรับแต่งฐานข้อมูล ซึ่งในขั้นตอนนี้จะมีปัญหาหลัก ๆ อยู่ 3 ข้อคือ หนึ่งฐานข้อมูลที่ได้อาจมี attributes จำนวนมากที่สามารถใช้ประโยชน์ได้แต่ถูกละเลย การเลือกกลุ่มของ attributes ที่จะใช้เป็นปัญหาที่สำคัญปัญหาหนึ่ง สอง ฐานข้อมูลที่ได้อาจมีจำนวนระเบียน (record) มากเกินไปกว่าที่จะสามารถทำการวิเคราะห์ให้เสร็จลงได้ในเวลาที่เหมาะสม ซึ่งในกรณีนี้ควรจะต้องทำการสุ่มข้อมูลตัวอย่างขึ้นมาใช้แทน สาม ข้อมูลบางอย่างอาจใช้ให้เกิดประโยชน์ได้ โดยการนำเสนอ ในรูปแบบของการวิเคราะห์แบบเฉพาะเจาะจง การทำ Data engineering นั้นจะมีการทำซ้ำขึ้นมาหลาย ๆ ครั้ง เพื่อทดสอบการใช้ attribute ที่แตกต่าง , ขนาดของกลุ่มตัวอย่างที่ต่างกัน เช่น ต้องการพยากรณ์อนาคตเมื่อเวลาผ่านไป 1 , 2 , 3 , หรือ 4 เดือน อาจจะพยากรณ์ได้โดยใช้เพียง attribute เป็นตัวพยากรณ์หรืออาจใช้ข้อมูลทุกอย่างที่มีเป็นตัวพยากรณ์ก็ได้ เป็นต้น

2.2.1.3 การดำเนินการทำเหมืองข้อมูล

เป็นขั้นตอนที่นำเอาวิธีการหรือเทคนิคการทำเหมืองข้อมูลตั้งแต่หนึ่งวิธีขึ้นไป มาทำการสกัดสาระสำคัญออกจากฐานข้อมูลที่มี เช่นการตอบคำถามว่าลูกค้าจะซื้อสินค้าต่อไปหรือไม่ อาจต้องทำการวิเคราะห์ตั้งแต่การจัดกลุ่มของลูกค้าและการจำแนกหน่วยหรือลูกค้าแต่ละคนว่าจะซื้อหรือไม่ซื้อสินค้าต่อไป ทั้งนี้ ในขณะที่ทำเหมืองข้อมูล อาจมีความจำเป็นต้องเข้าถึงข้อมูลอื่นในคลังข้อมูล รวมทั้งต้องแปลงข้อมูลรายการอื่นด้วยก็ได้

2.2.1.4 การแปลผล หรือการประยุกต์ใช้กับธุรกิจ

เป็นขั้นตอนที่นำเอาสารสนเทศที่ทำเหมืองได้มาวิเคราะห์ เพื่อตอบคำถามที่ผู้ตัดสินใจมีอยู่ ซึ่งการวิเคราะห์ในส่วนนี้จะครอบคลุมการกรองสารสนเทศที่เหมาะสมกับการส่งให้ผู้ใช้และการแปลผล เช่น ถ้าวัตถุประสงค์ของการทำเหมืองข้อมูลคือการสร้างตัวแบบการจำแนก

หน่วย ในขั้นตอนการแปลผลก็จะต้องพิจารณาความเชื่อถือได้ของตัวแบบที่ได้ด้วยวิธี เช่น Cross-validation เป็นต้น ถ้าผลที่ได้ไม่เป็นที่พอใจ ก็จะต้องทำขั้นตอนนี้ซ้ำอีก รวมทั้งขั้นตอนก่อนหน้าด้วย การแปลผลนี้อาจมองได้เป็นการประยุกต์วิทยาศาสตร์และเทคโนโลยีที่มีในการทำเหมืองข้อมูลให้เป็นผลทางธุรกิจ ทำให้สามารถทำการประเมินผลที่ได้จากการทำเหมืองข้อมูล

2.2.1.5 การปฏิบัติการในการทำเหมืองข้อมูล (Data Mining Operations)

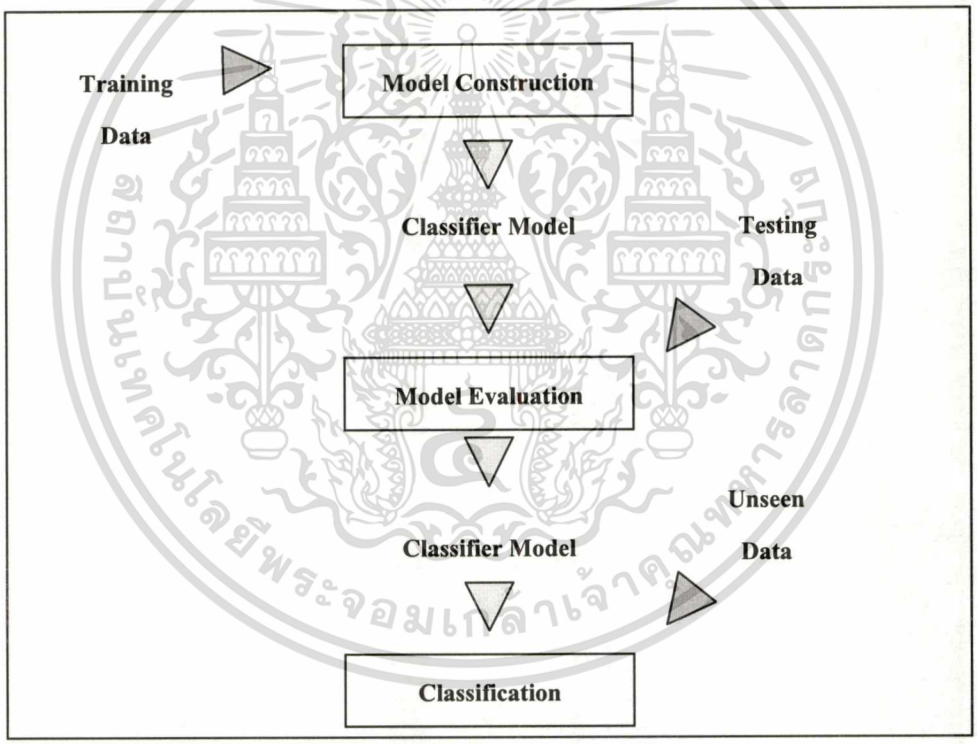
กิจกรรมหลักของการทำเหมืองข้อมูลคือ การสกัดข้อความรู้หรือรูปแบบที่มีอยู่ในฐานข้อมูลเพื่อสร้างสารสนเทศที่เป็นประโยชน์ต่อการตัดสินใจ การทำเหมืองข้อมูลโดยทั่วไปจะมีการปฏิบัติการ 4 ด้านด้วยกัน ได้แก่

- การสร้างตัวแบบการพยากรณ์และการจำแนก เป็นการปฏิบัติการที่ใช้สาระที่มีอยู่ในฐานข้อมูล ซึ่งเป็นข้อมูลที่เกิดขึ้นแล้ว มาสร้างตัวแบบเพื่อคาดการสิ่งที่จะเกิดในอนาคต วิธีการสร้างตัวแบบจะใช้วิธีการวิเคราะห์เชิงสถิติเป็นพื้นฐานและเสริมด้วยวิธีปฏิบัติของการทำเหมืองข้อมูล เพื่อให้ตัวแบบที่หาได้ง่ายต่อการทำความเข้าใจมากขึ้น ด้วยการใช้วิธีการเสนอตัวแบบในลักษณะการให้ทางเลือก การสร้างตัวแบบการพยากรณ์และการจำแนกนี้ โดยอาจใช้วิธีการจัดหน่วยเข้ากลุ่ม (Classification) ซึ่งเป็นวิธีการจำแนกหน่วยว่าเป็นสมาชิกของกลุ่มใด โดยมีการกำหนดกลุ่มไว้ล่วงหน้าแล้ว หรือการวิเคราะห์ความถดถอย (Regression) ซึ่งเป็นวิธีการหารูปแบบความสัมพันธ์ระหว่างตัวแปรต่างๆ เพื่อใช้ในการพยากรณ์หรืออธิบายตัวแปรที่สนใจ หรือการพยากรณ์ (Prediction) ซึ่งเป็นการกำหนดค่าที่คาดว่าจะเกิดขึ้นของตัวแปรที่สนใจโดยอาศัยรูปแบบที่ได้จากข้อมูลที่มีอยู่เดิม
- การวิเคราะห์ความเชื่อมโยงหรือความสัมพันธ์ เป็นการปฏิบัติการเพื่อสร้างความสัมพันธ์ระหว่างหน่วยต่าง ๆ ในฐานข้อมูล เช่น ผู้จัดการฝ่ายผลิตภัณฑ์ต้องการทราบว่าควรจัดสินค้าใดคู่กับสินค้าใดเพื่อการสั่งซื้อสินค้าเข้าร้าน เป็นต้น
- การแบ่งส่วนฐานข้อมูล เมื่อฐานข้อมูลมีขนาดใหญ่ขึ้นๆ และมีความหลากหลายมากขึ้นจะมีความจำเป็นต้องแบ่งฐานข้อมูลออกเป็นส่วนๆ โดยข้อมูลในแต่ละส่วนมีความเกี่ยวพันกันอย่างเหมาะสม ก่อนที่จะใช้วิธีปฏิบัติการอื่นต่อไป ทั้งนี้ เพื่อความสะดวกในการทำเหมืองข้อมูล รวมทั้งทำให้การสรุปสาระในแต่ละส่วนมีความหมายและใช้ประโยชน์ได้ตรงกับความต้องการ การแบ่งส่วนนี้อาจใช้การจัดกลุ่ม (Clustering) ซึ่งเป็นวิธีการกำหนดกลุ่มหรือประเภทที่มีความแตกต่างกันจำนวนหนึ่ง
- การค้นพบหน่วยผิดปกติ เป็นการปฏิบัติการเพื่อค้นพบว่ามีหน่วยที่แตกต่างจากหน่วยอื่นๆ ในฐานข้อมูลเป็นอย่างมาก หรือมีหน่วยผิดปกติหรือไม่ และสาเหตุของการเกิดหน่วยเช่นนั้นมาจากความผิดพลาดในการบันทึกข้อมูล หรือความผิดพลาดอื่นหรือเป็นลักษณะ

ผิดปกติจริงการค้นพบหน่วยผิดปกติและจัดการกับหน่วยดังกล่าวอย่างเหมาะสมในการทำเหมืองข้อมูลจะทำให้ข้อความรู้ที่ได้มีคุณค่ามากขึ้น และการใช้ประโยชน์เป็นไปอย่างมีประสิทธิผลยิ่งขึ้น

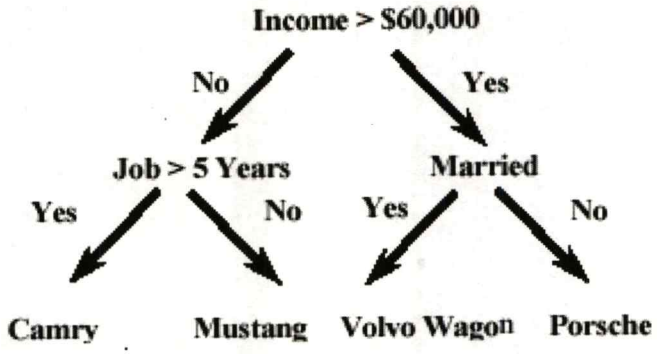
2.2.2 เทคนิคการจัดกลุ่มข้อมูลและการสร้างแบบจำลองต้นไม้

ในการพัฒนาระบบของโครงการที่ศึกษานี้จะใช้เทคนิคการจัดกลุ่ม (Classification) ช่วย ซึ่งเทคนิคการจัดกลุ่มนับเป็นกระบวนการสร้าง Model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ซึ่งเป็นการสร้างแบบจำลองพยากรณ์ (Predictive Model) ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้ หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ classification นี้แบ่งออกเป็น 3 ขั้นตอน ดังรูปที่ 2.3



รูปที่ 2.3 กระบวนการ Classification

สำหรับโครงสร้างแบบต้นไม้ของ Decision Treeเป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดง Attribute แต่ละกิ่งแสดงผลในการทดสอบ และ Leaf Node แสดงคลาสที่กำหนดไว้ รูปแบบของ Tree จะประกอบด้วย Node แรกสุดที่เรียกว่า Root Node จาก Root Node ก็จะแตกออกเป็น Node ลูก และที่ Node ลูกก็จะมีลูกของตัวเองซึ่ง Node ที่ระดับสุดท้ายจะเรียกว่า Leaf Node ดังรูปที่ 2.4



รูปที่ 2.4 แสดงตัวอย่างของดิซิชั่นทรี

จะเห็นว่า จาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบาย ถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่ม ซึ่งในแต่ละ Leaf Node นั้นอาจเป็นกลุ่มเดียวกัน ซึ่งเกิดจากเหตุผล ที่แตกต่างกันได้ โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น

“If Income = High and Married = No THEN Risk = Poor”

“If Income = High and Married = Yes THEN Risk = Good”

Decision tree เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่าย เทคนิค Decision tree จะจำกัดข้อมูลที่เป็นตัวแปรตาม (dependent variable) 1 ตัวต่อ 1 แบบจำลอง ถ้าต้องการพยากรณ์ตัวแปรตามหลาย ๆ ตัว จะต้องสร้างแบบจำลองสำหรับตัวแปรตามแต่ละตัว algorithm ของเทคนิคแบบ Decision tree ส่วนใหญ่ไม่รองรับข้อมูลแบบต่อเนื่อง (continuous data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (discrete data) เสียก่อน algorithm ที่เหล่านั้นได้ก่อน Chi-squared Automatic Interaction Detection (CHAID) , Classification and Regression Trees (CART) , C4.5 และ C5.0 algorithm เหล่านี้ส่วนมากมักเหมาะกับปัญหาแบบ classification Algorithm บางตัวปรับให้ใช้ได้กับปัญหาแบบ regression เช่น Classification and Regression Trees (CART) ซึ่งรองรับทั้งปัญหาในแบบ Classification และ regression นอกจากนี้ยังรองรับข้อมูลในแบบที่ต่อเนื่องด้วย โดยมีวิธีการที่ใช้สร้าง Decision Tree การนำข้อมูลมาสร้าง Tree มีขั้นตอนดังนี้

- หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกนำมาสร้างเป็น Root Node โดยจะมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf Node ถูกกำหนดไว้ก่อน
- นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกออกเป็นกลุ่มของตัวเอง
- แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root Node

- วนกลับไปทำที่ขั้นตอนแรก คือ หา Attribute ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวแบ่งต่อไป

โดยข้อจำกัดของ Decision Tree คือ

- การแบ่งกลุ่มแบบ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคา ต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน
- เมื่อ Algorithm เลือกว่าจะใช้ค่าไหนเป็นตัวแบ่งกลุ่มแล้วก็จะไม่สนใจค่าอื่นที่อาจมีความสำคัญเช่นเดียวกัน
- การจัดการกับข้อมูลที่ไม่ทราบค่า อาจมีผลกระทบกับผลลัพธ์ของ Decision Tree
- Tree ที่มีระดับชั้นมากเกินไป จะทำให้ข้อมูลที่ผ่าน Node แตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้น จะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์
- ปัญหาเรื่อง Over fitting / Overtraining เกิดจากการที่แบบจำลองได้เรียนรู้เข้าไปถึงรายละเอียดของข้อมูล มากเกินไปจะทำให้เกิด Node ที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูลที่ใช้ในการเรียนรู้ ซึ่งจะต้องหาวิธี การในการตัดกิ่งนี้ออกไปในธุรกิจหรือกิจการอื่นๆ ตามต้องการ

2.3 อัลกอรึทึ่ม ID3 และ C4.5

โดยจะขอกล่าวถึงอัลกอริทึ่มที่เกี่ยวข้องกับการพัฒนาระบบงานในโครงการนี้ซึ่งจะเริ่มจากอัลกอริทึ่ม ID3 ซึ่งเป็นพื้นฐานก่อนจะนำไปสู่อัลกอรทึ่ม C4.5 ดังนี้

2.3.1 อัลกอริทึ่ม ID3

ผู้พัฒนาอัลกอริธึ่ม ID3 คือ Ross Quinlan ในปี 1960 เป็นดิซิชั่นทรีที่สร้างอัลกอริทึ่มที่จัดกลุ่มโดยใช้หลักการของการใช้ทฤษฎีข่าวสาร ค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่ง ซึ่งเป็นการสร้าง tree แบบบนลงล่าง (Top-down approach) โดยกำหนดให้

S,T แทน Training Set และ T* แทน Training Set ของแอตทริบิวที่เราสนใจ

$|S_i|$ แทน จำนวนข้อมูลใน Class i ที่เลือก attribute A ซึ่งอยู่ใน Set S

|S| แทนจำนวนข้อมูลใน S

Intrinsic Info(S,A) แทน Information measure (ความสับสนที่วัดได้) หรือ entropy ของ Set S จากการเลือก attribute A

เมื่อนำสูตรนี้ไปประยุกต์ใช้กับ Training Set จะได้สมการ (2.1)

$$\text{Intrinsic Info}(S,A) = -\sum(|Si|/|S|) \log_2 \sum(|Si|/|S|) \qquad (2.1)$$

Intrinsic Info (T,A) เป็นการวัดค่าของ information เพื่อแบ่ง S โดยใช้ค่าที่เป็นไปได้ของ attribute A จะได้สมการ (2.2)

$$\text{Intrinsic Info}(T,A) = \sum(|Ti|/|T|) \times \text{Intrinsic Info}(S,A) \quad (2.2)$$

จากนั้น หา Gain(T,A) ซึ่งเป็นการวัดค่า Information ที่ได้รับ ถ้าเลือก attribute A ดังสมการ (2.3)

$$\text{Gain}(S,A) = \text{Intrinsic Info}(T^*,A) - \text{Intrinsic Info Info}(T,A) \quad (2.3)$$

โดย ID3 ถือว่า ค่าที่ได้ มีความสับสนน้อยที่สุดเอามาเป็นตัวแบ่งใน Decision Tree และ Gain คือ ความสบายใจ ซึ่งถ้า Gain มาก แสดงว่าสบายใจมากจึงเอาค่านั้นเป็นตัวแบ่ง

ข้อจำกัดของ ID3 คือ ID3 จะมีค่าของการ gain มีความโน้มเอียงมาก จากการที่ลือกใช้ค่าที่แตกต่างกันมากที่จะไป split เช่นการใช้ ค่า ID โดยนำ ID ซึ่งทราบอยู่แล้วว่า แต่ละ record จะไม่มีค่าซ้ำกันและ มีจำนวนมาก เช่นหากมีจำนวน 14 record จะต้องทำการสร้าง 14 กิ่ง โดยที่แต่ละกิ่งจะมีค่าเพียงอย่างเดียว ระหว่าง Yes กับ No โดยแต่ละกิ่งจะมีค่า Info([1,0]) หรือ Info([0,1]) ดังสมการที่ (2.4)

$$\begin{aligned}&\text{Info}([0,1], [0,1] \ldots [1,0], [0,1]) \\ &= (1/14)(\text{Info}([1,0])) + (1/14)(\text{Info}([0,1])) \ldots + (1/14)(\text{Info}([1,0])) \\ &\quad + (1/14)(\text{Info}([0,1])) \\ &= -((1/14)\log_2(1/14)) - ((1/14)\log_2(1/14)) \ldots - ((1/14)\log_2(1/14)) - ((1/14)\log_2(1/14)) \\ &= 0\end{aligned} \quad (2.4)$$

โดยค่า gain = 0.94 - 0 : จะได้ค่าเดิม คือไม่เกิดประโยชน์ต่อการทำ

2.3.2 อัลกอริทึม C4.5

อัลกอริทึมที่พัฒนาต่อจาก ID3 โดยพัฒนาขึ้นในปี 1993 โดยผู้พัฒนา ID3 คือ Quinlan โดยอ้างอิงเทคนิค Classification Decision เพื่อเรียกซ้ำสำหรับ Data Set ซึ่ง C4.5 เป็น algorithm ที่เพิ่มเติมความสามารถเข้าไปใน ID3 ซึ่งสิ่งที่เพิ่มเข้าไปนั้นคือการที่ C4.5 สามารถจัดกลุ่มให้กับ attribute ที่มีค่าเป็น numerical เพิ่มเติมจาก algorithm ID3 ที่สามารถจัดกลุ่มให้กับ attribute ที่มีค่าแบบ categories เท่านั้น ระบบ ID3 ใช้ประโยชน์จากข้อมูลในฐานะที่เป็น evaluation functions สำหรับclassification ด้วย evaluation function เพื่อใช้จัดการข้อมูลที่ ID3 ไม่สามารถจัดการได้

การจัดการข้อมูลที่ อัลกอริทึ่ม C4.5 เพิ่มเติมมาจากอัลกอริทึ่ม ID3 คือ

- ลดการพรุนนิ่งที่ผิดพลาด (error pruning)
- เลือก attribute ที่เหมาะสมมาเป็นเครื่องมือในการคัดเลือก
- ตรวจจับการเทรนนิ่งข้อมูลด้วยค่าที่ไม่ถูกต้อง
- ตรวจจับ attribute ที่เป็นข้อมูลต่อเนื่อง
- ปรับปรุงประสิทธิภาพการประมวลผลจากอัลกอริทึ่ม ID3

โดยสิ่งที่ C4.5 ต้องทำเพิ่มเติมคือ การหาค่า gain ratio ซึ่งจะใช้ตัวนี้เป็นตัวแบ่งซึ่ง Gain Ratio(S,A) เป็นการวัดค่าการแบ่งข้อมูล Set S โดยใช้ Attribute A ดังสมการที่ (2.5) และ (2.6) โดยที่ Split Intrinsic Info หรือ Split Info เป็นค่า Information ที่ได้จากการแบ่ง T ออกเป็น n subset

$$\text{Spilt Intrinsic Info}(S,A) = -\sum(\ |Si|/|S|\ ) \log_2 \sum(\ |Si|/|S|\ ) \tag{2.5}$$

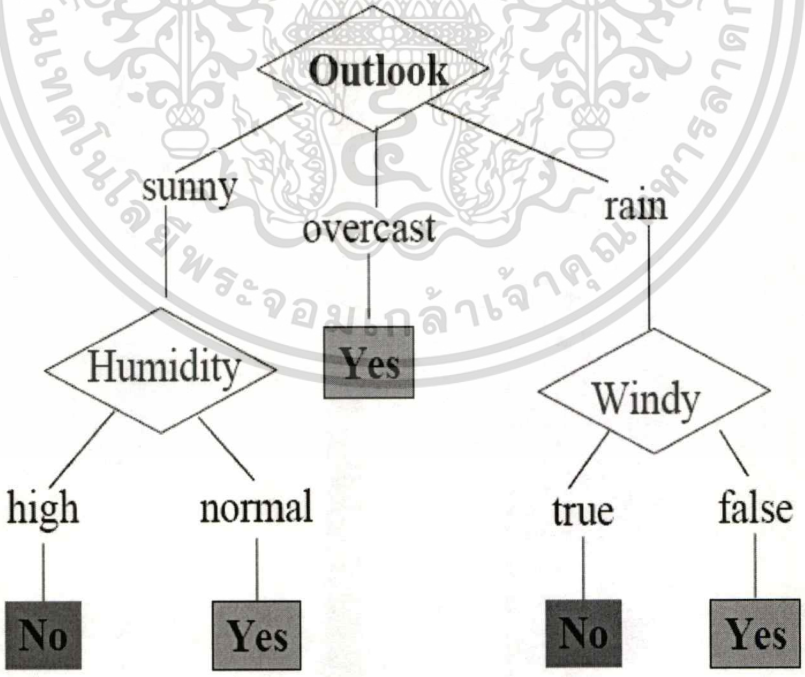
$$\text{Gain Ratio}(S,A) = \text{Gain}(S,A) / \text{Spilt Intrinsic Info}(S,A) \tag{2.6}$$

ต่อไปจะอธิบายการทำงานของอัลกอริทึ่มทั้งสอง โดยใช้ข้อมูลจากตารางที่ 2.1 และสามารถ นำมาสร้างเป็น Decision Tree ได้ดังรูปที่ 2.5 และจากตัวอย่างจะหาค่าความสับสนของ Attribute ที่พิจารณาทั้งหมด เพื่อหา Root Node โดยที่ข้อมูลประกอบด้วย class 2 class คือ Play (ควรเล่นเทนนิส) และ Don't Play(ไม่ควรเล่นเทนนิส) โดยข้อมูลอยู่ใน Class Play จำนวน 9 records และ Class Don't Play จำนวน 5 records จากนั้นพิจารณาใน แต่ละ Attribute เพื่อสร้าง Tree เพื่อแบ่งข้อมูลที่เหมาะสม ดังรูปที่ 2.5 และสามารถหาค่า Info ของ Class Play? ได้ดังนี้

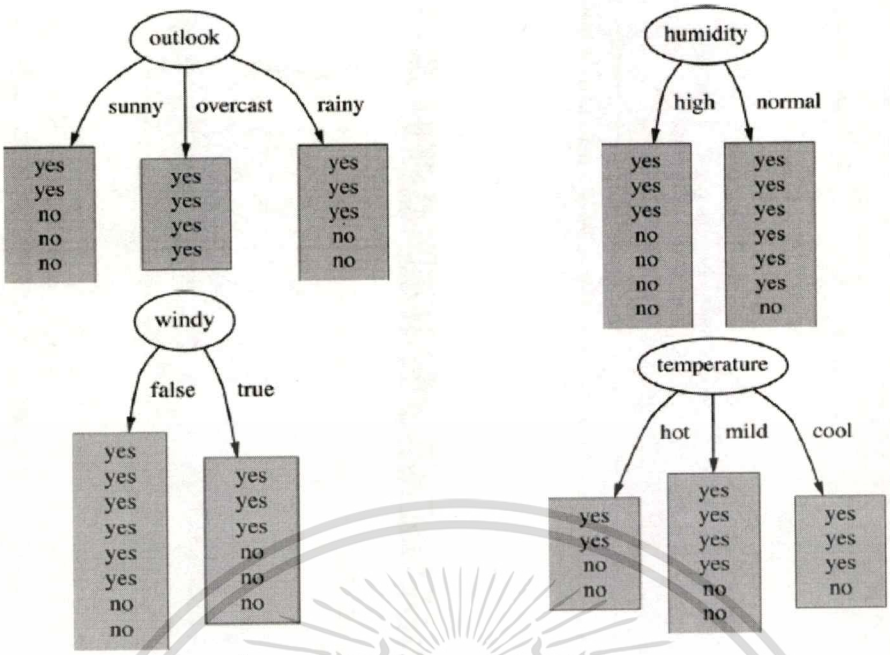
$$\begin{aligned}\text{Info (Play?)} &= -9/14 \log (9/14) - 5/14 \log (5/14) \\ &= 0.940 \text{ bits}\end{aligned} \tag{2.7}$$

ตารางที่ 2.1 แสดงข้อมูล Training Set ของปัจจัยในการเล่นเทนนิส

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No



รูปที่ 2.5 แสดง Tree ที่ได้จาก Training Set ของการตัดสินใจเล่นเทนนิส



รูปที่ 2.6 แสดง Attribute กับข้อมูล Class Play?

จากตัวอย่างจะหา Root node ได้ดังนี้

1) พิจารณา attribute outlook มี 3 Class คือ Sunny, Overcast และ Rainy

Info (Sunny) = -2/5 log (2/5) – 3/5log(3/5) = 0.971 bits (2.8)

Info (Overcast) = -1 log (1) – 0log(0) = 0 bits (2.9)

Info (Rainy) = -3/5 log (3/5) – 2/5log(2/5) = 0.971 bits (2.10)

Info (Outlook) (2.11)
= (5/14) x Info (Sunny) + (4/14) x Info (Overcast) +
(5/14) x Info (Rainy)
= 0.693 bits

Gain (Outlook) = Info (Play?) – Info(Outlook) = 0.940 – 0.693 (2.12)
= 0.247 bits

Spit Info (Outlook) (2.13)
= -(5/14) x log(5/14) - (4/14) x log(4/14) - (5/14) x log(5/14)
=1.577 bits

Gain Raito (Outlook) (2.14)
= Gain (Outlook) / Spilt Info (Outlook)
= 0.247/1.577 = 0.156 bits

2) พิจารณา attribute Temperature มี 3 Class คือ Hot, Mild และ Cool

Info (Hot) = -2/4 log(2/4) – 2/2 log(2/4) = 1 bits (2.15)

Info (Mild) = -2/6 log (2/6) – 4/6 log (4/6) = 0.918 bits (2.16)

Info (Cool) = -1/4 log (1/4) – 3/4 log (3/4) = 0.811 bits (2.17)

Info (Temperature) (2.18)
= (4/14) x Info (Hot) + (6/14) x Info (Mild) + (4/14) x Info (Cool)
= 0.911

Gain (Temperature) (2.19)
= Info (Play?) – Info (Temperature) = 0.940 – 0.911 = 0.029 bits

Spilt Info (Temperature) (2.20)
= -4/14 x log (4/14) – 6/14 x log (6/14) – 4/14 x log (4/14)
= 1.362 bits

Gain Raito (Temperature) (2.21)
= Gain (Temperature) / Spilt Info (Temperature)
= 0.029/1.362 = 0.021 bits

3) พิจารณา attribute Humidity มี 2 Class คือ High และ Normal

Info (High) = - 3/7 log (3/7) – 4/7 log (4/7) = 0.985 bits (2.22)

Info (Normal) = -6/7 log (6/7) – 1/6 log (1/6) = 0.592 bits (2.23)

Info (Humidity) = (7/14) x Info (High) + (7/14) x Info (Normal) (2.24)
= 0.788 bits

Gain (Humidity) = Info (Play?) – Info (Temperature) (2.25)
= 0.940 – 0.788 = 0.152 bits

Spilt Info (Humidity) (2.26)
= -7/14 x log (7/14) – 7/14 x log (7/14) = 1 bits

Gain Raito (Humidity) (2.27)
= Gain (Humidity) / Spilt Info (Humidity) = 0.152/1
= 0.152 bits

4) พิจารณา attribute Windy มี 2 Class คือ True และ False

Info (True) = -3/6 log (3/6) – 3/6 log (3/6) = 1 bits (2.28)

Info (False) = -6/8 log (6/8) - -2/8 log(2/8) = 0.811 bits (2.29)

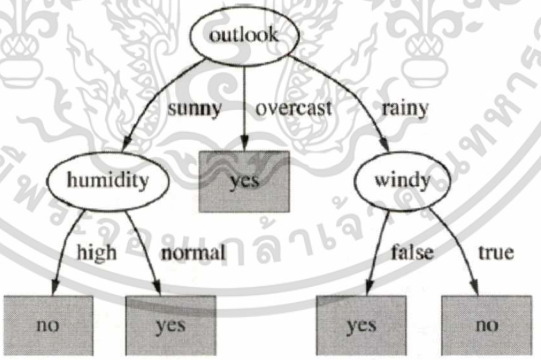
Info (Windy) = (6/14) x Info (True) + (8/14) x Info (False) (2.30)
= 0.892 bits

Gain (Windy) = Info (Play?) – Info (Windy) = 0.940 – 0.892 (2.31)
= 0.048 bits

Split Info (Windy) = (6/14) x log (6/14) + (8/14) x log (8/14) (2.32)
= 0.985 bits

Gain Raito (Windy)
= Gain (Windy) / Spilt Info (Windy) (2.33)
= 0.048/0.985 = 0.049 bits

จากทั้ง 4 attribute ค่า Gain Ratio ที่มากที่สุด คือattribute Outlook ดังนั้น Root node คือ Outlook ซึ่งจะได้ Tree ชุดแรก ดังรูปที่ 2.7



รูปที่ 2.7 แสดงดิซิสชั่นทรีที่มี root node ที่เหมาะสม

และหากต้องการเขียนเป็น Algorithm หรือเขียน Coding จะสามารถเขียนได้ดังนี้

IF outlook = ‘sunny’ AND humidity = ‘high’ THEN play = ‘no’

IF outlook = ‘sunny’ AND humidity = ‘normal’ THEN play = ‘yes’

IF outlook = ‘overcast’ THEN play = ‘yes’

IF outlook = ‘rainy’ AND windy = ‘false’ THEN play = ‘yes’

IF outlook = ‘rainy’ AND windy = ‘true’ THEN play = ‘no’

2.3.2.1 กรณีมีข้อมูลที่เป็น Unknown attribute values

- หาค่า Info (T) และ $Info_x(T)$ โดยพิจารณาจากข้อมูลที่ค่าของ Attribute A
- การหาค่า gain(x) โดย gain(x)=probability A is know x (Info(T)–Info(T))
- หาค่า Split Info(x) โดยพิจารณาจากกลุ่มของข้อมูลที่ไม่รู้ค่าของ A เป็น
- อีก 1 subset เช่น ถ้า attribute ที่จะมาทดสอบมีค่าที่เป็นไปได้ n ค่า Split Info(x)
- จะถูกคำนวณโดยแบ่งข้อมูลออกเป็น n+1 subset
- การแบ่ง Training Set สมมุติ attribute ที่เลือกจากขั้นตอนแรกมีค่าที่เป็นไปได้ O1, O2,…,On เมื่อข้อมูล 1 record ใน T ซึ่งมีค่า Oi ถูกกำหนดให้ subset Ti ค่าความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset อื่นเท่ากับ 0 และถ้าค่าใน attribute ไม่ทราบค่าความน่าจะเป็นจะมีค่าน้อยลง สำหรับข้อมูลแต่ละ record ในแต่ละ subset Ti weight จะเท่ากับความน่าจะเป็นของ Oi ที่จุดนั้นๆทำให้ |Ti| เป็นผลรวมของค่า weight w ซึ่งค่าใน attribute ไม่ทราบค่าจะถูกกำหนดให้แต่ละ subset Ti ด้วย weight ดังสมการ

$$W = \text{Probability of outcome Oi} \tag{2.34}$$

โดยความน่าจะเป็นคือ ผลรวมของ Weight ของข้อมูลทั้งหมดใน T ซึ่งมีค่า Oi หารด้วยผลรวมของ weight ของข้อมูลทั้งหมดใน T ซึ่งค่าใน attribute เป็นค่าที่ทราบค่า

- การใช้ Decision tree ที่ได้มาพยากรณ์กลุ่มของข้อมูล ในกรณีที่ค่าใน attribute ที่จะทดสอบที่ decision node เป็นค่าที่ไม่ทราบค่า ทำให้ไม่สามารถแบ่งข้อมูลได้ กรณีนี้ระบบจะสำรวจทุกเส้นทางที่เป็นไปได้ และรวมผลที่ได้จากการ classification ด้วยวิธีการทางคณิตศาสตร์ โดยผลที่ได้จะเกิดได้จากหลายเส้นทางจาก root ของ tree ไปยัง leaf node และ class ที่ได้จากการพยากรณ์จะเป็น class มีความน่าจะเป็นสูงสุด

จากตัวอย่าง สมมติ ค่าของข้อมูลมี Unknown value ดังตารางที่ 2.2

ตารางที่ 2.2 แสดงค่าข้อมูลที่มี Unknown value

Outlook	Play	Don't Play	Total
Sunny	2	3	5
Overcast	3	0	3
Rain	3	2	5
Total	8	5	13

$$\text{Info}(T) = -8/13 \times \text{Log}(8/13) - 5/13 \times \log(5/13) = 0.9691 \text{ bits} \quad (2.35)$$

$$\begin{aligned}\text{Info}_x(T) = {} & 5/13 \times (-2/5 \times \text{Log}(2/5) - 3/5 \times \text{Log}(3/5)) \\ & + 3/13 \times (-3/3 \times \text{Log}(3/3) - 0 \times \text{Log}(0/5)) \\ & + 5/13 \times (-3/5 \times \text{Log}(3/5) - 2/5 \times \text{Log}(2/5)) \\ = {} & 0.747 \text{ bits}\end{aligned} \quad (2.36)$$

$$\begin{aligned}&\text{Spilt Info}_x(T) \\ &= -(5/14) \times \log(5/14) - (3/14) \times \log(3/14) - (5/14) \times \log(5/14) \\ &= 1.537 \text{ bits}\end{aligned} \quad (2.37)$$

$$\text{Gain}(x) = 13/14 \times (0.961 - 0.747) = 0.199 \text{ bits} \quad (2.38)$$

$$\text{Gain Raio}(x) = 0.199/1.537 = 0.130 \text{ bits} \quad (2.39)$$

จะพบว่าค่าเดิม gain = 0.246 bits, split info = 1.577 และ gain ratio = 0.156 ในขณะที่ค่าใหม่ gain = 0.199 bits, split info = 1.537 และ gain ratio = 0.130 ซึ่งลดลงเล็กน้อย ทั้งนี้อาจจะเนื่องมาจาก unknown values มีจำนวนเพียง 1 ค่าทำให้ค่าต่างๆที่ได้ลดลงจากเดิมเพียงเล็กน้อย

2.3.2.2 กรณีมีข้อมูลที่เป็น แบบ Continuous attribute values

ข้อมูลที่มีค่าต่อเนื่องเช่น อุณหภูมิ เป็นต้น สมมติให้ A เป็นAttribute ชนิด Continuous attribute values การทดสอบค่าที่ attribute นี้ จะทำการแบ่งที่ละครึ่ง โดยทำการเปรียบเทียบค่า Threshold คือ ค่าที่ใช้แบ่ง ดังนี้

- เรียงลำดับ training set ด้วยค่าใน attribute A จากน้อยไปมาก และเลือกพิจารณาเฉพาะค่าที่ไม่ซ้ำกันมาพิจารณา จะได้ {c1,c2,…,cn}
- หาค่า Threshold ใดๆ ซึ่งค่า threshold จะอยู่ระหว่าง ci และ ci+1 โดยคำนวณจากค่ากลางของแต่ละช่วงดังนี้ (ci + ci+1)/2 โดย C4.5 จะเลือกค่าที่มากที่สุดใน attribute A แต่ต้องไม่เกินค่ากลางนั้นๆ จาก training Set เป็นค่า Threshold ของแต่ละช่วง เพื่อที่ว่าค่า Threshold ทั้งหมดที่ปรากฏอยู่ใน Tree จะเป็นค่าที่เกิดขึ้นจริงของข้อมูล
- หาค่า Threshold ที่เหมาะสม โดยพิจารณาจากค่า Threshold ที่มีค่า Information Gain สูงสุด

สมมติ ค่า Temperature ของข้อมูลเป็นแบบ Continuous attribute values ดังนี้

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

ใช้ข้อมูล (71 + 72)/2 = 71.5 ในการแบ่งข้อมูล จะได้ว่า

Temperature ≤ 71.5 มี ค่า Play เป็น Yes = 4, No = 2

Temperature > 71.5 มีค่า Play เป็น Yes = 5, No = 3

โดย

$$\text{Info}([4,2],[5,3]) = 6/13\ \text{Info}([4,2]) + 8/14\ \text{Info}([5,3]) = 0.939\ \text{bits} \quad (2.40)$$

จากนั้นก็ทำคำนวณกับข้อมูลต่อไปที่ละครึ่งจนหมด ดูค่า gain ที่มากที่สุดเป็นตัวแบ่งข้อมูลบน attribute ดังกล่าว

2.3.2.3 การ Pruning Decision Tree

เพื่อลดความผิดพลาดของข้อมูลในแต่ละกิ่งของ tree จึงต้องอาศัยการทำ Pruning ที่จะให้แต่ละ leaf node ที่ได้ไม่จำเป็นจะต้องประกอบด้วยข้อมูลที่อยู่ใน class เดียวกันทั้งหมด โดยแต่ละ leaf node จะมีการระบุการกระจายของข้อมูลแต่ละ class ไว้ ซึ่งจะบ่งบอกถึงความน่าจะเป็นที่ข้อมูลจะอยู่ใน class นั้น อัลกอริทึ่มของ C4.5 จะทำการ Pruning โดยการตัด Sub tree ที่จะทำโดยสมมุติว่า จะทำการแบ่งกลุ่ม Set ของข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Training set ซึ่งการคำนวณจะใช้สูตรทางสถิติซึ่งอยู่บนพื้นฐานของการกระจายแบบ Binomial

โดยกำหนดให้

N แทนจำนวนข้อมูลทั้งหมด ในแต่ละ leaf node

e แทนค่าความผิดพลาดของข้อมูล(ขอบเขตบน) เมื่อข้อมูลมีขนาดเท่ากับ N

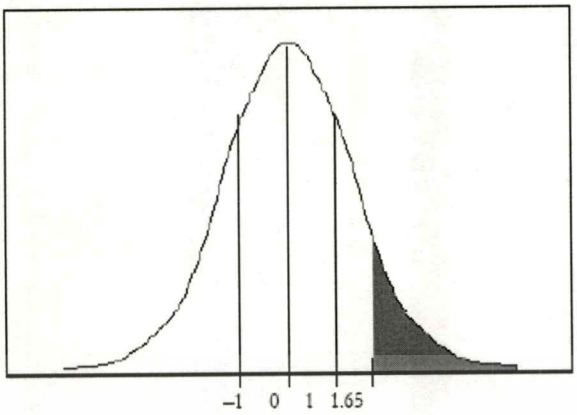
f แทนค่าความผิดพลาดที่อาจจะพยากรณ์ผิด จะมีค่าเท่ากับ ค่าที่ได้ผลลัพธ์ไม่ที่ต้องการ/ค่าของข้อมูลที่อยู่ใน leaf node นั้น

z แทนค่าการกระจายแบบ Binomial ซึ่งจะสัมพันธ์กับค่า Confidence Limits

Confidence Limits จะมีค่าเป็น percentage เป็นค่าความน่าจะเป็นสูงสุดที่จะเกิดความผิดพลาด

เช่น

Confidence limits = 25% (ค่า default ของ อัลกอริทึ่ม C4.5) จะได้ค่า z = 0.69 ดังรูปที่ 2.8



Pr[X ≥ z]	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
25%	0.69
40%	0.25

รูปที่ 2.8 แสดง ตัวอย่างค่า Binomial

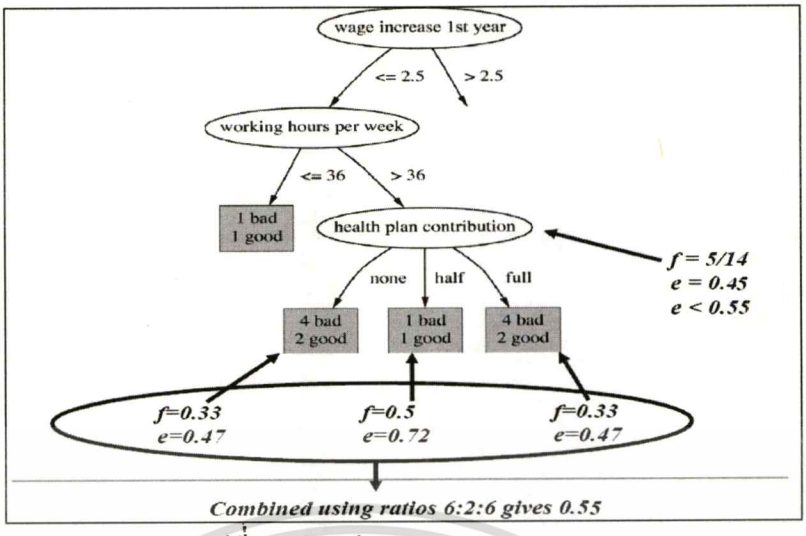
จากรูปที่ 2.8 ถ้าเลือก Confidence Limits = 5% จะได้ว่า

$$Pr[-1.65 < X < 1.65] = 100\% - (2 \times 5\%) = 90\% \quad (2.41)$$

และมีสมการหาค่าความผิดพลาดของการพยากรณ์ ดังนี้

$$e = [\, f + (z^2/2N) + z \sqrt{(f/N - f^2/N + z^2/4N^2)} \,] / (1 + z^2/N) \quad (2.42)$$

เมื่อตรวจสอบค่า e ของ parent node กับ leaf node พบว่า ค่า e ที่ parent node ≤ leaf node ก็แสดงว่าไม่ควรแตก leaf node เพราะ ค่าความผิดพลาดเมื่อไม่แตก leaf node จะน้อยกว่าการแตก leaf node ในทางกลับกัน หากค่า e ที่ parent node > leaf node ก็แสดงว่าควรแตก leaf node เพราะจะลดความผิดพลาดของข้อมูลที่จะพยากรณ์ได้ ซึ่งตัวอย่างของการ Pruning Tree ในรูปที่ 2.9 จะพบว่า ก่อนแตกกิ่ง health plan contribution ค่าความผิดพลาด(e) = 0.45 แต่เมื่อแตกกิ่ง none, half และ full แล้ว จะพบว่าค่าความผิดพลาดเพิ่มขึ้นเป็น 0.55 ดังนั้น จึงไม่สมควรแตกกิ่ง health plan contribution



รูปที่ 2.9 ตัวอย่าง Tree แสดงการ Pruning

พิจารณาที่ Health plan contribution โดยคำนวณหา e และ f ทีละ node โดยกำหนด Confidence Limits ที่ 25% ทำให้ได้ค่า z = 0.69

$$f(\text{none}) = 2/6 = 0.33 \tag{2.43}$$

$$e(\text{none}) = \frac{0.33 + (0.69)^2/12 + 0.69\sqrt{0.33 - (0.33)^2/6 + (0.69)^2/144)}}{1 + (0.69)^2/6} \tag{2.44}$$

$$= 0.47$$

$$f(\text{half}) = 1/2 = 0.5 \tag{2.45}$$

$$e(\text{half}) = \frac{0.5 + (0.69)^2/10 + 0.69\sqrt{0.5 - (0.5)^2/2 + (0.69)^2/16)}}{1 + (0.69)^2/2} \tag{2.46}$$

$$= 0.72$$

$$f(\text{full}) = 2/6 = 0.33 \tag{2.47}$$

$$e(\text{none}) = \frac{0.33 + (0.69)^2/12 + 0.69\sqrt{0.33 - (0.33)^2/6 + (0.69)^2/144)}}{1 + (0.69)^2/6} \tag{2.48}$$

$$= 0.47$$

และหา Combined using ratios เมื่อแตกกิ่ง health plan contribution ได้ว่า

$$e = e(\text{none}) + e(\text{half}) + e(\text{full}) / 3 = (0.47 + 0.72 + 0.47)/3 = 0.55 \tag{2.49}$$

บทที่ 3
การประยุกต์ใช้โปรแกรมเพื่อการพัฒนาระบบวิเคราะห์กลุ่มลูกค้าที่สร้างความคุ้มค่าให้กับธุรกิจการประกันภัย

โครงการพัฒนาระบบนี้อาศัยเทคนิคของการทำเหมืองข้อมูลแบบ Decision Tree โดยใช้ Oracle9i Server เพื่อพัฒนาระบบที่เหมาะสมกับข้อมูลเชิงสัมพันธ์ซึ่งเป็นฐานข้อมูลที่นิยมใช้ในทางธุรกิจ เครื่องมือที่ใช้คือ Oracle Developer 6i ซึ่งเป็นผลิตภัณฑ์ของ Oracle ที่มีรูปแบบเป็น GUI (Graphic User Interface) นอกจากนั้นฟังก์ชั่นต่างๆในการประมวลผลถูกสร้างเป็น Store Procedure บนฐานข้อมูลสำหรับการเรียกใช้จาก Application เพื่อสร้างแบบจำลองต้นไม้ รวมทั้งในการประมวลผลคำสั่งต่างๆจะใช้รูปแบบคำสั่ง ไดนามิค SQL เนื่องจากการเลือกตารางและข้อมูลต่างๆไม่แน่นอนขึ้นอยู่กับความต้องการของผู้ใช้ระบบเป็นหลัก

3.1 กำหนดวัตถุประสงค์

การค้นหาลูกค้าที่จะสร้างความคุ้มค่าให้กับธุรกิจการประกันภัยนั้น นอกจากจะอาศัยการออกสำรวจข้อมูลแล้ว การดึงเอาข้อมูลที่มีอยู่ในมือมาสร้างความคุ้มค่าก่อให้เกิดประโยชน์ในทางสถิติก็เป็นเรื่องที่ได้รับความสนใจอยู่ไม่น้อย ดังนั้นในกรณีศึกษานี้จึงต้องการค้นหาลักษณะของลูกค้าที่จะสามารถสร้างความคุ้มค่าให้องค์กรได้เป็นอย่างดี

ทั้งนี้เพื่อเป็นแนวทางให้ผู้ประกอบธุรกิจประกันภัยใช้สำหรับการจัดการกับลูกค้าได้อย่างมีประสิทธิภาพและเพื่อให้การจัดการกับลูกค้าได้ถูกต้องตรงกับกลุ่มเป้าหมาย รวมทั้งนำเสนอสารสนเทศที่จะช่วยประกอบการพัฒนาระบบลูกค้าสัมพันธ์ต่อไปในอนาคตได้

3.2 การเตรียมข้อมูล

จากการรวบรวมความต้องการรายงานที่สนับสนุนวัตถุประสงค์ดังกล่าว จึงต้องมีการจัดเตรียมข้อมูลในฐานข้อมูลในดาต้าแวร์เฮ้าส์ โดยมีรายละเอียดข้อมูลหลักที่เกี่ยวข้องดังต่อไปนี้

- ข้อมูลกรมธรรม์ ประกอบด้วยรายละเอียดเกี่ยวกับ กรมธรรม์ เช่น วันที่เริ่มความคุ้มครอง, วันที่สิ้นสุดความคุ้มครอง, เลขที่กรมธรรม์, ประเภทกรมธรรม์, ทุนประกันภัย, ความคุ้มครอง, เบี้ยประกันภัย, ผู้ถือกรมธรรม์ และอื่นๆ เป็นต้น
- ข้อมูลสินไหมทดแทน ประกอบด้วยรายละเอียดเกี่ยวกับ การเรียกร้องสินไหมทดแทน เช่น วันที่เกิดเหตุ, สถานที่เกิดเหตุ, ประเภทการจ่ายค่าสินไหมทดแทน, ค่าสินไหมทดแทน, เลขที่รับแจ้งเหตุ, เลขที่สินไหมทดแทน และอื่นๆ เป็นต้น

○ ข้อมูลลูกค้า ประกอบด้วยรายละเอียดเกี่ยวกับ ลูกค้า เช่น ชื่อ-นามสกุล, อายุ, เพศ, รายได้, วันเดือนปีเกิด, อาชีพ และอื่นๆ เป็นต้น

จากข้อมูลที่มีได้ทำการที่เลือกข้อมูลที่เป็นเหมาะสมมาใช้ในกระบวนการการทำดาต้าไมนิ่ง โดยรายละเอียดดังต่อไปนี้

ตารางที่ 3.1 แสดงข้อมูลที่จะนำมาทำดาต้าไมนิ่ง

Item	Attribute	Data Type	Description
1	Age	Number	อายุ
2	Sex	Character	เพศ
3	Occupation	Character	อาชีพ
4	Major	Character	ช่องทางหลักของลูกค้า
5	Quality	Character	ประเภทลูกค้า
6	Province	Character	ภูมิภาค
7	Premium	Number	เบี้ยประกันภัย
8	Sum Insured	Number	ทุนประกันภัย
9	Interest	Character	เข้าข่ายที่กำหนดความน่าสนใจ

ตารางที่ 3.2 แสดงค่าข้อมูลเพศของลูกค้าที่เป็นไปได้

Data Value of Sex	Description
Female	ผู้หญิง
Male	ผู้ชาย

ตารางที่ 3.3 แสดงค่าข้อมูลประเภทลูกค้าที่เป็นไปได้

Data Value of Quality	Description
General	ลูกค้าทั่วไป
VIP	ลูกค้าพิเศษ

ตารางที่ 3.4 แสดงค่าข้อมูลความน่าสนใจที่เป็นไปได้

Data Value of Interest	Description
Yes	เข้าข่ายน่าสนใจ
No	ไม่เข้าข่ายน่าสนใจ

ตารางที่ 3.5 แสดงค่าข้อมูลช่องทางการทำประกันภัยของลูกค้าที่เป็นไปได้

Data Value of Major	Description
Direct	ทำประกันโดยตรง
Agent	ทำประกันโดยผ่านตัวแทน
Broker	ทำประกันโดยผ่านนายหน้า

กำหนด Target Attribute ที่ใช้ในการสร้างแบบจำลองพยากรณ์ คือ Interest หรือ การเข้าข่ายที่กำหนดความน่าสนใจของลูกค้ามีค่าที่เป็นไปได้ดังตารางที่ 3.1 ข้างต้น จากนั้นจะต้องนำข้อมูลที่ได้มากลั่นกรองโดยกระบวนการคัดเลือกข้อมูล เนื่องจากข้อมูลที่มีอยู่ในปัจจุบันอาจจะมีบางส่วนที่ผิดพลาดหรือ แต่ละ Attribute ไม่สมบูรณ์เพียงพอ จึงต้องวิธีการในการคัดกรองข้อมูลที่ถูกต้องและมีความสมบูรณ์มาใช้ ซึ่งทำการสุ่มข้อมูลที่มีความสมบูรณ์ครบถ้วนจากฐานข้อมูลด้านประกันภัยประมาณ 1,000 รายการ ซึ่งทำการแปลงรูปแบบของข้อมูลให้เหมาะสมสำหรับการนำไปใช้งาน โดยมีการแปลงข้อมูลที่เป็นตัวเลขให้เป็นตัวอักษรเพื่อความรวดเร็วในการประมวลผล โดยสามารถแบ่งกลุ่มข้อมูลที่จะนำมาใช้ ได้ดังนี้

ตารางที่ 3.6 แสดงการแปลงอายุของลูกค้า

Age
>=60 Years
50 - 59
40 - 49
30 - 39
20 - 29
<= 19

ตารางที่ 3.7 แสดงการแปลงเบี้ยประกันของลูกค้า

Premium
>50,000
40,001 – 50,000
30,001 – 40,000
20,001 - 30,000
10,001 - 20,000
7,000 - 10,000
< 7,000

ตารางที่ 3.8 แสดงการแปลงอาชีพของลูกค้า

Class	Occupation
Low	กลุ่มประเภทไม่ค่อยมีความเสี่ยง เช่นพนักงานรับโทรศัพท์, แม่บ้าน, พนักงานบริษัททั่วไป เป็นต้น
Medium	กลุ่มประเภทมีความเสี่ยงปานกลาง เช่น พนักงานเซลล์แมน, นักเรียนต่ำกว่ามัธยม เป็นต้น
High	กลุ่มประเภทมีความเสี่ยงสูงเช่น พนักงานส่งเอกสาร, มอเตอร์ไซค์รับจ้าง เป็นต้น
Hazard	กลุ่มประเภทเสี่ยงกับอุบัติเหตุต่างๆสูงมาก เช่น กรรมการก่อสร้าง, ตำรวจ, นักการเมือง เป็นต้น

ตารางที่ 3.9 แสดงการแปลงทุนประกันภัย

Sum Insured
>5,000,000
3,000,001-5,000,000
1,00,001-3,000,000
500,000-1,000,000
<500,000

ตารางที่ 3.10 แสดงการแปลงภูมิภาคของลูกค้า

Province
กรุงเทพมหานคร
ภาคกลาง
ภาคเหนือ
ภาคใต้
ภาคตะวันตก
ภาคตะวันออก
ภาคตะวันออกเฉียงเหนือ

เมื่อได้ทำการจัดเตรียมข้อมูลเรียบร้อยแล้วจึงทำการแบ่งข้อมูลออกเป็น 2 ชุด คือ Training Data จำนวน 800 รายการ (ร้อยละ 80 ของข้อมูลทั้งหมด) เพื่อสร้างแบบจำลองต้นไม้ และ Testing Data อีกจำนวน 200 รายการ (ร้อยละ 20 ของข้อมูลทั้งหมด) เพื่อทดสอบแบบจำลองดังกล่าว ขั้นตอนถัดไปจะเป็นการจัดกลุ่มข้อมูลโดยใช้ระบบที่พัฒนาขึ้น

3.3 การออกแบบโปรแกรม

ในการออกแบบโปรแกรมจากตารางข้อมูล 3.1 ที่ได้นั้น กำหนดให้โปรแกรมทำการดึงข้อมูลในส่วนที่เป็นข้อมูลสำหรับสร้างแบบจำลอง (Training Data) ผ่าน Classification algorithm ที่เลือกใช้ในที่นี้คือ C4.5 เพื่อทำการ Classifier model ให้ได้แบบจำลองต้นไม้ และรูปแบบของกฎความสัมพันธ์ของข้อมูล ดังรูปที่ 3.1

สำหรับส่วนประกอบของโปรแกรมที่จะพัฒนาจะประกอบไปด้วยฟังก์ชั่นหลักดังนี้ Login, Generate Tree Model, Forecast, และ Save files โดยประกอบไปด้วยกิจกรรมในแต่ละฟังก์ชั่นต่างๆ ดังนี้

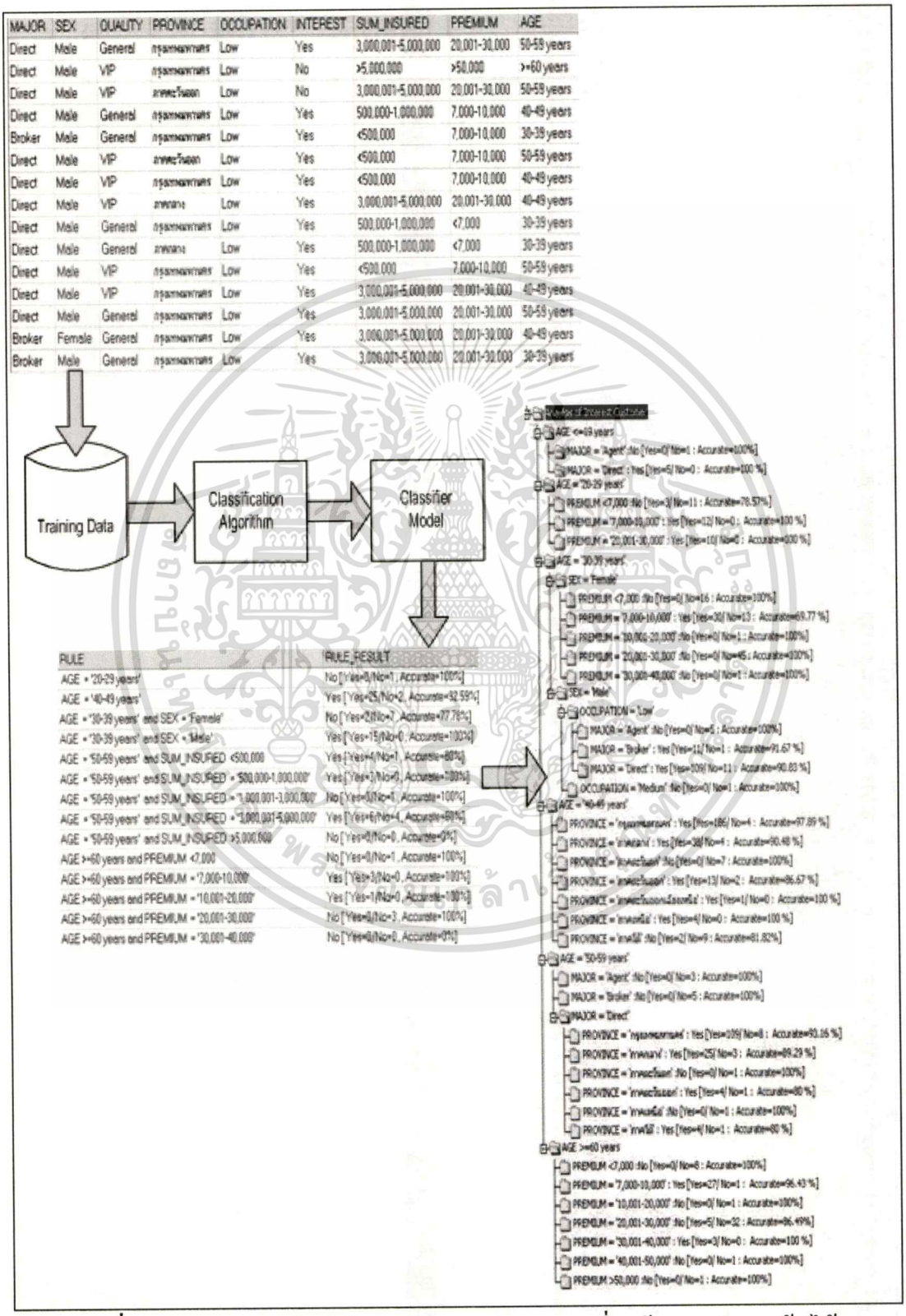
1) ฟังก์ชั่น Login เพื่อตรวจสอบสิทธิ์ในการใช้งานโปรแกรม ดังรูปที่ 3.2

2) ฟังก์ชั่น Generate Tree Model แบ่งเป็น การ Reload tree model หรือ Generate tree model ใหม่ ดังรูปที่ 3.3

3) ฟังก์ชั่น Forecast สามารถเลือกที่จะ Forecast ข้อมูล และบันทึกเป็นไฟล์เก็บไว้ดูภายหลังได้ โดยกำหนดรูปแบบข้อมูลเป็น *.csv file ดังรูปที่ 3.4

4) ฟังก์ชั่น Save files สามารถเลือกที่จะบันทึกข้อมูลเก็บไว้ดูภายหลังได้ โดยมีรูปแบบไฟล์ส่วนใหญ่เป็น *.txt file ยกเว้น Tree model ที่จะสามารถนำกลับมาใช้ได้ใหม่นั้นจะมีรูปแบบเป็น *.trm แบ่งได้เป็น รายละเอียดข้อมูล(Data), แบบจำลองต้นไม้(Tree) ,รูปแบบกฎที่ได้(Rule),

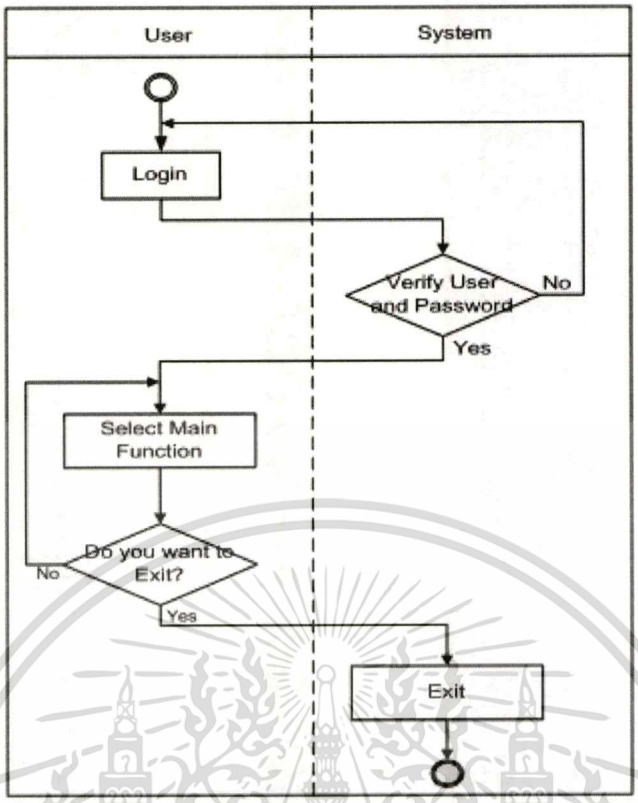
ผลลัพธ์การพยากรณ์ (Risk or Result), และแบบจำลองต้นไม้ที่สามารถนำกลับมาใช้ใหม่ได้ในการ Reload tree model(Tree model) ดังรูปที่ 3.5



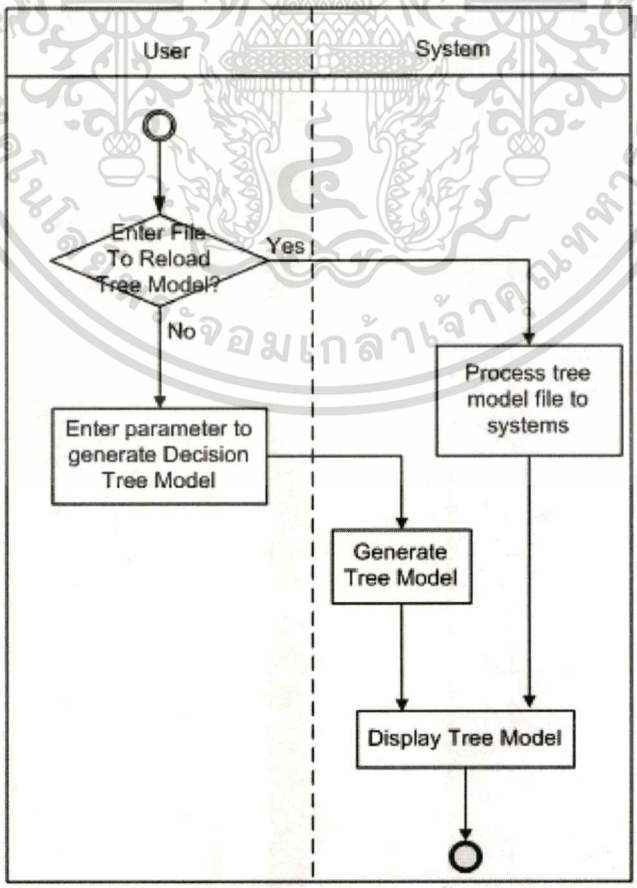
MAJOR	SEX	QUALITY	PROVINCE	OCCUPATION	INTEREST	SUM_INSURED	PREMIUM	AGE
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,000,000	20,001-30,000	50-59 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	No	>5,000,000	>50,000	>=60 years
Direct	Male	VIP	ภาคตะวันออก	Low	No	3,000,001-5,000,000	20,001-30,000	50-59 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	500,000-1,000,000	7,000-10,000	40-49 years
Broker	Male	General	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	30-39 years
Direct	Male	VIP	ภาคตะวันออก	Low	Yes	<500,000	7,000-10,000	50-59 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	40-49 years
Direct	Male	VIP	ภาคกลาง	Low	Yes	3,000,001-5,000,000	20,001-30,000	40-49 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	500,000-1,000,000	<7,000	30-39 years
Direct	Male	General	ภาคกลาง	Low	Yes	500,000-1,000,000	<7,000	30-39 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	50-59 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	Yes	3,000,001-5,000,000	20,001-30,000	40-49 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,000,000	20,001-30,000	50-59 years
Broker	Female	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,000,000	20,001-30,000	40-49 years
Broker	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,000,000	20,001-30,000	30-39 years

RULE	RULE_RESULT
AGE = '20-29 years'	No ['Yes=0,No=1 , Accurate=100%]
AGE = '40-49 years'	Yes ['Yes=25,No=2 , Accurate=92.59%]
AGE = '30-39 years' and SEX = 'Female'	No ['Yes=2,No=7 , Accurate=77.78%]
AGE = '30-39 years' and SEX = 'Male'	Yes ['Yes=15,No=0 , Accurate=100%]
AGE = '50-59 years' and SUM_INSURED <500,000	Yes ['Yes=4,No=1 , Accurate=80%]
AGE = '50-59 years' and SUM_INSURED = '500,000-1,000,000'	Yes ['Yes=3,No=0 , Accurate=100%]
AGE = '50-59 years' and SUM_INSURED = '1,000,001-3,000,000'	No ['Yes=0,No=1 , Accurate=100%]
AGE = '50-59 years' and SUM_INSURED = '3,000,001-5,000,000'	Yes ['Yes=6,No=4 , Accurate=60%]
AGE = '50-59 years' and SUM_INSURED >5,000,000	No ['Yes=0,No=0 , Accurate=0%]
AGE >=60 years and PREMIUM <7,000	No ['Yes=0,No=1 , Accurate=100%]
AGE >=60 years and PREMIUM = '7,000-10,000'	Yes ['Yes=3,No=0 , Accurate=100%]
AGE >=60 years and PREMIUM = '10,001-20,000'	Yes ['Yes=1,No=0 , Accurate=100%]
AGE >=60 years and PREMIUM = '20,001-30,000'	No ['Yes=0,No=3 , Accurate=100%]
AGE >=60 years and PREMIUM = '30,001-40,000'	No ['Yes=0,No=0 , Accurate=0%]

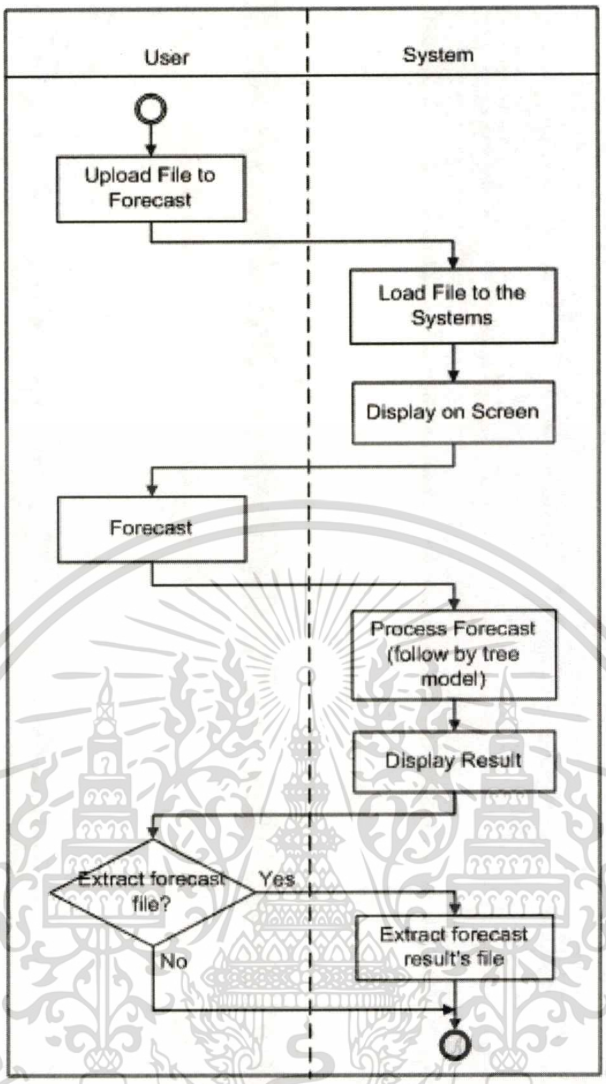
รูปที่ 3.1 การออกแบบโดยใช้ Classification algorithm เพื่อสร้างแบบจำลองต้นไม้



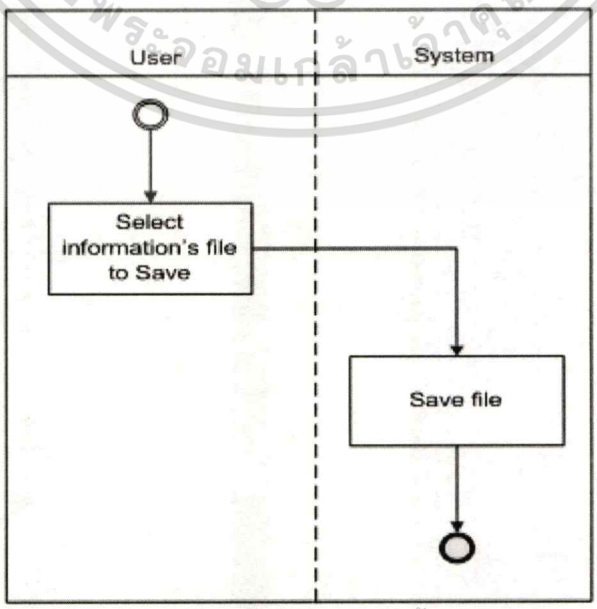
รูปที่ 3.2 แสดง Activity diagram ในการ Login เข้าสู่โปรแกรม



รูปที่ 3.3 แสดง Activity diagram ในการ Generate Tree Model



รูปที่ 3.4 แสดง Activity diagram ในการ Forecast



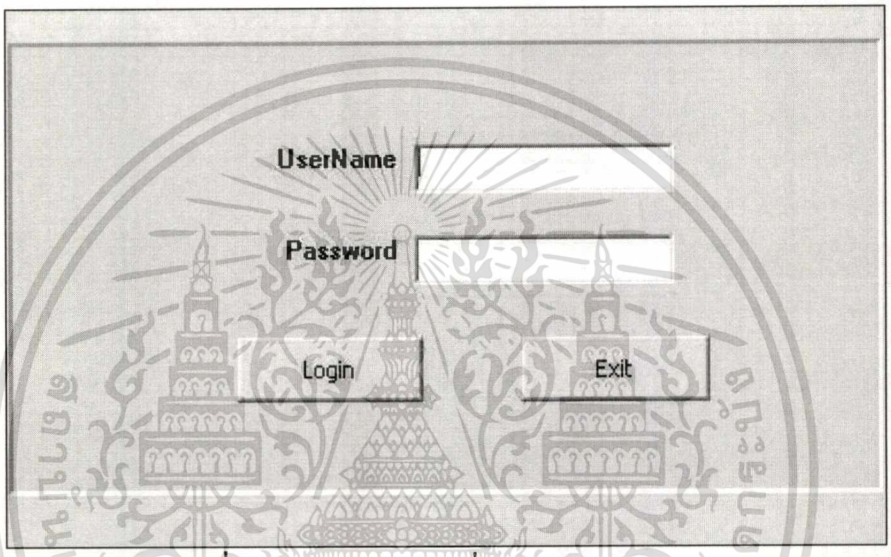
รูปที่ 3.5 แสดง Activity diagram ในการ Save File

3.4 การสร้างแบบจำลองต้นไม้โดยใช้โปรแกรมที่พัฒนาขึ้น

การทำงานของโปรแกรมจะประกอบไปด้วย 5 ส่วนหลักๆ ดังนี้

3.4.1 ส่วนติดต่อกับฐานข้อมูล

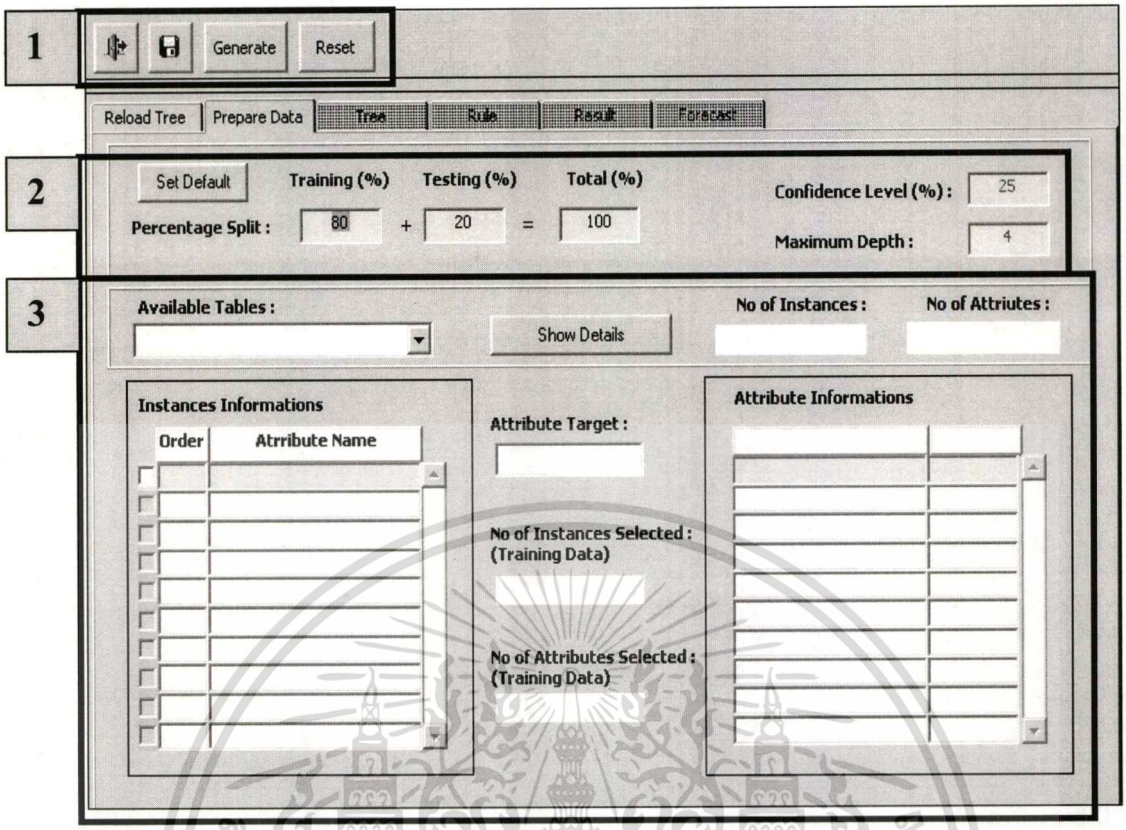
เป็นส่วนแรกที่จะทำการติดต่อกับฐานข้อมูล เพื่อเริ่มใช้งานระบบ โดยการ Log on เข้าสู่ระบบที่ต้องการทำการวิเคราะห์ ให้ระบุ User Name และ Password เพื่อตรวจสอบสิทธิ์การใช้งานระบบ ดังรูปที่ 3.6



รูปที่ 3.6 หน้าจอ Log in เพื่อเข้าสู่ระบบ

3.4.2 ส่วนจัดเตรียมหรือเลือกข้อมูล

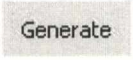
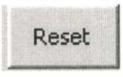
เป็นส่วนที่สองในการที่จะแสดงข้อมูลรายละเอียดของสิ่งที่ต้องการวิเคราะห์ ซึ่งตารางข้อมูลที่จะนำมาวิเคราะห์จะถูกกำหนดไว้แล้วว่ามีตารางใดบ้าง โดยแต่ละตารางจะมี Attribute ที่สนใจ (Target attribute) ที่กำหนดไว้แล้ว ซึ่งจะกำหนดเป็น List of values ดังรูปที่ 3.7



รูปที่ 3.7 หน้าจอหลักเพื่อเลือกข้อมูลที่จะนำมาวิเคราะห์

จากหน้าจอนี้จะแบ่งหลักๆ ได้เป็น 3 ฟังก์ชั่นหลัก

ฟังก์ชั่นที่ 1 แสดงในส่วนการใช้งานระบบ

-  ปุ่ม Exit สำหรับการออกจากระบบ
-  ปุ่ม Save สำหรับเลือกบันทึกรูปแบบข้อมูลต่างๆเป็น Text File
-  ปุ่ม Generate สำหรับประมวลผลสร้างแบบจำลองรูปต้นไม้
-  ปุ่ม Reset เพื่อเคลียร์ค่าให้กลับคืนสู่สถานะเริ่มต้นใช้งานใหม่

ฟังก์ชั่นที่ 2 แสดงส่วนของ การกำหนดเงื่อนไขในการสร้างแบบจำลองต้นไม้

- Percentage split

กำหนดจำนวนร้อยละของข้อมูลในตารางที่แบ่งไว้สำหรับ Training Data และ Testing Data โดยค่ามาตรฐานจะกำหนดว่า Training Data 80% และ Testing Data 20%

- Confidence Level (%)

กำหนดระดับความเชื่อมั่นที่จะยอมให้เกิดความผิดพลาดของการพยากรณ์ได้คิดเป็นร้อยละ โดยค่ามาตรฐานของอัลกอริทึม C4.5 จะกำหนดไว้ที่ 25%

- Maximum Depth

กำหนดความลึกของแบบจำลองต้นไม้ที่ต้องการโดยค่ามาตรฐานจะกำหนดไว้ที่ระดับ 4

- ปุ่ม [Set Default]

ใช้สำหรับตั้งค่า เงื่อนไข ตามมาตรฐานที่ได้กำหนดไว้

ฟังก์ชั่นที่ 3 แสดงส่วนของตาราง, ข้อมูล และ Attribute ที่สนใจ

- Available tables

แสดง รายการตารางที่สามารถเลือกวิเคราะห์ได้

- No of instances

จำนวนข้อมูลทั้งหมดในตารางที่เลือก

- No of attributes

จำนวนAttribute ทั้งหมดในตารางที่เลือก

- Instances information

แสดงรายละเอียดของข้อมูลในตารางที่เลือก ซึ่ง User สามารถคลิ๊ก Check box เพื่อเลือกเฉพาะAttribute ที่ต้องการมาสร้างแบบจำลอง

- Attribute target

แสดง Attribute ที่สนใจ

- No of instances (Training Data)

จำนวนข้อมูลทั้งหมดในตารางที่เลือกที่เป็นข้อมูล Training Data

- No of attributes (Training Data)

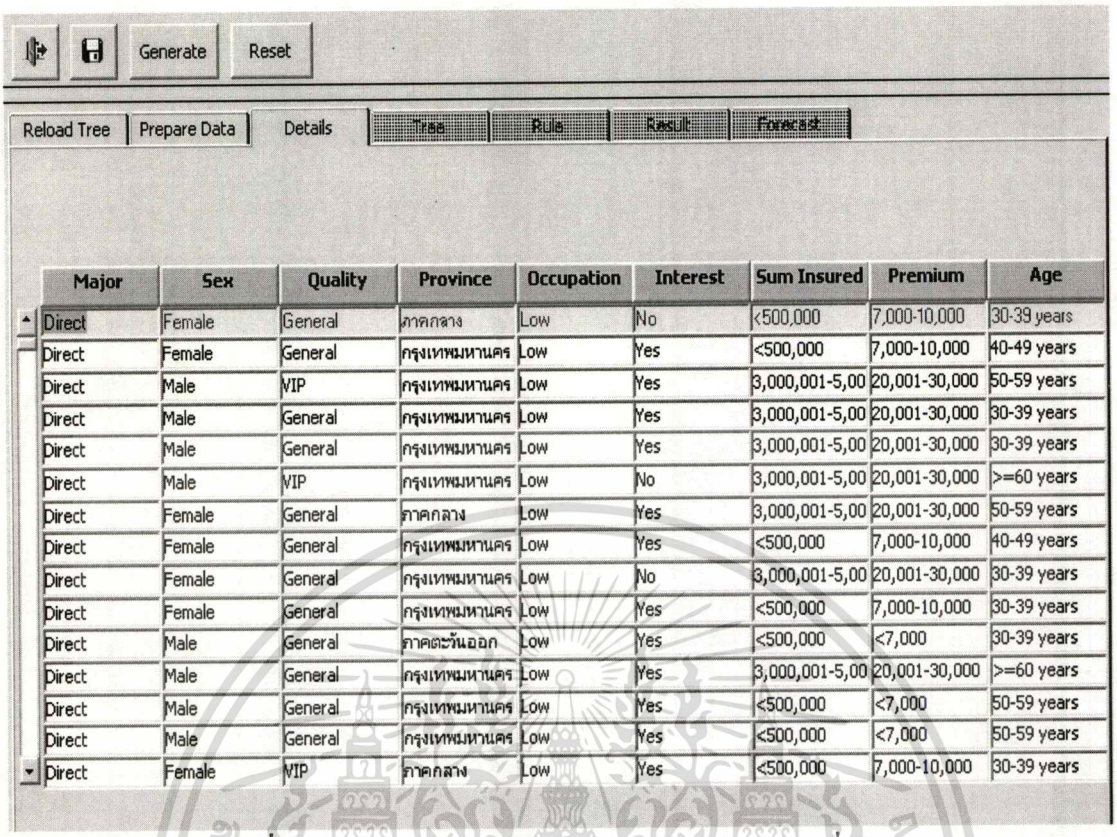
จำนวนAttribute ทั้งหมดในตารางที่เลือกที่เป็นข้อมูล Training Data

- Attributes information

แสดงรายละเอียดข้อมูลในแต่ละAttribute ที่เลือก จาก Instances information

- Show Details

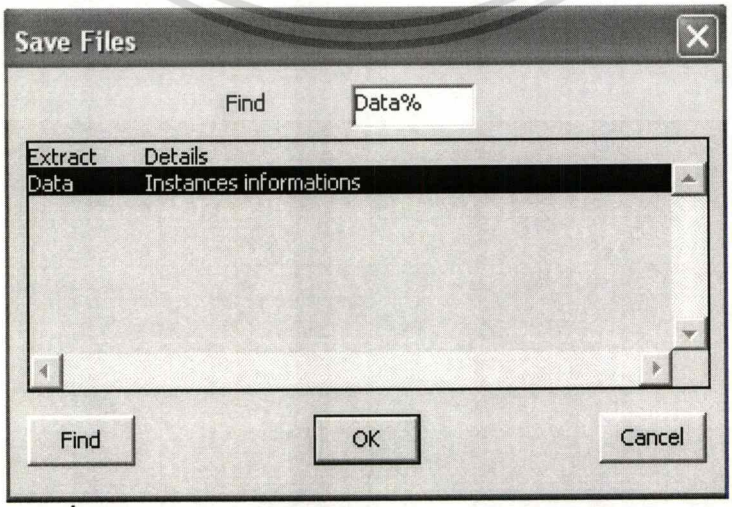
แสดง รายการข้อมูลทั้งหมดในตารางที่เลือก ดังรูปที่ 3.8



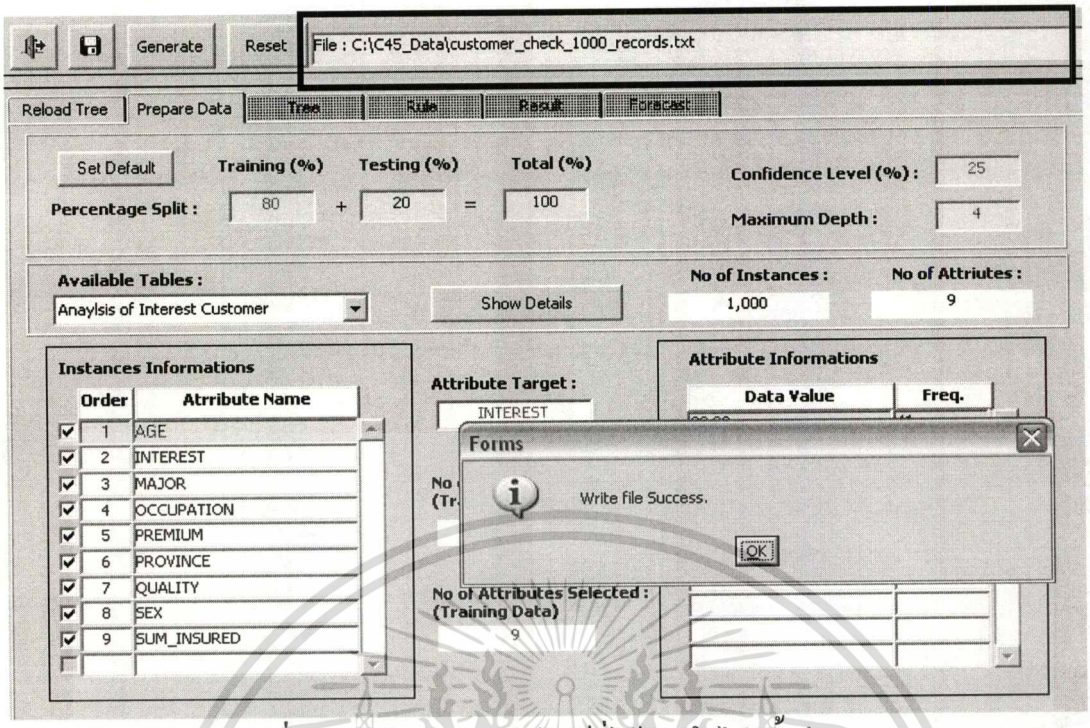
Major	Sex	Quality	Province	Occupation	Interest	Sum Insured	Premium	Age
Direct	Female	General	ภาคกลาง	Low	No	<500,000	7,000-10,000	30-39 years
Direct	Female	General	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	40-49 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	Yes	3,000,001-5,00	20,001-30,000	50-59 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,00	20,001-30,000	30-39 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,00	20,001-30,000	30-39 years
Direct	Male	VIP	กรุงเทพมหานคร	Low	No	3,000,001-5,00	20,001-30,000	>=60 years
Direct	Female	General	ภาคกลาง	Low	Yes	3,000,001-5,00	20,001-30,000	50-59 years
Direct	Female	General	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	40-49 years
Direct	Female	General	กรุงเทพมหานคร	Low	No	3,000,001-5,00	20,001-30,000	30-39 years
Direct	Female	General	กรุงเทพมหานคร	Low	Yes	<500,000	7,000-10,000	30-39 years
Direct	Male	General	ภาคตะวันออก	Low	Yes	<500,000	<7,000	30-39 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	3,000,001-5,00	20,001-30,000	>=60 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	<500,000	<7,000	50-59 years
Direct	Male	General	กรุงเทพมหานคร	Low	Yes	<500,000	<7,000	50-59 years
Direct	Female	VIP	ภาคกลาง	Low	Yes	<500,000	7,000-10,000	30-39 years

รูปที่ 3.8 หน้าจอแสดงรายละเอียดของข้อมูลในตารางที่เลือก

จากนั้น สามารถกดปุ่ม Generate Tree เพื่อทำการสร้างแบบจำลองต้นไม้ เพื่อวิเคราะห์ข้อมูลหากต้องการบันทึกข้อมูลในตารางที่เลือก สามารถกดปุ่ม Save จะแสดงหน้าจอดังรูปที่ 3.7 และระบบจะแสดงไฟล์ที่บันทึกที่ด้านบนในส่วนที่ 1 ดังรูปที่ 3.9 เพื่อเลือกบันทึกข้อมูลโดยที่ยังไม่ได้ทำการสร้างแบบจำลองได้ ซึ่งข้อมูลที่บันทึกจะเป็นเฉพาะในส่วนของข้อมูลดิบในตารางที่เลือกไว้เท่านั้น



รูปที่ 3.9 หน้าจอบันทึกข้อมูลก่อนการสร้างแบบจำลองต้นไม้



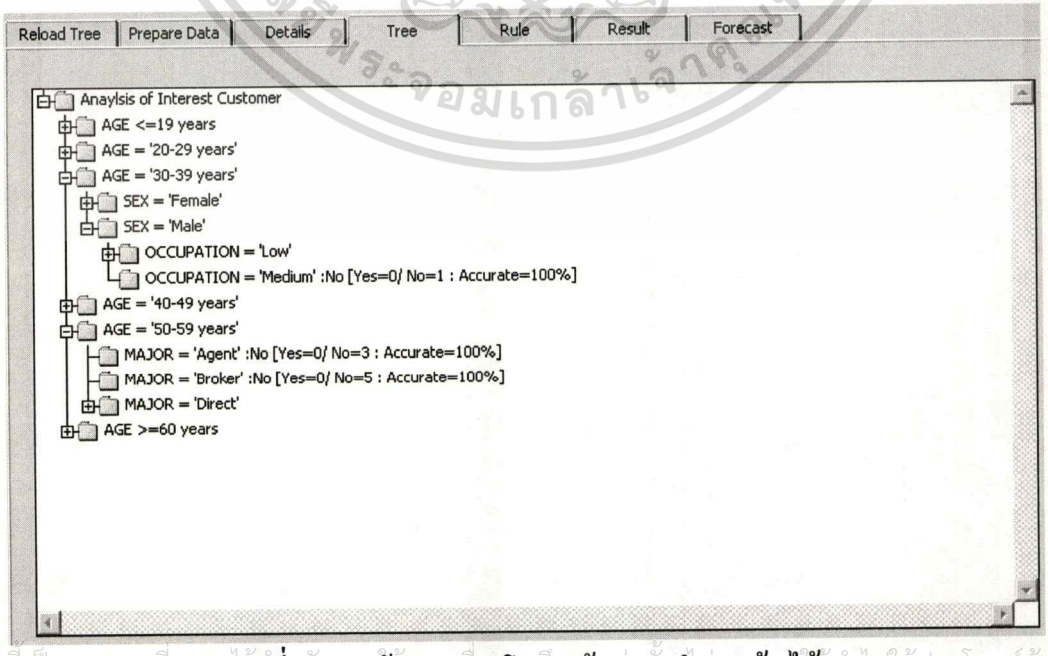
รูปที่ 3.10 หน้าจอแสดงผลไฟล์ที่ได้บันทึกไปครั้งล่าสุด

3.4.3 ส่วนการแสดงผลลัพท์

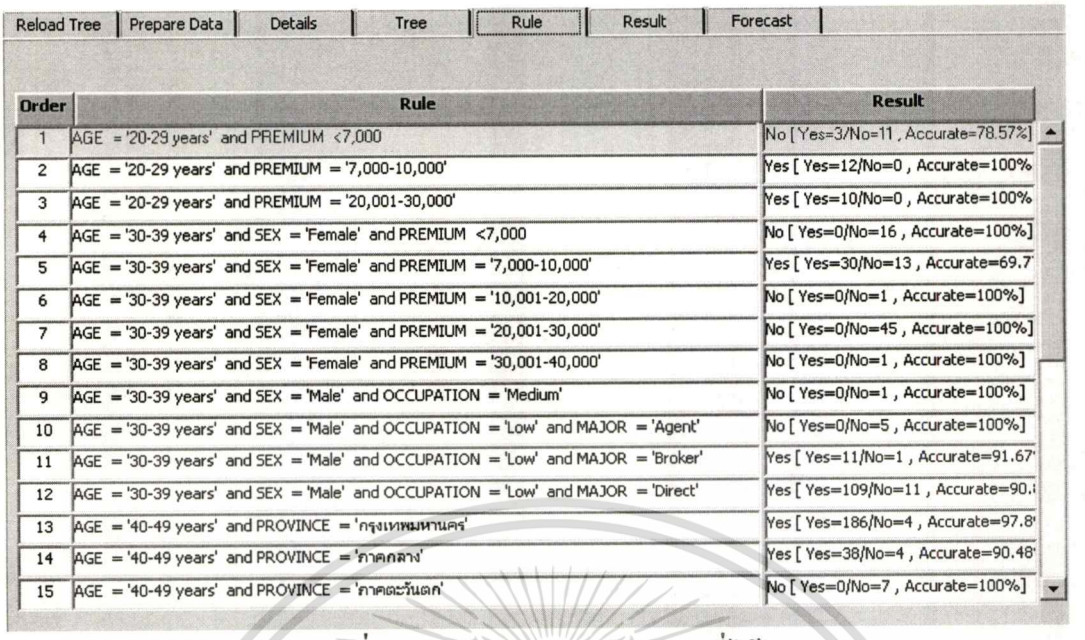
เป็นส่วนที่เกิดขึ้นหลังจากระบบได้ทำการสร้างแบบจำลองพยากรณ์เรียบร้อยแล้ว โดยสามารถแสดงผลได้ใน 3 รูปแบบด้วยกัน ดังนี้

แบบที่ 1 โครงสร้างแบบจำลองต้นไม้ ดังรูปที่ 3.11

แบบที่ 2 รูปแบบของกฎ ดังรูปที่ 3.12



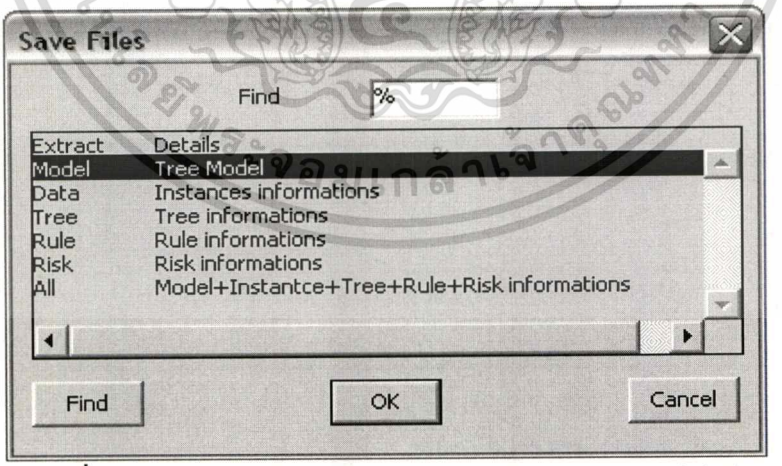
รูปที่ 3.11 หน้าจอแสดงโครงสร้างแบบจำลองต้นไม้



Order	Rule	Result
1	AGE = '20-29 years' and PREMIUM <7,000	No [Yes=3/No=11 , Accurate=78.57%]
2	AGE = '20-29 years' and PREMIUM = '7,000-10,000'	Yes [Yes=12/No=0 , Accurate=100%
3	AGE = '20-29 years' and PREMIUM = '20,001-30,000'	Yes [Yes=10/No=0 , Accurate=100%
4	AGE = '30-39 years' and SEX = 'Female' and PREMIUM <7,000	No [Yes=0/No=16 , Accurate=100%]
5	AGE = '30-39 years' and SEX = 'Female' and PREMIUM = '7,000-10,000'	Yes [Yes=30/No=13 , Accurate=69.7
6	AGE = '30-39 years' and SEX = 'Female' and PREMIUM = '10,001-20,000'	No [Yes=0/No=1 , Accurate=100%]
7	AGE = '30-39 years' and SEX = 'Female' and PREMIUM = '20,001-30,000'	No [Yes=0/No=45 , Accurate=100%]
8	AGE = '30-39 years' and SEX = 'Female' and PREMIUM = '30,001-40,000'	No [Yes=0/No=1 , Accurate=100%]
9	AGE = '30-39 years' and SEX = 'Male' and OCCUPATION = 'Medium'	No [Yes=0/No=1 , Accurate=100%]
10	AGE = '30-39 years' and SEX = 'Male' and OCCUPATION = 'Low' and MAJOR = 'Agent'	No [Yes=0/No=5 , Accurate=100%]
11	AGE = '30-39 years' and SEX = 'Male' and OCCUPATION = 'Low' and MAJOR = 'Broker'	Yes [Yes=11/No=1 , Accurate=91.67
12	AGE = '30-39 years' and SEX = 'Male' and OCCUPATION = 'Low' and MAJOR = 'Direct'	Yes [Yes=109/No=11 , Accurate=90.
13	AGE = '40-49 years' and PROVINCE = 'กรุงเทพมหานคร'	Yes [Yes=186/No=4 , Accurate=97.8
14	AGE = '40-49 years' and PROVINCE = 'ภาคกลาง'	Yes [Yes=38/No=4 , Accurate=90.48
15	AGE = '40-49 years' and PROVINCE = 'ภาคตะวันตก'	No [Yes=0/No=7 , Accurate=100%]

รูปที่ 3.12 หน้าจอแสดงรูปแบบกฎที่ได้

โดยสามารถกดปุ่ม Save เพื่อบันทึกผลซึ่งจะสามารถเลือกบันทึกข้อมูลในแต่ละรูปแบบได้ตามต้องการ ดังรูปที่ 3.13 โดยในแต่ละรูปแบบที่จะบันทึกจะทำการบันทึกแยกเป็นคนละ Text files หากเลือก All เพื่อบันทึกข้อมูลทุกรูปแบบ ระบบก็จะบันทึกแยก Text files ไว้ให้ โดย Path Default จะเป็น C:\C45_data และระบบจะสร้างชื่อไฟล์อัตโนมัติ สำหรับ Data, Tree, Rule และ Risk (หรือ Result) ของข้อมูลที่ได้นำมาวิเคราะห์

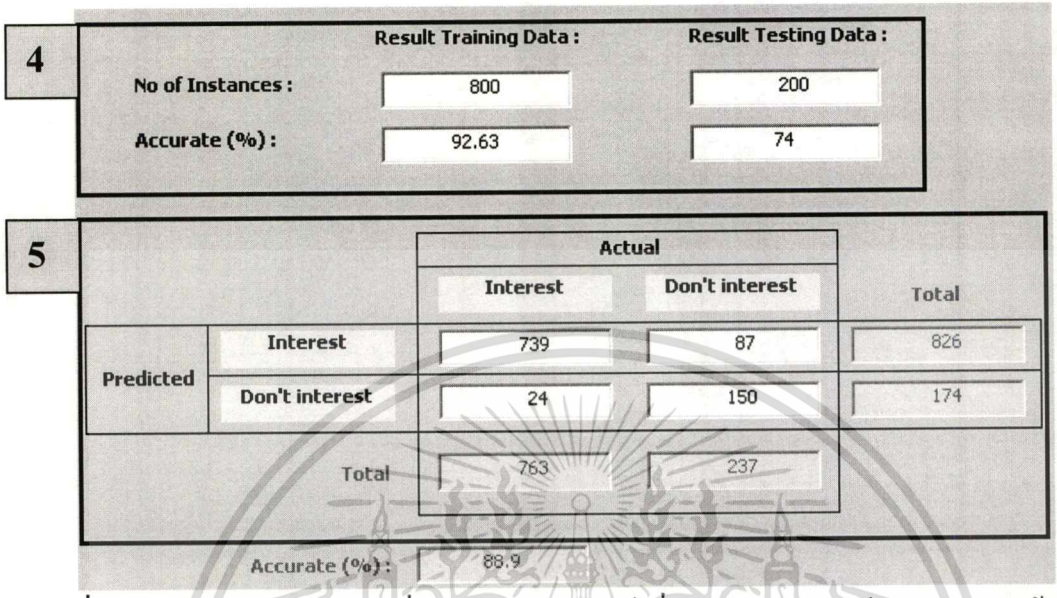


รูปที่ 3.13 หน้าจอบันทึกข้อมูลหลังการสร้างแบบจำลองต้นไม้

3.4.4 ส่วนการประเมินผลการพยากรณ์ของแบบจำลองที่สร้างขึ้น

เป็นส่วนที่ทำการประเมินประสิทธิภาพของแบบจำลองที่สร้างขึ้นว่ามีความน่าเชื่อถือมากน้อยเพียงใด โดยการนำข้อมูลที่ได้แบ่งไว้แล้วทั้ง Training Data และ Testing Data มาตรวจสอบกับ กฎ

ที่ได้จากแบบจำลอง เพื่อดูว่าผลลัพธ์ถูกต้องกับค่าความเป็นจริงมากน้อยเพียงใด โดยการแสดงผลลัพธ์ ดังรูปที่ 3.14



รูปที่ 3.14 หน้าจอแสดงผลลัพธ์ที่ได้จากการพยากรณ์ เมื่อ Training Data ร้อยละ 80 ของข้อมูลทั้งหมด

จากส่วนนี้ จะสามารถแบ่งฟังก์ชั่นการทำงานหลักๆ ได้อีก 2 ฟังก์ชั่น

ฟังก์ชั่นที่ 4 เป็นการสรุปผลการพยากรณ์ แสดงผลแยกระหว่าง Training Data และ Testing Data

- No of instances

  รายการข้อมูลจากตารางที่วิเคราะห์

- Accurate (%)

  ร้อยละ ความแม่นยำในการพยากรณ์

ฟังก์ชั่นที่ 5 เป็นการสรุปผลการพยากรณ์ แสดงผลรวมทั้ง Training Data และ Testing Data

- Predicted

  จำนวนข้อมูลที่ทำนาย ว่า น่าสนใจ และไม่น่าสนใจ

- Actual

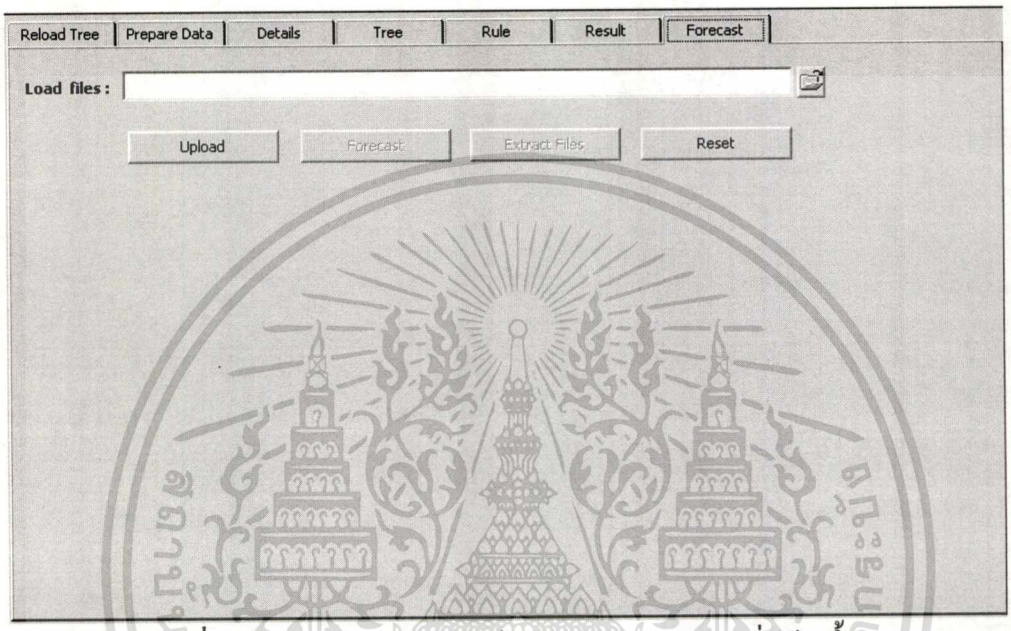
  จำนวนข้อมูลจริงและผลลัพท์ที่ได้กำหนดไว้ว่า น่าสนใจ และไม่น่าสนใจ

- Accurate (%)

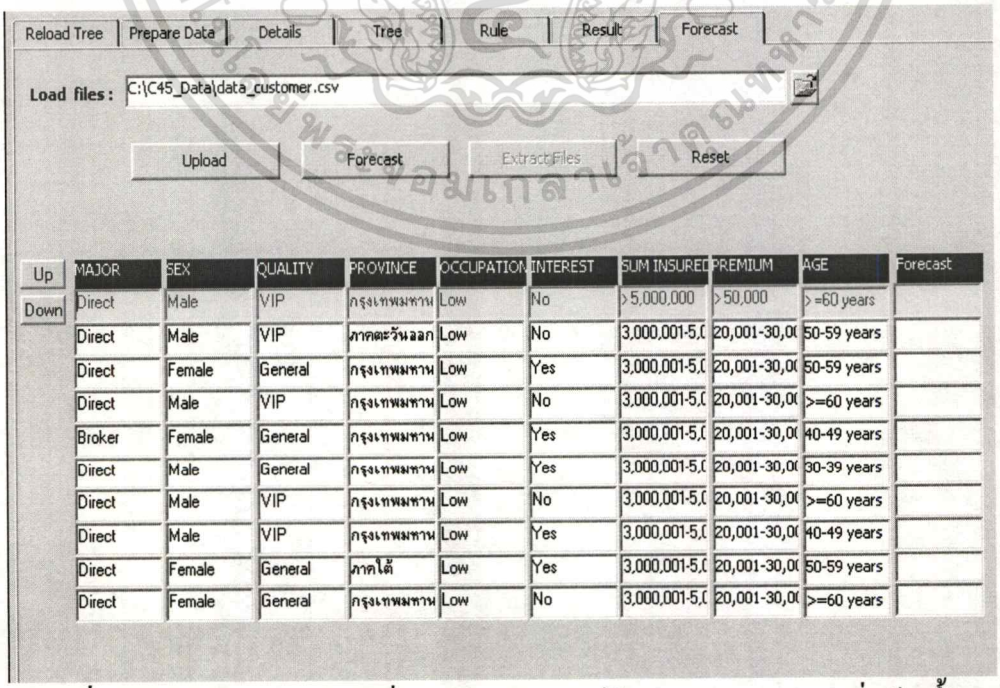
  ร้อยละความแม่นยำในการพยากรณ์ข้อมูลทั้งหมดจากตารางที่วิเคราะห์

3.4.5 ส่วนการพยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้น

เป็นส่วนที่สุดท้ายที่ของระบบ เพื่อพยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้นว่ามีอยู่ในกลุ่มที่เข้าข่ายความน่าสนใจหรือไม่ โดยทำการเลือกไฟล์ข้อมูลที่จะพยากรณ์ (*.csv file) เข้าสู่ระบบ จาก นั้นจึงทำการ กดปุ่ม Upload บันทึกข้อมูลดังกล่าวเข้าระบบ ดังรูปที่ 3.15 โปรแกรมก็จะแสดงผลข้อมูลที่มีการโหลดเข้า ดังรูปที่ 3.16



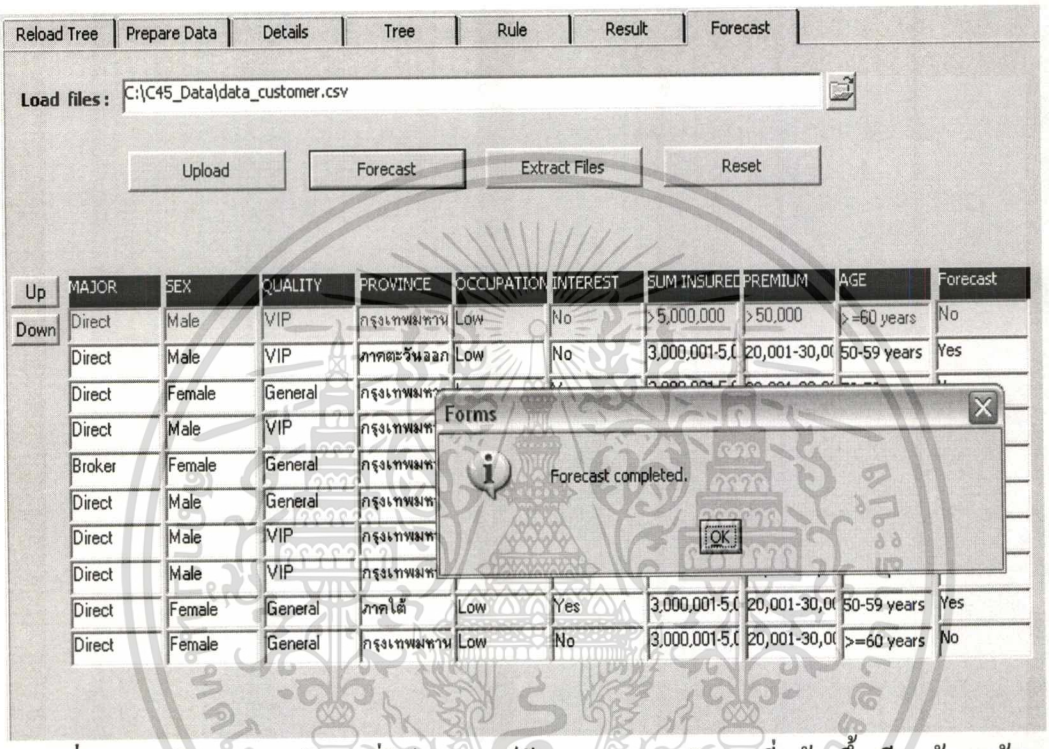
รูปที่ 3.15 หน้าจอการพยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้น



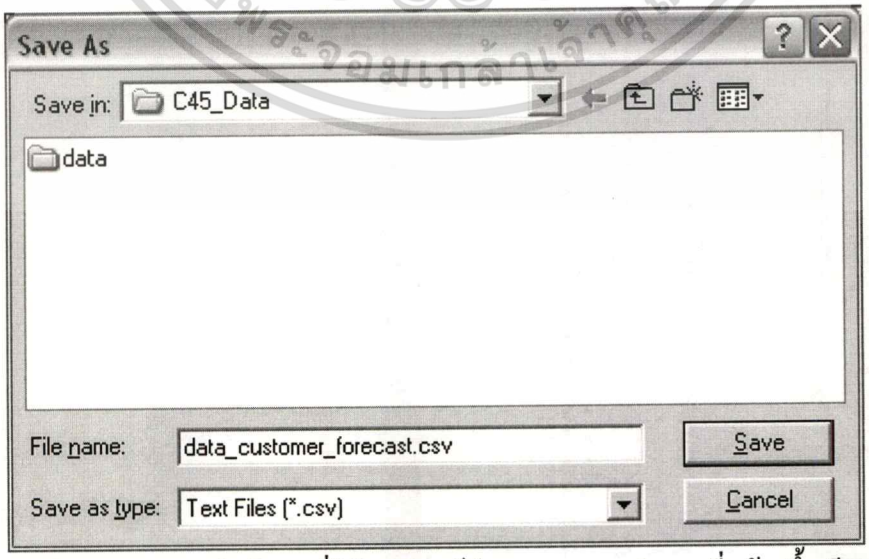
MAJOR	SEX	QUALITY	PROVINCE	OCCUPATION	INTEREST	SUM INSURED	PREMIUM	AGE	Forecast
Direct	Male	VIP	กรุงเทพมหาน	Low	No	> 5,000,000	> 50,000	> =60 years	
Direct	Male	VIP	ภาคตะวันออก	Low	No	3,000,001-5,(	20,001-30,0(	50-59 years	
Direct	Female	General	กรุงเทพมหาน	Low	Yes	3,000,001-5,(	20,001-30,0(	50-59 years	
Direct	Male	VIP	กรุงเทพมหาน	Low	No	3,000,001-5,(	20,001-30,0(	>=60 years	
Broker	Female	General	กรุงเทพมหาน	Low	Yes	3,000,001-5,(	20,001-30,0(	40-49 years	
Direct	Male	General	กรุงเทพมหาน	Low	Yes	3,000,001-5,(	20,001-30,0(	30-39 years	
Direct	Male	VIP	กรุงเทพมหาน	Low	No	3,000,001-5,(	20,001-30,0(	>=60 years	
Direct	Male	VIP	กรุงเทพมหาน	Low	Yes	3,000,001-5,(	20,001-30,0(	40-49 years	
Direct	Female	General	ภาคใต้	Low	Yes	3,000,001-5,(	20,001-30,0(	50-59 years	
Direct	Female	General	กรุงเทพมหาน	Low	No	3,000,001-5,(	20,001-30,0(	>=60 years	

รูปที่ 3.16 หน้าจอแสดงข้อมูลที่จะทำการพยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้น

เมื่อทำการโหลดข้อมูลเพื่อใช้พยากรณ์เรียบร้อยแล้ว ผู้ใช้สามารถกดปุ่ม Forecast เพื่อทำการพยากรณ์ข้อมูล เมื่อระบบพยากรณ์เรียบร้อยแล้วจะแสดงข้อความว่า “Forecast Complete” ดังรูปที่ 3.17 และให้ผู้ใช้สามารถเลือกกดปุ่ม Extract File เพื่อบันทึกข้อมูลที่พยากรณ์เรียบร้อยแล้วเป็นไฟล์นำมาเรียกดูภายหลังได้ ดังรูปที่ 3.18 หรือเลือกกดปุ่ม Reset เพื่อเคลียร์ค่าให้สามารถเลือกโหลดไฟล์ใหม่อีกครั้ง



รูปที่ 3.17 หน้าจอแสดงข้อมูลที่ได้พยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้นเรียบร้อยแล้ว



รูปที่ 3.18 หน้าจอแสดงบันทึกข้อมูลที่ได้พยากรณ์ข้อมูลจากแบบจำลองที่สร้างขึ้นเรียบร้อยแล้ว

บทที่ 4
สรุปผลการศึกษา และ ข้อเสนอแนะ

4.1 สรุปผลการศึกษา

โครงงานพัฒนาระบบฉบับนี้มีวัตถุประสงค์หลักเพื่อที่จะดึงขีดความสามารถของข้อมูลที่แฝงอยู่จากข้อมูลในอดีตมาใช้ให้เกิดประโยชน์สูงสุด โดยอาศัยอัลกอริทึ่ม C4.5 มาประยุกต์ใช้กับงานในธุรกิจประกันภัยในการสร้างแบบจำลองพยากรณ์ลักษณะของลูกค้าที่จะสามารถ สร้างความคุ้มค่าให้กับองค์กรได้ หนึ่งในแบบจำลองพื้นฐานและเป็นที่นิยมอย่างแพร่หลายก็คือ decision tree ซึ่งเป็นแบบจำลองที่เข้าใจง่าย แพร่หลายและมีความยืดหยุ่นสูง ดังนั้นโครงงานนี้จึงได้หยิบยกเอาแบบจำลองดังกล่าวมาทำการศึกษาและประยุกต์ใช้งาน โดยมีวัตถุประสงค์เพื่อให้องค์กรสามารถรับรู้ถึงลักษณะของกลุ่มลูกค้าที่มีคุณค่ากับองค์กร และสามารถนำสารสนเทศที่ได้ไปใช้ประโยชน์เพิ่มเติมเช่น ประกอบการตัดสินใจในการวางแผนงานต่างๆได้อีกทางหนึ่ง

ผลการพยากรณ์ข้อมูลด้วยแบบจำลองต้นไม้ ของกลุ่มลูกค้าที่สร้างความคุ้มค่าให้กับธุรกิจการประกันภัย ได้ Root node คือ อายุของลูกค้า โดยมีลักษณะหรือประเภทของลูกค้าที่เข้าข่ายที่จะสร้างความคุ้มค่าให้กับธุรกิจประกันภัย จากการสุ่มตัวอย่างข้อมูล 1,000 รายการ ได้แก่

- ลูกค้าที่อายุ ไม่เกิน 19 ปี และเป็นประเภทลูกค้าประกันตรง
- ลูกค้าที่อายุ ระหว่าง 20-29 ปี และ เบี้ยประกันภัยอยู่ระหว่าง 7,000 – 10,000 บาท
- ลูกค้าที่อายุ ระหว่าง 20-29 ปี และ เบี้ยประกันภัยอยู่ระหว่าง 20,001 – 30,000 บาท
- ลูกค้าที่อายุ ระหว่าง 30-39 ปี, เพศหญิง และ เบี้ยประกันภัยอยู่ระหว่าง 7,000 – 10,000 บาท
- ลูกค้าที่อายุระหว่าง 30-39 ปี, เพศชาย, ประเภทอาชีพความเสี่ยงต่ำ และช่องทางลูกค้าคือ นายหน้า
- ลูกค้าที่อายุระหว่าง 30-39 ปี, เพศชาย, ประเภทอาชีพความเสี่ยงต่ำ และช่องทางลูกค้าคือ ประกันตรง
- ลูกค้าที่อายุระหว่าง 40-49 ปี และภูมิลำเนา กรุงเทพมหานคร, ภาคกลาง, ภาคตะวันออก, ภาคตะวันออกเฉียงเหนือ และ ภาคเหนือ
- ลูกค้าที่อายุระหว่าง 50-59 ปี, ช่องทางลูกค้าประเภทประกันตรง และภูมิลำเนา กรุงเทพมหานคร, ภาคกลาง, ภาคใต้ และ ภาคตะวันออก
- ลูกค้าที่อายุมากกว่า 60 ปี และเบี้ยประกันภัยระหว่าง 7,000 – 10,000
- ลูกค้าที่อายุมากกว่า 60 ปี และเบี้ยประกันภัยระหว่าง 30,001 – 40,000

- ลูกค้าที่อายุมากกว่า 60 ปี, เบี้ยประกันภัยระหว่าง 20,001 – 30,000 และภูมิลำเนาภาคตะวันตก, ภาคเหนือ และ ภาคใต้

จากผลจากการศึกษาโครงงานที่ได้พัฒนาระบบนี้ ยังมีอีกปัจจัยหนึ่งที่มีผลสำคัญในการพยากรณ์ด้วยก็คือ ความถูกต้องครบถ้วน และปริมาณของข้อมูล เพราะสิ่งเหล่านี้ล้วนเป็นปัจจัยหลักที่ส่งผลกระทบต่อผลการความแม่นยำในการพยากรณ์ด้วย ดังนั้นจากแบบจำลองที่ได้สามารถสรุปผลการทดลอง โดยการคำนวณหาค่าความถูกต้องในการพยากรณ์จากการวิเคราะห์ข้อมูลที่เป็นไปตามกฎที่ได้จากแบบจำลองเปรียบเทียบกับข้อมูลจริงว่า มีการพยากรณ์ตรงกับข้อมูลจริงคิดเป็นร้อยละของข้อมูลทั้งหมด ซึ่งสรุปได้ว่าค่าความถูกต้องในการพยากรณ์ข้อมูลทั้งหมดจากแบบจำลองที่ได้คือ ร้อยละ 88.9 ดังรูปที่ 3.14

4.2 ข้อเสนอแนะ

ระบบที่พัฒนาขึ้นนี้สามารถนำไปใช้งานกับธุรกิจอื่นๆที่ใช้ฐานข้อมูล Oracle ได้ด้วย เนื่องจากในการพัฒนาได้ใช้ภาษาหลักคือ PL/SQL พัฒนาบนฐานข้อมูล Oracle รวมทั้งใช้เครื่องมือที่ของ Oracle เช่น Developer6i อีกด้วย สำหรับในเรื่องการนำไปใช้งานจริงจำเป็นต้องอาศัยผู้ที่มีความเชี่ยวชาญ มีความเข้าใจในธุรกิจนั้นๆ ตรวจสอบความถูกต้องของข้อมูลและเลือกAttribute ที่เหมาะสมเพื่อมาทำการสร้างแบบจำลอง ก็จะทำให้ผลการพยากรณ์มีความน่าเชื่อถือมากขึ้น แต่เนื่องจากข้อมูลที่นำมาทดสอบของโครงการพัฒนาระบบนี้ อาจจะมีจำนวนไม่มากพอที่จะใช้วัดประสิทธิภาพการทำงานจริงได้

แม้ว่าระบบที่พัฒนาขึ้นจากหลักการของอัลกอริทึม C4.5 ซึ่งเป็น Decision Tree ที่สามารถสร้างแบบจำลองเป็นโครงสร้างต้นไม้ที่แตกกิ่งออกไป และกำหนดมาเป็นกฎในรูปแบบที่เข้าใจได้ง่ายแล้วก็ตาม แต่ก็ยังมีข้อเสียของการพยากรณ์ที่ต้องคำนึงถึงด้วย คือ ปริมาณของข้อมูล ซึ่งความผิดพลาดส่วนใหญ่มักจะเกิดจากความไม่สมบูรณ์และปริมาณของข้อมูลที่ไม่เพียงพอ โดยเฉพาะอย่างยิ่งในข้อมูลที่มีปริมาณมากๆ

บรรณานุกรม

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมานนท์. **Data Mining**. [Online]. Available : http://micro.se-ed.com/content/mc187/mainframe.ASP?tar=MC187_92.asp. 1998.

ศูนย์เทคโนโลยีสารสนเทศ. **Data warehouse of AZCP Second Edition.** Bangkok : Amarin Printing Group. 2547.

Andrew Kusiak. **Decision Tree Algorithm.** [Online]. Available : http://www.icaen.uiowa.edu/~ankusiak. 1999.

Cabena, Hadjinian, Stadler, Verhees and Zanasi. **Discovery Data Mining From Concept to Implementation.** New Jersey : Prentice Hall. 1998.

Department of Insurance. **Department of Insurance of Thailand.** [Online]. Available : http://www.doi.go.th/egate.htm. 2005.

DWreview Co.Ltd. **Data Warehousing Review.** [Online]. Available : http://www.dwreview.com. 2005.

Groth R. **Data Mining a hands on approach for business professionals.** : Prentice Hall Publishing. 1997.

J.R Quinlan. **C4.5: Programs for Machine Learning.** Morgan Kauffman : San Mateo. C.A. 1993.

Michael Nashvili. 2004. **Decision Trees.** [Online]. Available : http://www.decisiontrees.net

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวนุสรา จิรเจริญจิตต์
วัน เดือน ปีเกิด	16 ตุลาคม 2520 ที่กรุงเทพมหานคร
ประวัติการศึกษา	2542 วิทยาศาสตร์บัณฑิต สาขาคณิตศาสตร์ประยุกต์
	มหาวิทยาลัยธรรมศาสตร์
ประวัติการทำงาน	ปัจจุบัน – นักวิเคราะห์ระบบ บริษัท อลิอันซ์ ซี.พี. ประกันภัย