ระบบการพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรี

Developing Analysis Churn Management System in Mobile Services by Decision Tree

โดย

นางสาวดวงหทัย วงศ์สวัสดิ์สุริยะ

รหัสประจำตัวนักศึกษา 44067483

อาจารย์ที่ปรึกษา

ผศ. ดร. วรพจน์ กรีสุระเดช

H002157

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อ	การพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรี
นักศึกษา	นางสาว ดวงหทัย วงศ์สวัสดิ์สุริยะ
อาจารย์ที่ปรึกษา	ผศ. ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

ระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีนี้ จะเป็นระบบที่ช่วยสนับสนุนข้อมูลในด้านการตัดสินใจให้กับผู้บริหารในการวิเคราะห์หาสาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้า ซึ่งข้อมูลที่นำมาวิเคราะห์จะเกี่ยวข้องกับ โปรโมชั่น ค่าใช้จ่าย อายุของลูกค้า เป็นต้นโดยนำหลักการของ Decision Tree และอัลกอริทึม SLIQ เข้ามาใช้ ในการพัฒนาระบบนั้นจะนำ VB6 มาใช้เป็น Front-end โดยจะทำการติดต่อกับฐานข้อมูลที่เป็น SQL Server 2000

Title	Developing Analysis Churn Management System in Mobile Services by Decision Tree
Student	Miss. Duanghatai Wongsawatsuriya
Advisor	Dr. Worapoj Kresuradej
Level of Study	Master of Science in information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

Decision Tree System, the system to analyze the reason of customer changing the service of mobile phone system, is the supporting data system for the management team in order to make decision and analyze the reason of the changing mobile phone's service from customers. To analyze this, the information, for example like promotion, expense, and demographic of customer will be conducted and analyze under the concept of Decision Tree and Algorithm SLIQ. While VB6 will be used as Front-end for the development of the system by connecting with database, which is SQL Server 2000.

กิตติกรรมประกาศ

ในการพัฒนาระบบงานในครั้งนี้ทางผู้จัดทำขอขอบพระคุณ ผศ. ดร. วรพจน์ กรีสุระเดช เป็นอย่างสูง ที่เป็นผู้ให้ความรู้และให้คำแนะนำในการพัฒนาระบบงาน ขอขอบคุณเพื่อนๆ IS 12.2 ทุกคนที่คอยเป็นแรงกระตุ้นให้ผู้จัดทำมีความกระตือรือร้นที่จะพัฒนาระบบ ขอขอบคุณพี่ๆ เพื่อนๆ ทุกคนใน DTAC ที่ให้คำแนะนำและเป็นที่ปรึกษาที่ดีในการพัฒนาระบบ พร้อมทั้งให้ความช่วยเหลือ และคำวิจารณ์อีกมากมาย และท้ายที่สุดนี้ขอขอบคุณ คุณพ่อ คุณแม่ ที่เป็นพลังเกื้อหนุนที่ดีที่สุด ที่ทำให้ผู้จัดทำมีวันนี้

ขอบคุณมากค่ะ

ดวงหทัย วงศ์สวัสดิ์สุริยะ

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

บทที่

สารบัญ(ต่อ)

สารบัญ(ต่อ)

หน้า

บทที่

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

รูปที่

สารบัญภาพ (ต่อ)

หน้า

รูปที่

บทที่ 1
บทนำ

1.1 ความเป็นมา

ภายใต้สภาพแวดล้อมทางด้านธุรกิจและการตลาดที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา ทำให้บริษัทต่างๆ ต้องเผชิญกับการแข่งขันในด้านธุรกิจกันอย่างรุนแรงเพื่อที่จะเพิ่มกำไร ส่วนแบ่งทางการตลาด (Market Share) ให้มากขึ้น ซึ่งแนวคิดหนึ่งที่ใช้ในการสร้างความประทับใจในสินค้า และสร้างสัมพันธ์ภาพกับลูกค้าปัจจุบัน ทั้งลูกค้าที่เป็นตลาดธุรกิจด้วยกัน (B-2-B) และลูกค้าในตลาดผู้บริโภคคือ การบริหารงานลูกค้าสัมพันธ์ (Customer Relationship Management: CRM) ซึ่งในความเป็นจริงแล้วประโยชน์หรือคุณค่าของ CRM มีมากกว่าการรักษาลูกค้าปัจจุบัน (Customer Retention) บริษัทยังสามารถนำ CRM มาใช้เพื่อดึงลูกค้าเก่าให้กลับมาซื้อสินค้าหรือบริการจากบริษัทอีก (Customer Win-back Strategy) ตลอดจนใช้ในการหาลูกค้าใหม่ (New Customer Acquisition)

ในปัจจุบันธุรกิจการแข่งขันในด้านผู้ให้บริการโทรศัพท์เคลื่อนที่นั้นค่อนข้างรุนแรง มีการแย่งลูกค้ากันมาก ซึ่งเป็นที่ทราบกันดีอยู่แล้วว่าการรักษาลูกค้าเก่าให้อยู่กับบริษัทนั้นง่ายกว่าการหาลูกค้าใหม่ ดังนั้นแต่ละองค์กรจึงจำเป็นที่จะต้องนำข้อมูลของลูกค้ามาวิเคราะห์ว่าทำอย่างไรลูกค้าถึงจะพอใจในสินค้าและบริการของเรา และข้อมูลสำคัญอีกส่วนที่ต้องทราบก็คือ เหตุใดลูกค้าจึงยกเลิกการใช้สินค้าและบริการของบริษัทเรา ซึ่งศัพท์ทางการตลาดจะใช้คำว่า "Churn Management"

Churn Management คือการจัดการในเรื่องการเปลี่ยนใจของลูกค้าในการใช้บริการสินค้าหรือการยกเลิกการใช้บริการ ซึ่งการจัดการในเรื่องนี้ได้นั้นจำเป็นที่จะต้องมีข้อมูลและเทคนิคที่จะนำมาใช้ในการวิเคราะห์ซึ่งดาต้าไมนิ่งก็เป็นวิธีการหนึ่งที่เหมาะสมในการวิเคราะห์เรื่อง Churn Management

ในโครงงานนี้จะกล่าวถึง การพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีโดยนำอัลกอริทึมที่ชื่อว่า SLIQ เป็นหลักในการเขียนโปรแกรม ซึ่งระบบนี้จะช่วยสนับสนุนข้อมูลในด้านการตัดสินใจให้กับผู้บริหารในการวิเคราะห์หาสาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้า โดยผลการวิเคราะห์จะแสดงออกมารูปของแบบจำลองต้นไม้ (Tree Model)

1.2 วัตถุประสงค์ของโครงงาน

1. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับแนวคิดและขั้นตอนการทำดาต้าไมนิ่ง
2. เพื่อศึกษาและทำความเข้าใจในหลักการตลาดเบื้องต้น เพื่อเป็นความรู้ในการทำงานและเปิดโลกทัศน์ให้กว้างขึ้น
3. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับการใช้แนวคิด Classification โดยนำหลักการดิซิชั่นทรีและอัลกอริทึมที่ชื่อว่า SLIQ (Supervised Learning in Quest) เข้ามาใช้
4. เพื่อศึกษาถึงแนวทางและความเป็นไปได้ในการนำแนวคิดดิซิชั่นทรีมาใช้ในการวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้า ว่าสามารถนำมาวิเคราะห์ข้อมูลได้จริงและมีประสิทธิภาพหรือไม่
5. เพื่อเป็นแนวทางในการประยุกต์ใช้ในการสนับสนุนการตัดสินใจให้ผู้บริหารในการวางกลยุทธ์ในด้านการตลาด

1.3 ขอบเขตของโครงงาน

1.3.1 หน้าที่การทำงานของระบบ

ระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีมีขอบเขตการทำงานหลักดังต่อไปนี้

- ระบบจะเปิดให้ผู้ใช้สามารถนำข้อมูลที่จะนำมาวิเคราะห์ได้ 2 ทางคือ ฐานข้อมูลที่ผู้ใช้ระบบมีอยู่แล้วซึ่งในระบบนี้จะกำหนดให้ใช้ได้เฉพาะฐานข้อมูลที่มาจาก SQL Server 2000 เท่านั้น และการโหลดข้อมูลจาก Text File ซึ่งมีรูปแบบตามที่ผู้พัฒนาระบบกำหนดไว้เท่านั้น ซึ่งรายละเอียดจะกล่าวไว้ในบทที่ 5 คู่มือการใช้งานระบบ
- ระบบจะทำการวิเคราะห์ข้อมูลโดยใช้แนวคิด Classification โดยนำหลักการดิซิชั่นทรีและอัลกอริทึมที่ชื่อว่า SLIQ (Supervised Learning in Quest) และแสดงผลลัพธ์ออกมาในรูปของแบบจำลองต้นไม้เท่านั้น
- ในส่วนของการทดสอบผลของการวิเคราะห์แบบจำลองต้นไม้ นั้นจะใช้แนวคิดดังนี้คือ แบ่งข้อมูลออกเป็น 2 ส่วน ข้อมูลส่วนแรก (Training Data)ใช้ในการสร้างแบบจำลองต้นไม้ ข้อมูลส่วนที่สองจะถูกแบ่งไว้ใช้ในการทดสอบ (Testing Data) แบบจำลองต้นไม้ที่ระบบสร้างขึ้น

1.3.2 ข้อมูลที่นำมาวิเคราะห์

ข้อมูลที่นำมาใช้วิเคราะห์ในการพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีมีรายละเอียดดังนี้

- ข้อมูลที่ใช้เป็นข้อมูลการปิดบริการของลูกค้าในองค์กรจริง ซึ่งข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลของลูกค้าที่ปิดบริการในระบบแบบใช้ก่อน จ่ายเงินทีหลัง (Post-Paid) ระหว่างวันที่ 01/07/2002 – 31/12/2002 จำนวน 1,500 เรคคอร์ด
- ข้อมูลที่นำมาใช้ในการวิเคราะห์นั้นจะต้องเป็นข้อมูลที่ผ่านการเตรียมข้อมูลมาแล้ว (Data Preparation) ตามหลักการคาต้าไมนิ่ง

1.4 ขั้นตอนและแผนงานในการพัฒนา

1.4.1 การกำหนดระยะเวลาในการดำเนินงาน

- กำหนดระยะเวลาในการพัฒนาระบบ โดยคำนึงถึงการดำเนินงานจริงดังนี้

ตารางที่ 1.1 ตารางการทำงาน

		Name	Duration	Start	Finish	Resource Names
1		**Project : Data Mining**	**96.94 days?**	**Thu 13/11/03**	**Fri 26/3/04**	**Duanghatai W.**
2		Sending Proposal	1 day?	Thu 13/11/03	Fri 14/11/03	
3		Analysis & Design Program Environment	4.94 days?	Mon 17/11/03	Fri 21/11/03	
4		Design Database	5 days	Mon 24/11/03	Mon 1/12/03	
5		Design Interface	5.94 days?	Tue 2/12/03	Tue 9/12/03	
6		Prepare Environment	2.94 days?	Wed 10/12/03	Fri 12/12/03	
7		Coding Part of Program (Load Data)	3.94 days?	Tue 16/12/03	Fri 19/12/03	
8		**Develop Document (Progress Report 2 Copies)**	8.94 days?	Mon 15/12/03	Thu 25/12/03	
9		Implement & Testing Program	51.94 days?	Fri 26/12/03	Mon 8/3/04	
10		**Develop Document (Original Version 4 Copies)**	17.94 days?	Tue 20/1/04	Thu 12/2/04	
11		**Presentation**	4.94 days?	Mon 8/3/04	Fri 12/3/04	
12		**Develop Document (Complete version)**	9.94 days?	Mon 15/3/04	Fri 26/3/04	

1.4.2 การออกแบบระบบ

- ศึกษาทฤษฎีและความต้องการของระบบโดยอ้างอิงจากทฤษฎีคาต้าไมนิ่ง แนวคิด Classification โดยนำหลักการหลักการดิซิชั่นทรีและอัลกอริทึมที่ชื่อว่า SLIQ (Supervised Learning in Quest) เพื่อวิเคราะห์ถึงหน้าที่การทำงานหลักที่โปรแกรมเกี่ยวกับการวิเคราะห์ข้อมูลต้องมี และทำการหาขอบเขตของระบบที่จะพัฒนาขึ้น
- กำหนดเครื่องมือและทรัพยากรที่ใช้ในการพัฒนาระบบ
- ออกแบบหน้าจอของระบบ

- ออกแบบฐานข้อมูล
- ออกแบบโครงสร้างของโปรแกรมดังนี้ ส่วนการนำเข้าข้อมูล ส่วนการวิเคราะห์ข้อมูล ส่วนการทดสอบผลการวิเคราะห์

1.4.3 การเขียนโปรแกรมและทดสอบระบบ

- เขียนโปรแกรมตามที่ได้ออกแบบไว้
- ทำการทดสอบระบบที่ได้พัฒนาขึ้น โดยแบ่งการทดสอบออกเป็น 3 ส่วนดังนี้
  - ทดสอบส่วนย่อยของโปรแกรม (Unit Testing) ว่าสามารถทำงานได้ถูกต้องตามที่ออกแบบไว้หรือไม่
  - ทดสอบการทำงานร่วมกันของแต่ละฟังก์ชั่น (Integrate Testing) ว่าสามารถทำงานได้สอดคล้องถูกต้องหรือไม่
  - ทดสอบภาพรวมของระบบ (System Test) ว่าสามารถทำงานได้ถูกต้องตรงกับความต้องการหรือไม่

1.4.4 การทำเอกสารประกอบระบบ

- ทำเอกสารประกอบการออกแบบระบบ
- ทำเอกสารคู่มือการใช้งานระบบ

1.4.5 การติดตั้งระบบ

- ทำการติดตั้งระบบ
- ทำการบำรุงดูแลรักษาระบบ

1.5 เครื่องมือที่ใช้ในการพัฒนาระบบงาน

- เครื่องมือที่ใช้ในการเขียนโปรแกรมคือ VB 6
- ระบบฐานข้อมูลคือ MS SQL Server 2000
- เครื่องคอมพิวเตอร์ที่นำมาใช้ในการพัฒนา คือ Microsoft Windows XP Professional 2002 Pentium 4 CPU 1.8 GHz 256 MB

1.6 ประโยชน์ที่คาดว่าจะได้รับจากโครงการ

1. ผู้พัฒนาระบบมีความเข้าใจในหลักการของดาต้าไมนิ่งแบบ Classification โดยใช้ดิซิชั่นทรีและอัลกอริทึมที่ชื่อว่า SLIQ มากยิ่งขึ้น
2. ผู้ใช้ระบบสามารถนำผลการวิเคราะห์มาใช้ในการวางกลยุทธ์ทางด้านธุรกิจได้
3. สามารถนำระบบที่จัดทำขึ้นมาช่วยเพิ่มประสิทธิภาพในการทำงานขององค์กร
4. เป็นกรณีศึกษาเพื่อเป็นแนวทางในการพัฒนาระบบงานอื่นๆ ต่อไป

บทที่ 2
ทฤษฎีดาต้าไมนิ่ง

ในส่วนของบทที่ 2 นี้จะกล่าวถึงความหมายและทฤษฎีของดาต้าไมนิ่ง (Data Mining) พร้อมทั้งแนวคิดหลักๆ และขั้นตอนในการทำดาต้าไมนิ่ง

2.1 ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่ง คือ ขั้นตอนการทำงานที่ใช้ในการค้นหาองค์ความรู้จากแหล่งข้อมูลขนาดใหญ่ ซึ่งองค์ความรู้ที่ได้นั้นจะต้องถูกต้อง มีเหตุมีผล และสามารถนำไปใช้ได้ เพื่อเป็นข้อมูลที่ช่วยในการตัดสินใจ ซึ่งดาต้าไมนิ่งอาจเรียกอีกอย่างหนึ่งได้ว่า Knowledge Discovery in Databases (KDD)

2.2 ขั้นตอนการทำดาต้าไมนิ่ง

การทำดาต้าไมนิ่ง มี 5 ขั้นตอนดังนี้

2.2.1. ขั้นตอนการกำหนดวัตถุประสงค์ในการวิเคราะห์ (Business Objective Determination) คือ ขั้นตอนในการกำหนดวัตถุประสงค์และปัญหาที่ต้องการทราบ เพื่อนำมาใช้ในการวิเคราะห์ข้อมูล

2.2.2 ขั้นตอนการเตรียมข้อมูล (Data Preparation) คือขั้นตอนในการเตรียมข้อมูลเพื่อที่จะทำให้ข้อมูลมีคุณภาพและเหมาะสมในการทำไมนิ่ง และเป็นขั้นตอนที่มักใช้เวลามากที่สุดเนื่องจากต้องมีการพิจารณาเกี่ยวกับข้อมูลและวัตถุประสงค์ในการทำ ชนิด ประเภท จำนวนและอายุของข้อมูล ซึ่งประกอบด้วยขั้นตอนย่อย 4 ขั้นตอนดังนี้

2.2.2.1 Data cleaning คือ ขั้นตอนในการเลือกข้อมูลที่ต้องการและเอาข้อมูลที่ไม่ต้องการออกจากแหล่งข้อมูล ซึ่งข้อมูลส่วนใหญ่นั้นจะไม่สมบูรณ์ (Incomplete), ค่าของข้อมูลผิดไปจากค่าที่ควรจะเป็น (noisy), ไม่ครบ (Missing Value) และไม่สอดคล้องกัน (Inconsistent) ซึ่งในการเลือกข้อมูลนั้นจำเป็นที่จะต้องเข้าใจความหมายทราบประเภทของข้อมูลและค่าที่สามารถเป็นไปได้ ซึ่งตัวแปรของข้อมูลจะแบ่งออกเป็น 2 ลักษณะดังนี้

- แบบ Categorical คือค่าของตัวแปรที่สามารถแบ่งเป็นกลุ่มหรือหมู่ และบอกถึงสิ่งนั้นๆ ได้อย่างชัดเจน ซึ่งแบ่งได้ออกเป็น 2 ประเภท คือ
  - Nominal: คือตัวแปรที่ลำดับของข้อมูลไม่มีความสำคัญเช่น เพศ (ชาย, หญิง)
  - Ordinal: คือตัวแปรที่ลำดับของข้อมูลมีความสำคัญ เช่น ระดับความน่าเชื่อถือของลูกค้า (ดี, ปานกลาง, ไม่ดี)
- แบบ Quantitative เป็นตัวแปรที่บอกค่าของความแตกต่างในด้านขนาดหรือปริมาณ ซึ่งแบ่งได้เป็น 2 ประเภทดังนี้
  - Continuous: เป็นตัวแปรที่เก็บค่าตัวเลขเป็นจำนวนจริง เช่น ภาษี, ค่าใช้จ่ายของบริษัทในแต่ละเดือน
  - Discrete: เป็นตัวแปรที่เก็บค่าตัวเลขเป็นจำนวนเต็ม เช่น จำนวนพนักงานในบริษัท

ซึ่งในที่นี้จะกล่าวถึงในกรณีที่ค่าของข้อมูลขาดหายไป (Missing Value) จะมีทางเลือกในการแก้ไขดังนี้

1. ถ้าฟิลด์ใดหายก็ไม่สนใจฟิลด์นั้น โดยไม่นำมาประมวลผลด้วย ซึ่งวิธีการนี้ไม่ค่อยมีประสิทธิภาพ
2. ทำการหาข้อมูลเพื่อนำมาเติมข้อมูลที่ขาดหายไป ซึ่งวิธีการนี้อาจไม่เหมาะสมในกรณีที่แหล่งข้อมูลมีขนาดใหญ่
3. ใช้ค่าคงที่กลาง (Global constant) เติมลงไปในฟิลด์ที่ข้อมูลหาย เช่น เติมค่า Unknown แต่วิธีนี้อาจทำให้ผลของการคำนวณไม่ถูกต้องเท่าที่ควร
4. ใช้ค่าเฉลี่ยของฟิลด์นั้นเติมลงไปในข้อมูลที่ขาด เช่น ถ้าข้อมูลในฟิลด์เงินเดือนพนักงานของเรคคอร์ดหนึ่งหายไป ก็ใช้ค่าเฉลี่ยเงินเดือนพนักงานเติมแทน
5. ใช้ค่าเฉลี่ยของข้อมูลในคลาสเดียวกันเติมในฟิลด์ของเรคคอร์ดที่หายไป
6. ใช้ค่าความน่าจะเป็นที่คิดว่าน่าจะเกิดขึ้นมากที่สุด มาเติมในค่าที่ขาดหายไป ซึ่งอาจจะนำ Regression, Bayesian, decision tree มาใช้ซึ่งวิธีนี้เป็นวิธีที่นิยมใช้มากที่สุด

2.2.2.2 Data integration คือ การรวมข้อมูลจากหลายๆ แหล่งข้อมูลเข้าด้วยกัน สิ่งที่ต้องคำนึงถึงก็คือความถูกต้องตรงกันในเรื่องของสกีมา (Schema integration) เช่น ปัญหาในเรื่องการกำหนดเอนติตี้ (Entity Identification Problem) เช่น ในกรณี

A.cust-id = B.cust-# และความขัดแย้งในเรื่องของหน่วยวัดข้อมูล ซึ่งการรวมข้อมูลนั้นจะต้องมั่นใจได้ว่าข้อมูลนั้นไม่ซ้ำซ้อน (Redundancy)

2.2.2.3 Data transformation คือ ขั้นตอนที่ข้อมูลจะถูกเปลี่ยนรูปและรวมให้อยู่ในรูปแบบที่เหมาะสมแก่การทำดาต้าไมนิ่งซึ่งมีแนวคิดการทำงานดังนี้

- Smoothing คือการเอาข้อมูลที่ผิดไปจากค่าที่ควรจะเป็น (Noise) ออกจากข้อมูลซึ่งวิธีที่ใช้คือ Binning, การจัดกลุ่ม (Clustering), regression
- Aggregation คือวิธีในการรวมหรือสรุปข้อมูล เช่นข้อมูลการขายประจำวันเราอาจจะทำการสรุปข้อมูลให้เป็นรายปีหรือรายเดือน
- Generalization คือการแปลงข้อมูลดิบหรือข้อมูลที่อยู่ในระดับต่ำให้อยู่ในระดับที่สูงกว่าตามลำดับขั้น (Hierarchies) เช่น ฟิลด์ถนน เราอาจจะนำข้อมูลไปรวมอยู่ในฟิลด์ City และ ฟิลด์อายุ เราอาจจะแบ่งฟิลด์อายุออกเป็น อายุน้อย, วัยกลางคน, อายุมาก
- Normalization คือการกำหนดช่วงของค่าให้แคบลง เช่น กำหนดให้ข้อมูลอยู่ในช่วง –1.0 – 1.0 หรือ 0.0 – 1.0
- Attribute construction คือการสร้างแอตทริบิวต์ใหม่เพื่อทำให้ข้อมูลมีความถูกต้องและมีความเข้าใจมากยิ่งขึ้น เช่น เราอาจจะสร้างแอตทริบิวต์ Area ใหม่โดยได้มาจากแอตทริบิวต์ Width และ High นำมาคูณกัน เป็นต้น

2.2.2.4 Data Reduction คือการลดรูปของข้อมูลให้อยู่ในลักษณะที่มีการสรุปมากขึ้นและมีขนาดเล็กลงเพื่อนำมาใช้ในการวิเคราะห์ ซึ่งวิธีการที่ใช้ในการทำ Data reduction มีดังนี้

- Data Cube Aggregation คือวิธีการในการสรุปข้อมูลและมองโครงสร้างในลักษณะของรูปลูกบาศก์ เช่น ถ้าเรามีข้อมูลยอดขายสินค้ารายเดือนในแต่ละปี เราสามารถที่จะทำการสรุปข้อมูลให้อยู่ในรูปของยอดขายสินค้ารายปี ซึ่งการสรุปรวมข้อมูลแบบนี้ทำให้ง่ายในการวิเคราะห์และช่วยลดขนาดของข้อมูล
- Dimension reduction คือการนำข้อมูลที่ไม่มีความสัมพันธ์กับข้อมูลที่ต้องการวิเคราะห์ออกไป เช่น ถ้าเราต้องการวิเคราะห์ยอดขาย CD ของบริษัท เราอาจไม่จำเป็นต้องเอาข้อมูลเบอร์โทรศัพท์ของลูกค้ามาวิเคราะห์
- Data Compression จะนำกลไกในการ Encode มาใช้ลดขนาดของข้อมูล

- Numerosity reduction คือการใช้ข้อมูล Parametric มาเก็บแทนข้อมูลจริง
- Discretization and concept hierarchy generation คือค่าของข้อมูลจะถูกแทนที่ด้วยช่วงของค่าใน Conceptual Level ที่สูงกว่า

2.2.3 ขั้นตอนการไมนิ่งข้อมูล (Data Mining) เป็นขั้นตอนในการประมวลผลข้อมูลตามวิธีและอัลกอริทึมที่เรากำหนดไว้ ซึ่งการทำไมนิ่งนั้นมีแนวคิดหลักๆ 4 แนวคิดดังนี้

2.2.3.1 Predictive Modeling (การสร้างแบบจำลองพยากรณ์) คือการสร้างรูปแบบจำลองที่ใช้ในการวิเคราะห์ฐานข้อมูลเพื่อการตัดสินใจ ดังนั้นรูปแบบและข้อมูลจะต้องมีความถูกต้องเพื่อที่ทำให้ผลลัพธ์ที่ออกมามีความถูกต้อง ซึ่งการทำงานจะแบ่งเป็น 2 ส่วนดังนี้

- Training Phase เป็นขั้นตอนที่สร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีตโดยใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมด
- Testing Phase คือขั้นตอนการทดสอบแบบจำลองที่ได้สร้างขึ้นว่ามีประสิทธิภาพหรือไม่ โดยนำข้อมูลในส่วนที่เหลือ 20% จากช่วง Training Phase มาใช้ทดสอบแบบจำลอง

Predictive Modeling แบ่งเป็น 2 รูปแบบคือ

2.2.3.1.1. Classification เป็นการสร้างแบบจำลองเพื่อทำนายกลุ่มของข้อมูลที่เราสนใจ ซึ่งกลุ่มต่างๆ จะมีการกำหนดไว้ล่วงหน้า เช่น การทำนายการให้กู้ยืมเงินของธนาคารแก่ลูกหนี้ว่ามีความเสี่ยงหรือปลอดภัยซึ่งแนวคิดใน Classification มีหลายรูปแบบดังนี้ Decision Tree, Bayesian Classification, Bayesian Belief Network, Neural Network,k-nearest neighbor Classification Case-Base Reasoning, Genetic Algorithms, Rough Set, Fuzzy Logic

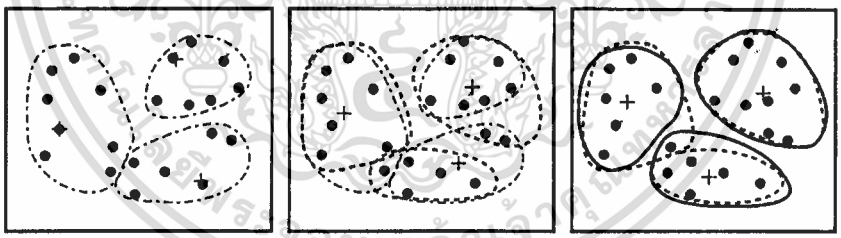
2.2.3.1.2. Value Prediction เป็นการประมาณการข้อมูลทางตัวเลขในฐานข้อมูล โดยใช้เทคนิคทางสถิติ หรือเรียกอีกอย่างหนึ่งว่าการทำ Scoring ซึ่งเป็นตัวเลขที่บอกถึงความเป็นไปได้ของข้อมูล เช่น บริษัทขายรถยนต์ต้องการประเมินข้อมูลการอนุมัติการกู้เงินของลูกค้า ซึ่งเราจะต้องนำตัวแปรต่างๆ มาประกอบการพิจารณา เช่น อายุ, ประวัติการใช้รถ, สถานะทางสังคม, วุฒิการศึกษา และ อาชีพ เป็นต้น

2.2.3.2 Database Segmentation หรือ Clustering นั้นเป็นวิธีในการแบ่งกลุ่มของข้อมูล โดยจัดข้อมูลที่มีความคล้ายคลึงกันอยู่ด้วยกัน ส่วนข้อมูลที่มีความแตกต่างกันก็อยู่คนละกลุ่ม ซึ่งเราสามารถแบ่งเทคนิคในการทำ Clustering ออกเป็น 4 ประเภทดังนี้

- Partitationing methods เช่น k-Means, k-modes, k-Medoids, CLARANS (Clustering LARge Applications) Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

ซึ่งในรายงานฉบับนี้จะกล่าวถึง K-Means และ Hierarchical methods เท่านั้น ซึ่งแนวคิดแบบ Partitioning methods มีหลักในการทำงานดังนี้คือ สมมุติว่าในฐานข้อมูลมีออบเจกต์อยู่ n ตัว และทำการแบ่งกลุ่มออกเป็น k ส่วนโดยที่ $k \leq n$ หลังจากนั้นก็ทำการแบ่งออบเจกต์ทีละตัวเข้ากับส่วน (Partition) ที่เรากำหนดไว้ โดยที่ข้อมูลที่คล้ายคลึงกันก็จะอยู่ในส่วน (Partition) เดียวกัน ส่วนข้อมูลที่ต่างกันก็จะอยู่คนละส่วน (Partition) ซึ่งวิธีการที่ใช้ในการแบ่งข้อมูลมีดังนี้

- **K-means**



รูปที่ 2.1 K-means

จากรูปที่ 2.1 จะอธิบายการทำงานของ k-means ได้ดังนี้

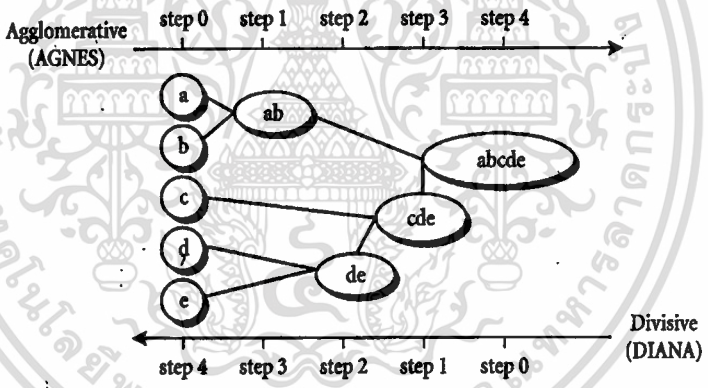
จาก (a) กำหนดจำนวนกลุ่ม (Cluster = k) ที่ต้องการซึ่งจากรูปกำหนดให้ k = 3 และทำการเลือกข้อมูลมา k ตัว (ซึ่งจากตัวอย่างเลือกข้อมูลมา 3 ตัวเพื่อมาใช้เริ่มเป็นจุดศูนย์กลางของกลุ่ม) ซึ่งข้อมูลที่เป็นจุดศูนย์กลางจะมีเครื่องหมาย + กำกับอยู่ต่อจากนั้นจะนำเอาข้อมูลแต่ละตัวมาจัดอยู่ในกลุ่ม โดยมีหลักว่าระยะห่างจากตัวข้อมูลกับศูนย์

กลางของแต่ละกลุ่ม ใดมีค่าน้อยที่สุดก็ให้สมาชิกเข้าไปอยู่ในกลุ่มนั้น ทำเช่นนี้กับข้อมูลจนครบ N ตัว

จาก (b) เมื่อทำการจัดข้อมูล N ตัวให้อยู่ในกลุ่มได้แล้วก็ให้ทำการคำนวณหาศูนย์กลางใหม่ โดยคำนวณจากค่าเฉลี่ยของออบเจกต์ (Object) ในแต่ละกลุ่ม และนำเอาข้อมูลแต่ละตัวมาเทียบกับศูนย์กลางใหม่ในแต่ละกลุ่ม ถ้าข้อมูลในกลุ่มใดมีค่าใกล้เคียงกับศูนย์กลางอื่นมากกว่าก็ทำการย้ายกลุ่ม

จาก (c) ให้ทำการคำนวณหาศูนย์กลางใหม่คือทำซ้ำขั้นตอน (b) ไปเรื่อยๆ จนไม่มีการเปลี่ยนแปลงข้อมูลของกลุ่มใดกลุ่มหนึ่งอีก

- **Hierarchical methods** เป็นการจัดกลุ่มของข้อมูลให้อยู่ในรูปของโครงสร้างต้นไม้ (Tree Diagram) ดังรูปที่ 2.2



รูปที่ 2.2 Hierarchical methods

Hierarchical methods จะแบ่งออกได้เป็น 2 ประเภท

1. Agglomerative hierarchical clustering ใช้หลักการ bottom-up เริ่มต้นโดยการแทนที่แต่ละออบเจกต์ (Object) ใน Cluster เดียวกันและทำการรวม Cluster กลุ่มเล็กเข้ากับ Cluster กลุ่มใหญ่ จนกระทั่งทุก Object รวมอยู่ใน 1 Cluster หรือจนกระทั่งสิ้นสุดเงื่อนไขที่ตั้งไว้

2. Divisive hierarchical clustering ใช้หลักการ top-down เป็นการมองย้อนกลับของ Agglomerative เริ่มต้นโดยจะทำการแบ่ง Cluster ใหญ่ให้กลายเป็น Cluster ย่อยจนกระทั่งแต่ละออบเจกต์ (Object) เป็นตัวของมันเองหรือสิ้นสุดเงื่อนไขที่พอใจ

2.2.3.3 Link Analysis คือ แนวคิดในการค้นหาความสัมพันธ์ระหว่างเรคคอร์ดแต่ละตัวหรือกลุ่มของเรคคอร์ดในฐานข้อมูล ซึ่งความสัมพันธ์นั้นจะเรียกว่า Association วิธี Link Analysis เหมาะในการวิเคราะห์ความสัมพันธ์ระหว่างสินค้าหรือบริการที่ลูกค้ามีแนวโน้มว่าจะใช้ร่วมกัน เช่น Cross Selling, Target Marketing และ Stock Price Movement ซึ่ง Link Analysis นั้นมี 3 วิธีการดังนี้ Association discovery, Sequential pattern discovery และ Similar time sequence discovery

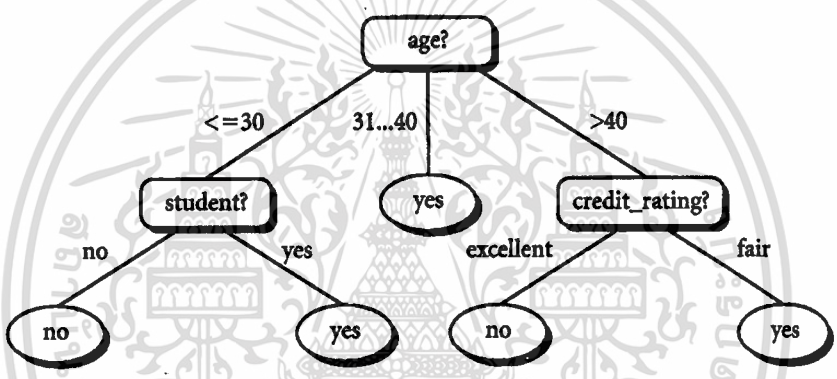
2.2.3.4 Deviation Detection เป็นแนวคิดใหม่ซึ่งมีความสำคัญ Deviation detection ที่ใช้กันมีอยู่ 2 วิธีคือ Statistics กับ Visualization techniques ที่จะใช้ตรวจหาความผิดปกติจากข้อมูลปกติหรือ ข้อมูลที่คาดว่าจะเป็น ซึ่งในปัจจุบันเทคนิคการทำ Visualization สามารถทำได้ง่าย ซึ่งจะแสดงผลการวิเคราะห์เป็นกราฟิกทำให้สามารถตรวจจับความผิดปกติได้ง่าย ตัวอย่างของ Deviation detection ได้แก่ การตรวจสอบความผิดปกติของการใช้บัตรเครดิต, การอ้างสิทธิ์การประกัน, การตรวจสอบคุณภาพ (Quality Control)

2.2.4 ขั้นตอนการประเมินรูปแบบดาต้าไมนิ่ง (Pattern evaluation) คือ ขั้นตอนในการประเมินรูปแบบดาต้าไมนิ่งที่เราได้ทำลงไปซึ่งการวัดนั้นส่วนใหญ่จะใช้วิธีการทางสถิติ ซึ่งผลที่ได้จากการประมวลผลนั้นต้องมีความน่าเชื่อถือ เข้าใจง่าย น่าสนใจและมีประโยชน์

2.2.5. ขั้นตอนการแสดงผล (Knowledge presentation) คือ ขั้นตอนในการนำเสนอองค์ความรู้ที่ได้ให้กับผู้ใช้ ซึ่งควรมีการนำ Graphical User Interface (GUI) เข้ามาใช้เพื่อให้ผู้ใช้เข้าใจข้อมูลที่เราแสดงได้ง่าย ควรให้ผู้ใช้การค้นหาข้อมูลเอง (มีระบบการค้นข้อมูล) มีระบบค้นหา (Search)

บทที่ 3
รายละเอียดและขั้นตอนการทำดิซิชั่นทรี

ดิซิชั่นทรีเป็นรูปแบบการตัดสินใจในแบบมองจากบนลงล่าง (Top-Down) ซึ่งโครงสร้างของ ดิซิชั่นทรีจะประกอบไปด้วยรูตโหนด (Root Node) ซึ่งจะเป็นโหนดบนสุด และแตกออกไปเป็นโหนดลูก (Child Node) ซึ่งแต่ละโหนดอาจมีลูกมากกว่า 2 โหนดก็ได้ ส่วนโหนดที่อยู่ระดับล่างสุด เราเรียกว่า ลีฟโหนด (Leaf Node) ดังรูปที่ 3.1



รูปที่ 3.1 ดิซิชั่นทรี

จากรูปที่ 3.1 จะแสดงข้อมูลลูกค้าที่ซื้อคอมพิวเตอร์โดยที่อินเทอร์นอลโหนด (Internal Node) แสดงด้วยรูปที่สี่เหลี่ยมผืนผ้าปลายมน จะแสดงเงื่อนไขในการทดสอบแอตทริบิวต์ (Attribute) และกิ่ง (Branch) ของแบบจำลองต้นไม้ (Tree Model) จะแสดงผลของการทดสอบ และลีฟโหนด (แสดงด้วยรูปวงรี) จะแสดงคลาส (Class) หรือกลุ่มของออบเจกต์ (Object) ซึ่ง ดิซิชั่นทรีมีขั้นตอนในการดำเนินงานดังนี้

1. ทำการเลือกแอตทริบิวต์ (Test Attribute ซึ่งจากรูปที่ 3.1 คือ Age) เพื่อใช้ทดสอบกลุ่มข้อมูลตัวอย่าง เพื่อมากำหนดเป็นรูตโหนดโดยมีการกำหนดผลของการทำไว้ล่วงหน้าเป็นลีฟโหนด (ในที่นี้คือกรณีที่ลูกค้ามีอายุ 31-40 ปี ให้จัดอยู่ในคลาส yes)
2. นำกลุ่มข้อมูลมาแบ่งข้อมูลตาม Test Attribute ที่เราได้กำหนดไว้
3. ถ้ากลุ่มข้อมูลใดอยู่ในกลุ่มเดียวกับ Test Attribute ให้ทำการคืนค่าโหนด (Leaf Node)

4. ถ้าข้อมูลไม่ได้อยู่ในกลุ่มเดียวกับ Test Attribute ข้างต้นให้ทำการเลือก Test Attribute ตัวอื่นๆ ในกลุ่มของแอตทริบิวต์ลิสส์ (เช่น Student, Credit_rating) เพื่อทำการทดสอบต่อไป
5. ย้อนกลับไปทำข้อ 2 ใหม่ จนครบทุกแอตทริบิวต์ในแอตทริบิวต์ลิสส์

3.1 อัลกอริทึม SLIQ

ในอดีตที่ผ่านมาปัญหาที่สำคัญอันหนึ่งของการทำดาต้าไมนิ่งโดยใช้อัลกอริทึมที่เกี่ยวกับการทำ Classification คือ อัลกอริทึมส่วนใหญ่จะถูกออกแบบให้ใช้พื้นที่หน่วยความจำขนาดใหญ่ในการทำงาน ซึ่งทำให้จำนวนชุดของข้อมูลที่ใช้ในการวิเคราะห์จะถูกจำกัดโดยจำนวนหน่วยความจำ ซึ่งในรายงานฉบับนี้จะศึกษาเกี่ยวกับอัลกอริทึมที่เรียกว่า SLIQ (Supervised Learning in Quest) ซึ่งเป็นอัลกอริทึมในการทำ Classification โดยใช้แนวความคิดดิซิชั่นทรี SLIQ นั้นเหมาะกับข้อมูลขนาดใหญ่ SLIQ สามารถจัดการกับข้อมูลได้ทั้งแบบตัวเลข (Numeric attributes) และข้อมูลที่จัดเป็นหมวดหมู่(Categorical attributes) โดยใช้เทคนิคใหม่ ในการ Pre-sorting data คือการเรียงลำดับข้อมูลในช่วงของการสร้างแบบจำลอง (Tree-growth phase) ที่เรียกว่า Breadth-first growing เพื่อที่จะย้ายข้อมูลบางส่วนลงไปเก็บในดิสก์แทนและในส่วนของการปรับแต่งแบบจำลองต้นไม้ (Tree-pruning) ก็นำเทคนิค MDL (Minimum Description Length Principle) มาใช้ซึ่งจะมีผลทำให้เพิ่มความถูกต้องของแบบจำลองมากยิ่งขึ้น ซึ่งทั้งหมดนี้จะทำให้ SLIQ สามารถจัดการกับการแบ่งกลุ่มข้อมูลได้โดยไม่ต้องคำนึงถึง ขนาด หรือชนิดของข้อมูลอีกต่อไป ซึ่งจะทำให้เราสามารถสร้างระบบในการทำดาต้าไมนิ่งได้ดีขึ้น

3.2 การทำงานของ SLIQ

ในการทำงานของดิซิชั่นทรีนั้นจะแบ่งเป็น 2 ขั้นตอนคือ

3.2.1 การสร้างแบบจำลองต้นไม้ (Tree Building)

3.2.2 การปรับแต่งแบบจำลองต้นไม้ (Tree Pruning)

ซึ่งในแต่ละขั้นตอนนั้นสามารถอธิบายวิธีการทำงานได้ดังต่อไปนี้

3.2.1 การสร้างแบบจำลองต้นไม้ (Tree Building)

การสร้างแบบจำลองต้นไม้นั้นเป็นขั้นตอนแรกในการทำ ดิซิชั่นทรีโดยทำการแบ่งกลุ่ม ข้อมูลที่ใช้ในการสร้างแบบจำลองต้นไม้ (Training Data) ซึ่งชุดข้อมูลจะถูกแบ่งเป็นสองส่วนหรือมากกว่า 2 ส่วนก็ได้ตามแอตทริบิวต์ที่กำหนดไว้ การทำงานในขั้นตอนนี้จะเป็นการทำงานแบบรี

เดอร์ซีฟ (Recursive) คือทำซ้ำเป็นรอบๆ จนกระทั่งแต่ละตัวอย่างนั้นอยู่ในคลาสเดียว ตามขั้นตอนการทำงานดังนี้

```
Make Tree(Trainging Data T)
    Partition(T)
    Partition(Data S)
    If (All points in S are in the same class) then return;
    Evaluate splits for each attribute A
    Use best split found to partition S into S1 and S2;
    Partition(S1);
    Partition(S2);
```

ในการสร้างแบบจำลองต้นไม้นั้นมี 2 ขั้นตอนที่สำคัญคือ

1. ทำการคัดเลือกแอตทริบิวต์ที่เหมาะสมในการวิเคราะห์ข้อมูล และทำการแบ่งกลุ่มแอตทริบิวต์ออก เพื่อหาจุดที่ดีที่สุด (Best Split) ในการเลือกและแบ่งแอตทริบิวต์ ซึ่งเราจะใช้สูตรคำนวณในการหาคือ gini ซึ่งมีสูตรดังนี้

$$Gini(T) = 1 - \sum p^2_j$$

T: ชุดของข้อมูลที่เก็บตัวอย่างจาก n คลาส

p_j : ความถี่สัมพันธ์ของคลาส j ใน T

ถ้าชุดของข้อมูล T ได้แบ่งเป็น 2 สับเซต ด้วยขนาด N1 และ N2 คลาส ค่า $Gini_{split}$ จากจำนวน N คลาสจะเป็นดังนี้

$$Gini_{split}(T) = N_1/N\ gini(T_1) + N_2/N\ gini(T_2)$$

2. สร้างจุดแบ่งโดยใช้ค่า $Gini_{split}$ ที่ต่ำที่สุดมาเป็นตัวกำหนดจุดแบ่ง ซึ่งในการกำหนดจุดแบ่งนั้นจะขึ้นอยู่กับว่าแอตทริบิวต์นั้นเป็นแอตทริบิวต์ประเภทไหนดังนี้

- การแบ่งข้อมูลสำหรับแอตทริบิวต์ที่เป็นตัวเลข (Numeric) เป็นการแบ่งข้อมูลแบบเป็น 2 ทาง (Binary Split) จากรูปแบบ $A \leq v$ โดยที่ v เป็นตัวเลขจำนวนจริงของแอตทริบิวต์ A ซึ่ง

ให้ทำการเรียงข้อมูลที่นำมาใช้ทดสอบซึ่งขึ้นอยู่กับค่าของแอตทริบิวต์ A ที่พิจารณา เช่นทำการเรียงค่า *v1, v2,...,vn* และเมื่อหาค่า Best Split ได้ว่าเป็นเรคคอร์ดที่ v_i ก็ให้นำค่ากลางระหว่าง $vi - v_{i+1}$ มาเป็น Best Split

- การแบ่งข้อมูลสำหรับแอตทริบิวต์ที่แบ่งเป็นหมวดหมู่ (Categorical) ถ้าให้ *S(A)* เป็นชุดของค่าที่เป็นไปได้ในแอตทริบิวต์ A ดังนั้นการแบ่งข้อมูลจะอยู่ในรูปของ $A \in S'$, ซึ่ง $S' \subset S$ ดังนั้นจำนวนของซับเซตที่เป็นไปได้ n จำนวนจะเท่ามีค่าเท่ากับ 2^n สำหรับแอตทริบิวต์ที่มีค่าเป็นไปได้ n ค่า โดยจะต้องมีการกำหนดค่า MAXSETSIZE คือจำนวน n สูงสุดที่ทำให้การหาค่า Best Split มีประสิทธิภาพ โดยทั่วไปจะกำหนด MAXSETSIZE = 10

3.2.1.1 ขั้นตอนการสร้างแบบจำลองต้นไม้ (Tree Model) ใน SLIQ

แบ่งออกเป็น 2 ขั้นตอนดังนี้

1. Pre-Sorting และ Breadth-First Growth

เป็นขั้นตอนที่นำข้อมูลมาเรียงลำดับ ซึ่งข้อมูลที่เป็นตัวเลขนั้นการเรียงลำดับข้อมูลถือเป็นสิ่งที่จำเป็นที่สุด ดังนั้นเทคนิคแรกของการทำ ดิซิชั่นทรีโดยใช้อัลกอริทึม SLIQ คือต้องนำข้อมูลที่เป็น Numeric มาเรียงลำดับ (การเรียงลำดับนั้นให้เรียงเพียงครั้งแรกครั้งเดียว)

ในการทำงานของ Pre-Sorting นั้นเราต้องทำการจัดโครงสร้างข้อมูลให้เป็นดังนี้

- ทำการกำหนดคลาสลิสต์ (Class List) ขึ้นมาก่อนซึ่งในคลาสลิสต์นั้นจะประกอบไปด้วย 2 ฟิลด์ ดังนี้คือคลาสเลเบิล (Class Label) และลีฟโหนด ซึ่งคลาสเลเบิลคือค่าของแอตทริบิวต์ที่ใช้ในการทำนาย เช่น Y = Churn, N = Non Churn
- สร้างแอตทริบิวต์ลิสส์ (Attribute List) ซึ่งในแอตทริบิวต์ลิสส์นั้นจะประกอบไปด้วยค่าแอตทริบิวต์และค่าดัชนี (Index) และลำดับที่ I ในคลาสลิสต์จะมีผลต่อลำดับที่ I ของข้อมูลในแต่ละลีฟโหนดของ ดิซิชั่นทรีจะแสดงผลของการแบ่งข้อมูลใน Training Set (T) ในการสร้างแบบจำลองต้นไม้ครั้งแรกจะกำหนดให้ลีฟโหนดในคลาสลิสต์ชี้ไปที่รูตโหนด

2. ขั้นตอนการแบ่งโหนด (Node Splits)

ขั้นตอนการแบ่งโหนดมีอัลกอริทึมในการทำดังนี้

EvaluateSplits()

For each attribute list of *A* DO

Traverse attributed list of *A*

For each value *v* in the attribute list **DO**

Find the corresponding entry in the class list, and hence the corresponding class and the leaf node (say *l*)

Update the class histogram in the leaf *l*

IF *A* is a numeric attribute **then**

Compute splitting index for test ($A \leq v$) for leaf *l*

IF *A* is a categorical attribute **then**

For each leaf of the tree **do**

Find subset of A with best split

ในการแบ่งโหนดนั้นจะนำ Gini เทคนิคเข้ามาช่วย ซึ่งในการคำนวณสำหรับแต่ละแอตทริบิวต์นั้นต้องนำการกระจายความถี่ของค่าในแต่ละคลาสมาคิดด้วย ซึ่งก็คือการทำฮิสโทแกรม (Histogram) ถ้าแอตทริบิวต์เป็นค่า Numeric ตัวฮิสโทแกรมจะมีรูปแบบดังนี้ <class, frequency> ถ้าแอตทริบิวต์เป็น Categorical ตัวฮิสโทแกรมจะมีรูปแบบดังนี้ <attribute value, class, frequency>

3. เมื่อทำการหา Best Split ได้แล้วก็ให้ทำการสร้าง Tree และทำการปรับค่า Class List

3.2.1.2. ตัวอย่างการทำดิซิชั่นทรีโดยใช้ SLIQ

ข้อมูลที่นำมาใช้คือข้อมูลของลูกค้าที่ปิดบริการในระบบแบบใช้ก่อน จ่ายทีหลัง (Post-Paid) ระหว่างวันที่ 01/07/2002 – 31/12/2002 ข้อมูลตัวอย่างที่นำมาวิเคราะห์ประกอบด้วยแอตทริบิวต์ดังนี้

- Subr_Age คือ อายุการใช้งานโทรศัพท์มือถือมีหน่วยเป็นเดือน
- Invc_levl คือ ระดับค่าบริการรายเดือนเฉลี่ยย้อนหลัง 3 เดือนก่อนปิดบริการ

ตารางที่ 3.1 ตารางแสดงระดับอัตราค่าใช้บริการ

อัตราค่าใช้บริการ	ระดับ
<1,000	L
1,000-2,000	M
2,000 บาทขึ้นไป	H

ตารางที่ 3.2 ตารางแสดงข้อมูลตัวอย่าง

index	subr_age	Invc_levl	class
1	62	H	N
2	11	H	Y
3	13	M	N
4	10	M	N
5	7	L	N
6	32	L	N
7	26	M	N
8	9	H	Y
9	17	M	N
10	15	H	N

ขั้นตอนการสร้างแบบจำลองต้นไม้มีดังนี้

1. จากตัวอย่างให้นำข้อมูล Numeric แอตทริบิวต์มาเรียงลำดับ(จะทำการเรียงข้อมูลในตอนแรกเพียงครั้งเดียว) ซึ่งจะได้ข้อมูลดังนี้

   - แอตทริบิวต์ที่เป็น Numeric ซึ่งในที่นี้คือ subr_age

ตารางที่ 3.3 ตารางแสดง subr_age ที่เรียงลำดับแล้ว

subr_age	index
7	5
9	8
10	4
11	2
13	3
15	10
17	9
26	7
32	6
62	1

- แอตทริบิวต์ที่เป็น Category ไม่ต้องทำการเรียงลำดับ แต่ให้ใส่ Class List Index เหมือน Numeric

ตารางที่ 3.4 ตารางแสดงค่า Invc_levl และ Class List

Invc_levl	index		class	Leaf
L	5	1	N	N1
H	8	2	Y	N1
M	4	3	N	N1
H	2	4	N	N1
M	3	5	N	N1
H	10	6	N	N1
M	9	7	N	N1
M	7	8	Y	N1
L	6	9	N	N1
H	1	10	N	N1

2. เมื่อทำการเรียงข้อมูลแล้วให้ทำการสร้างตารางฮิสโทแกรมแล้วคำนวณ Gini Index ของแต่ละเรคคอร์ด ซึ่งการสร้างตารางฮิสโทแกรมนั้นให้นำค่าคลาสมากำหนดเป็นคอลัมน์แล้วนับค่าที่อยู่เหนือและอยู่ล่างเรคคอร์ดนั้นตามค่าคลาสที่กำหนดไว้ ในตัวอย่างนี้จะใช้ Subr_age แอตทริบิวต์มากำหนดเป็น รูต ซึ่งเป็น Numeric แอตทริบิวต์ ดังนั้นจึงมีขั้นตอนการทำตารางฮิสโทแกรมดังนี้

เรคคอร์ดที่ 1. จะมีค่าฮิสโทแกรมดังนี้

ตารางที่ 3.5 ฮิสโทแกรมของ Numeric แอตทริบิวต์

	Y	N	sum
C above	0	1	1
C below	2	7	9

ซึ่งนำค่าในตารางฮิสโทแกรมมาคำนวณค่าในสูตร Gini Index และ Gini Split ดังนี้

$Gini(S1) = 1-[(0/1)^2+(1/1)^2] = 0$

$Gini(S2) = 1-[(2/9)^2+(7/9)^2] = 0.345$

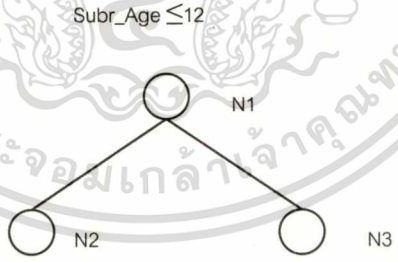
$Gini_{split} = 1/10(0) +9/10(0.345) = 0.311$

ทำการคำนวณดังนี้จนครบทุกเรคคอร์ด จะได้ค่า Ginisplit ดังนี้

ตารางที่ 3.6 Gini Split ของ subr_age

index	subr_age	class	gini split
5	7	N	0.311111
8	9	Y	0.275
4	10	N	0.304762
2	11	Y	0.2
3	13	N	0.24
10	15	N	0.266667
9	17	N	0.285714
7	26	N	0.3
6	32	N	0.311111
1	62	N	0.32

จากตารางที่ 3.6 จะพบว่า Index ที่ 2 มีค่าน้อยสุดเท่ากับ 0.2 เพราะฉะนั้นเราจะใช้ Index 2 เป็นจุดแบ่ง (Split Point) ดังนั้นเราจะใช้ค่ากลางระหว่าง 11 และ 13 ซึ่งก็คือ (11+13)/2 = 12 เป็นค่า Best Split ซึ่งสามารถนำมาสร้างเป็นแบบจำลองต้นไม้ได้ดังรูปที่ 3.2



รูปที่ 3.2 การแบ่ง Best Split ที่รูตโหนด

3. ทำการปรับค่าคลาสลิสต์

ในการปรับค่าคลาสลิสต์นั้นมีอัลกอริทึมในการทำดังนี้

UpdateLabels()

For each attribute *A* used in a split **do**

Traverse attribute list of *A*

For each value *v* in the attribute list **do**

Find the corresponding entry in the class list (say *e*)

Find the new class *c* to which *v* belongs by applying

The splitting test at node referenced from *e*

Update the class label for *e* to *c*

Update node referenced in *e* to the child corresponding to the class c

ดังนั้นเราสามารถแบ่งข้อมูลไปไว้ในโหนด N2 และ N3 ได้ดังนี้

ตารางที่ 3.7 เรคคอร์ดในโหนด N2

Index.	Subr_Age	Class	Leaf
5	7	N	N2
8	9	Y	N2
4	10	N	N2
2	11	Y	N2

ตารางที่ 3.8 เรคคอร์ดในโหนด N3

Index	Subr_Age	Class	Leaf
3	13	N	N3
10	15	N	N3
9	17	N	N3
7	26	N	N3
6	32	N	N3
1	62	N	N3

พิจารณาจากตารางที่ 3.8 จะพบว่าค่าที่อยู่ในลีฟโหนดN3 มีค่าเป็น N เพียงค่าเดียวดังนั้นจึงไม่ต้องทำการแบ่งต่อแล้ว ส่วนในตารางที่ 3.7 Node N2 ยังมีค่าทั้ง Y และ N อยู่จึงต้องทำการแบ่งค่าต่อโดยใช้แอตทริบิวต์ที่เหลือซึ่งก็คือ Invc_levl ให้นำมาหาค่า Gini Index แบบ Categorical แอตทริบิวต์

ตารางที่ 3.9 Invc_levl ในโหนด N2

Index	Invc_levl	class	Leaf
5	7	L	N
8	9	H	Y
4	10	M	N
2	11	H	Y

จากตารางให้ทำการสร้างตารางฮิสโทแกรมและหาค่า Gini index ดังนี้

ตารางที่ 3.10 ฮิสโทแกรมของ Category แอตทริบิวต์

	Y	N
L	0	1
M	0	1
H	2	0

จากตัวอย่างจะทำการหา Gini Split ของ H

$gini(S1) = 1-(2/2)^2 + (0/2)^2 = 0$

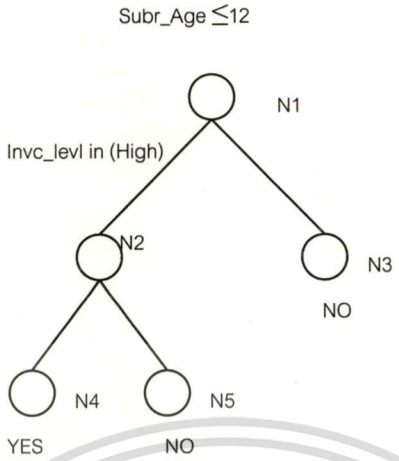
$gini(S2) = 1-(0/2)^2 + (2/2)^2 = 0$

$Gini_{split(H)} = (2/4)gini(S1) + (2/4)gini(S2) = 0$

ดังนั้น $Gini_{split(L)} = 0.208333$

$Gini_{split(M)} = 0.333333$

จากการคำนวณจะพบว่า $Gini_{split(H)}$ มีค่าน้อยสุดดังนั้นจึงนำมาใช้เป็นจุดแบ่ง ซึ่งสามารถสร้างแบบจำลอง (Tree Model) โดยใช้ Invc_levl = High ดังรูปที่ 3.3



รูปที่ 3.3 ผลดิซิชั่นทรี

ค่าของลีฟโหนดบนคลาสลิสต์จะเปลี่ยนแปลงค่าต่างๆ ตามการแบ่งกลุ่มที่เปลี่ยนแปลงไป จากตารางข้างล่าง จะเห็นว่าทุกลีฟโหนด มีค่าคลาสลิสต์อยู่ในกลุ่มเพียงหนึ่งกลุ่มเท่านั้น ดังนั้นจึงไม่ต้องทำการแตกโหนดอีกต่อไป

ตารางที่ 3.11 ลีฟโหนด

Index.	Invc_levl	Class	Leaf
8	H	Y	N4
2	H	Y	N4

Index.	Invc_levl	Class	Leaf
5	L	N	N5
4	M	N	N5

3.2.2. การปรับแต่งแบบจำลองต้นไม้ (Tree Pruning)

เป็นขั้นตอนในการปรับแบบจำลองต้นไม้เพื่อทำให้แบบจำลองนั้นมีขนาดเล็กลง และช่วยลดข้อมูลที่ไม่เกี่ยวข้องกับการวิเคราะห์ออกไป (Noise Data) ซึ่งในขั้นตอนนี้จะต้องทำการตรวจสอบ

แบบจำลองที่ได้สร้างไป และเลือกแบบจำลองต้นไม่ย่อย (Sub tree) ที่มีข้อผิดพลาดน้อยที่สุด ซึ่งมีวิธีหาค่าผิดพลาด อยู่ 2 วิธีคือ

- ใช้ชุดของ Training data เดิมวิธีนี้เรียกว่า Cross-Validation โดยนำ Training data มาแบ่งออกเป็น หลายๆตัวอย่างแล้วสร้างแบบจำลองต้นไม้จากตัวอย่างเหล่านี้ จากนั้นก็ใช้แบบจำลองนี้ในการประเมินหาข้อผิดพลาด (Error) กับ แบบจำลองที่เราได้สร้างขึ้นซึ่งถึงแม้ว่าวิธีการนี้จะทำให้ได้แบบจำลองที่ขนาดเล็กลงและมีความถูกต้องสูง แต่ว่าต้นทุนในการสร้างแบบจำลองแต่ละแบบจำลองนั้นค่อนข้างสูง ซึ่งทำให้วิธีนี้ไม่ค่อยเหมาะกับชุดของ Training data ขนาดใหญ่
- ใช้ข้อมูลที่เป็น Training data ใหม่ วิธีการนี้จะแบ่ง Training data ออกเป็น 2 ส่วน ส่วนหนึ่งจะใช้ในการสร้างแบบจำลองต้นไม้ (Tree Model) อีกส่วนหนึ่งจะใช้ในการปรับแต่ง (pruning) ซึ่งข้อมูลที่ใช้ในการปรับแต่งก็ควรจะเลือกข้อมูลที่มีการกระจายตามความเป็นจริง เพราะถ้าเลือกข้อมูลขนาดไม่เหมาะสมหรือเลือกข้อมูลไม่ถูก จะมีผลทำให้ไปลดขนาดและความถูกต้องของ Training data ในการสร้างแบบจำลองต้นไม้ได้

เพื่อป้องกันการสร้างแบบจำลองต้นไม้ที่เกินพอดีจึงได้นำหลักการของ MDL มาใช้คือเป็นการปรับแต่งแบบจำลองต้นไม้ในระหว่างการสร้างแบบจำลอง และจะหาแบบจำลองต้นไม้ย่อย (Sub Tree) ที่มีการใช้จำนวนบิต (Bit) ในการ Encode น้อยที่สุด ซึ่งต้นทุนของการ Encode มีสูตรดังนี้

$$\log\left(\frac{n+k-1}{k-1}\right) + \log\frac{n!}{n_1!\cdots n_k!}$$

โดยที่ให้เซต S ประกอบไปด้วย n เรคคอร์ดซึ่งแต่ละเรคคอร์ดอยู่ในคลาส k โดยที่ n_i เท่ากับจำนวนเรคคอร์ดในคลาส i

จากสูตรข้างต้นเทอมแรกของสูตรคือจำนวนของบิตในการระบุการแบ่งคลาสซึ่งก็คือจำนวนของของเรคคอร์ดในคลาส 1,..., k เทอมที่สองคือจำนวนของบิตที่ต้องการ Encode ในคลาสสำหรับแต่ละเรคคอร์ดซึ่งก็คือ n_i เรคคอร์ดใน คลาสเลเบิล i ซึ่งสูตรนี้จะไม่ถูกต้องถ้าบางค่าของ n_i มีค่าเข้าใกล้ศูนย์หรือ n ดังนั้นจึงมีอีกสูตรในการคำนวณคือ

$$C(S)=\sum_i n_i \log\frac{n}{n_i}+\frac{k-1}{2}\log\frac{n}{2}+\log\frac{\pi^{k/2}}{\tau(k/2)}$$

ในเทอมแรกคือ $n * E(S)$ ซึ่ง $E(S)$ คือ Entropy ของเซต S

ต้นทุนในการ Encode แบบจำลองต้นไม้ประกอบด้วย 3 ต้นทุนดังนี้

1. ต้นทุนในการ Encode โครงสร้างของแบบจำลองต้นไม้
2. ต้นทุนในการ Encode แต่ละจุดแบ่ง โดยพิจารณาจากประเภทของแอตทริบิวต์และค่าของจุดแบ่ง
3. ต้นทุนในการ Encode คลาสของข้อมูลในแต่ละ Leaf โหนดของแบบจำลองต้นไม้

- อัลกอริทึมในการปรับแต่งแบบจำลองต้นไม้ (Pruning Algorithm)

เราสามารถหาค่าต้นทุนที่ต่ำที่สุดของแบบจำลองต้นไม้ย่อย (Sub tree) ตามอัลกอริทึมดังต่อไปนี้

```
Procedure computeCost&Prune(Node N):
  /* S is the set of data records for N*/
1. If N is a leaf return (C(S) + 1)
   /* N1 and N2 are N's children */
2. minCost1 := computeCost&Prune(N1)
3. minCost2 := computeCost&Prune(N2)
4. minCostN := min{C(S) + 1,
       Csplit(N) + 1 + minCost1 + minCost2};
5. if minCostN = C(S) + 1
6.     prune child nodrs N1 and N2 from tree
7. return minCostN
```

บทที่ 4
วิธีดำเนินการศึกษา

ในการศึกษาโครงการนี้ จะเป็นการนำทฤษฎีดาต้าไมนิ่งแนวคิดดิชชั่นทรีโดยใช้อัลกอริทึมที่ชื่อว่า SLIQ มาพัฒนาเป็นโปรแกรมประยุกต์ เพื่อใช้ในวิเคราะห์หาสาเหตุการเปลี่ยนผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้า โดยแบ่งดำเนินการศึกษาเป็น 6 ขั้นตอน คือ

4.1 การศึกษาทฤษฎีที่เกี่ยวข้อง

ในการพัฒนาระบบงานนี้มีทฤษฎีที่ผู้พัฒนาต้องทราบดังนี้

- แนวคิด Customer Relationship Management (CRM) เบื้องต้นโดยเฉพาะในส่วนของ Churn Management เพื่อใช้ในการหาข้อมูล
- ทฤษฎีดาต้าไมนิ่ง โดยเฉพาะในแนวคิด Classification ในส่วนของดิชชั่นทรี
- อัลกอริทึมที่ใช้ในการสร้างแบบจำลองต้นไม้ (Tree Model) ซึ่งในการพัฒนาระบบนี้จะใช้ อัลกอริทึมที่ชื่อว่า SLIQ
- ความรู้ในด้านการพัฒนาโปรแกรมซึ่งในการพัฒนาโครงงานนี้ได้ใช้ VB.6 และ MS SQL Server 2000

4.2 การรวบรวมข้อมูลที่เกี่ยวข้อง

ในการรวบรวมข้อมูลที่จะนำมาใช้วิเคราะห์นั้น ทางผู้จัดทำได้ทำหนังสือขอข้อมูลไปถึงแผนกการตลาดขององค์กรที่ให้บริการด้านโทรศัพท์เคลื่อนที่แห่งหนึ่ง ซึ่งข้อมูลที่ได้เป็นข้อมูลของลูกค้าในระบบแบบใช้ก่อนจ่ายทีหลัง (Post-Paid) ที่ปิดบริการตั้งแต่วันที่ 01/07/2002 – 31/12/2002 จำนวน 1,500 เรคคอร์ด ซึ่งข้อมูลที่ได้นั้นอาจจะเก่าไปเมื่อเทียบกับข้อมูลปัจจุบัน ซึ่งรายละเอียดของข้อมูลที่ขอนั้นจะกล่าวไว้ในส่วนของฐานข้อมูล

4.3 การศึกษาความต้องการของระบบ

จากการศึกษาความต้องการของระบบนั้น พบว่าระบบนั้นต้องมีหน้าที่หลักในการทำงาน 6 หน้าที่ดังต่อไปนี้

1. ระบบต้องสามารถนำข้อมูลเข้าได้ 2 ทาง ดังนี้
   - เท็กซ์ไฟล์ (Text File) ซึ่งเท็กซ์ไฟล์นั้นจะต้องมีรูปแบบตามที่กำหนด ดังนี้ ชื่อตัวแปร (Attribute) จะอยู่ในแถวแรกและข้อมูลแต่ละตัวจะขั้นด้วยไปป์ (|)
   - ข้อมูลจากฐานข้อมูล ซึ่งข้อมูลจะต้องมาจากฐานข้อมูลที่เป็น MS SQL Server 2000 เท่านั้น
2. ผู้ใช้ระบบสามารถนำตัวแปรเข้ามาวิเคราะห์ได้ไม่เกิน 10 ตัวแปร (รวมตัวแปร ที่เป็นคลาสเลเบิล) และสามารถกำหนดได้ว่าจะนำตัวแปรใดเข้ามาวิเคราะห์ ซึ่งต้องสามารถเพิ่มหรือลดจำนวนตัวแปรที่นำมาวิเคราะห์ได้ในภายหลัง
3. ระบบต้องสามารถแบ่งข้อมูลออกเป็น 2 ส่วนจากแหล่งข้อมูลเดียวกัน
   - ส่วนที่ 1 คือ ส่วนของข้อมูลที่นำมาใช้ในการสร้างแบบจำลองต้นไม้ (Training Data)
   - ส่วนที่ 2 คือ ส่วนของข้อมูลที่นำมาใช้ทดสอบแบบจำลองต้นไม้ว่าที่เราได้สร้างขึ้น (Testing Data)
4. ผู้ใช้ระบบสามารถกำหนดเงื่อนไขในการสร้างแบบจำลองต้นไม้ได้ดังนี้
   - ผู้ใช้สามารถกำหนดระดับของแบบจำลองต้นไม้ที่จะแตกได้
   - ผู้ใช้สามารถกำหนดจำนวนข้อมูลที่น้อยที่สุดในแต่ละโหนดที่นำมาวิเคราะห์ได้
5. ระบบสามารถคำนวณผลการวิเคราะห์ได้ถูกต้องน่าเชื่อถือ ตามแนวคิด Classification โดยนำหลักการดิซิชั่นทรีและอัลกอริทึมที่ชื่อว่า SLIQ (Supervised Learning in Quest) เข้ามาใช้
6. ระบบแสดงค่าความเชื่อมั่นเพื่อให้ผู้ใช้ระบบใช้ประกอบการตัดสินใจ โดยค่าความเชื่อมั่นมีวิธีในการคิดดังนี้

$$\text{ค่าความเชื่อมั่น} = \frac{\text{จำนวนข้อมูลที่อยู่ในคลาสนั้น}}{\text{จำนวนข้อมูลในโหนดทั้งหมด}} \times 100\%$$

7. ระบบแสดงผลลัพธ์ให้ผู้ใช้ระบบสามารถเข้าใจได้ง่าย ซึ่งในที่นี้จะแสดงในรูปแบบจำลองต้นไม้ (Tree Model)

ซึ่งผู้จัดทำได้นำหน้าที่หลักทั้ง 7 ข้อนั้นมาใช้ในการวิเคราะห์และออกแบบระบบซึ่งจะได้กล่าวในหัวข้อถัดไป

4.4 การวิเคราะห์และออกแบบระบบ

ในการวิเคราะห์และออกแบบระบบนั้นแบ่งเป็น 2 ส่วนดังนี้

- การวิเคราะห์และออกแบบส่วนของฐานข้อมูล
- การวิเคราะห์และออกแบบส่วนของหน้าจอ

4.4.1 การวิเคราะห์และออกแบบส่วนของฐานข้อมูล

ในส่วนของฐานข้อมูลหลักประกอบด้วยตารางดังต่อไปนี้

1. ตาราง : DM_DATA_DETL เป็นตารางที่เก็บรายละเอียดชุดของข้อมูลและค่าพารามิเตอร์ที่นำมาใช้ในการวิเคราะห์ ซึ่งประกอบด้วยฟิลด์ดังต่อไปนี้

ตารางที่ 4.1 Table: DM_DATA_DETL

Field	Type	Format	Req.	PK or FK	FK Reference Table
inpt_id	Int(4)	9999	Y	PK	
text_path	Varchar(80)	xxxxx	Y		
load_dttm	Datetime (4)	DD/MM/YYYY hh:mm:ss	Y		
id_data_load	Int(4)	9999	Y		
db_textfile	Char(1)	x	Y		
class_vlue1	Char(10)	xxxxx	Y		
class_vlue2	Char(10)	xxxxx	Y		
vald_tree	Char(1)	x	Y		
testing_set	Int(4)	9999	Y		
class_list_var_id	Int(4)	9999	Y		
max_levl	Int(4)	9999	Y		
min_data	Int(4)	9999	Y		
total_recd	Int(4)	9999	Y		
total_load	Int(4)	9999	Y		

คำอธิบายฟิลด์ของ Table: DM_DATA_DETL

ตารางที่ 4.2 คำอธิบายฟิลด์ของ Table: DM_DATA_DETL

Field	Content	Description
inpt_id	Input ID	รหัสชุดข้อมูล
text_path	Text Path	ที่เก็บชุดของข้อมูลที่นำมาวิเคราะห์ (Path)
load_dttm	Load Datetime	วันที่และเวลาที่ชุดของข้อมูลนั้นถูกนำเข้ามาวิเคราะห์
id_data_load	Data Load	เก็บ Inpt_ID เดิมของแบบจำลองต้นไม้ก่อนมีการแก้ไข เพื่อถ้ามีการแก้ไขจะได้ไม่ต้องมีการสร้างข้อมูลบางตารางใหม่
db_textfile	Database/Textfile	เก็บว่าข้อมูลมาจากฐานข้อมูลหรือเท็กซ์ไฟล์ ถ้ามาจากฐานข้อมูลจะมีค่าเป็น "Y" ถ้ามาจากเท็กซ์ไฟล์จะมีค่าเป็น "N"
class_vlue1	Class Value No.1	ค่าคลาสเลเบิลที่ 1
class_vlue2	Class Value No.2	ค่าคลาสเลเบิลที่ 2
vald_tree	Validation Tree	เก็บว่า Input ID นี้จะมีการแบ่งข้อมูลเป็น Training Data และ Testing Data หรือไม่ ซึ่งเก็บ 2 ค่าดังนี้ Y : Validation คือ มีการแบ่งข้อมูลออกเป็น 2 ส่วน N : ไม่ Validating คือ ไม่มีการแบ่งข้อมูลเพื่อนำไว้ใช้ในการทดสอบ
testing_set	Testing Set	ตัวเลขเปอร์เซ็นต์ของ Testing Data
max_levl	Maximun Level	จำนวนระดับสูงสุดที่แบบจำลองต้นไม้สามารถแตกได้
min_data	Minimum Level	จำนวนข้อมูลน้อยสุดในแต่ละโหนด
total_recd	Total Record	จำนวน Training Data ทั้งหมด
total_load	Total Load	จำนวนข้อมูลทั้งหมดที่นำมาใช้ในการวิเคราะห์

2. ตาราง: DM_DATA_LOAD เป็นตารางที่ใช้ในการเก็บข้อมูลที่จะนำมาวิเคราะห์ โดยที่ข้อมูลที่อยู่ในตารางนี้จะถือเป็นข้อมูลดิบ (Original Data) ซึ่งประกอบด้วยฟิลด์ดังต่อไปนี้

ตารางที่ 4.3 Table: DM_DATA_LOAD

Field	Type	Format	Req.	PK or FK	FK Reference Table
inpt_id	Int(4)	9999	Y	PK,FK	DM_DATA_DETL
var_id	Int(4)	9999	Y	PK,FK	DM_DATA_HEAD
seqn_recd	Int(4)	9999	Y	PK	
var_vlue	Varchar(50)	xxxxx	Y		

คำอธิบายฟิลด์ของ Table: DM_DATA_LOAD

ตารางที่ 4.4 คำอธิบายฟิลด์ของ Table: DM_DATA_LOAD

Field	Content	Description
inpt_id	Input ID	รหัสชุดข้อมูล
var_id	Variable ID	รหัสของตัวแปรแต่ละตัวที่นำมาวิเคราะห์ ค่าเริ่มต้นคือ 1
seqn_recd	Sequence Record	ลำดับที่ของข้อมูล
var_vlue	Variable Value	ค่าของข้อมูลที่นำมาวิเคราะห์

3. ตาราง : DM_DATA_HEAD เป็นตารางที่เก็บชื่อตัวแปรและประเภทของข้อมูลวิเคราะห์ซึ่งประกอบด้วยฟิลด์ดังต่อไปนี้

ตารางที่ 4.5 Table: DM_DATA_HEAD

Field	Type	Format	Req.	PK or FK	FK Reference Table
inpt_id	Int(4)	9999	Y	PK,FK	DM_DATA_DETL
var_id	Int(4)	9999	Y	PK	
var_name	Varchar(50)	xxxxx	Y		
var_type	Int(4)	9999	Y		
selected	Char(1)	X	Y		

คำอธิบายฟิลด์ของ Table: DM_DATA_HEAD

ตารางที่ 4.6 คำอธิบายฟิลด์ของ Table: DM_DATA_HEAD

Field	Content	Description
inpt_id	Input ID.	รหัสชุดข้อมูล
var_id	Variable ID.	รหัสของตัวแปรแต่ละตัวที่นำมาวิเคราะห์
var_name	Variable Name	ชื่อของตัวแปรที่นำมาวิเคราะห์
var_type	Variable Type	ประเภทของตัวแปรที่นำมาวิเคราะห์ ซึ่งในที่นี้เก็บเป็นค่า 0 และ 1 โดยมีความหมายดังนี้ 0 : ตัวแปรที่มีค่าเป็น Numeric 1 : ตัวแปรที่มีค่าเป็น Category
selected	Select	ตัวแปรที่ถูกเลือกเพื่อนำมาวิเคราะห์ 0 : ตัวแปรที่ไม่ถูกเลือกมาวิเคราะห์ 1 : ตัวแปรถูกเลือกมาวิเคราะห์ 2 : ตัวแปรที่ถูกเลือกมาเป็นคลาสเลเบิล

4. ตาราง :DM_DATA_RUN เป็นตารางที่เก็บรายละเอียดของข้อมูลที่ใช้ในการวิเคราะห์และเก็บค่าลีฟโหนด (Leaf Node) ที่ข้อมูลแต่ละตัวนั้นอยู่ ซึ่งประกอบด้วยฟิลด์ดังต่อไปนี้

ตารางที่ 4.7 Table: DM_DATA_RUN

Field	Type	Format	Required	PK or FK	FK Reference Table
inpt_id	Int(4)	9999	Y	PK,FK	DM_DATA_DETL
seqn_recd	Int(4)	9999	Y	PK,FK	DM_DATA_LOAD
var_1	Char(50)	xxxxx	N		
var_2	Char(50)	xxxxx	N		
var_3	Char(50)	xxxxx	N		
var_4	Char(50)	xxxxx	N		
var_5	Char(50)	xxxxx	N		
var_6	Char(50)	xxxxx	N		
var_7	Char(50)	xxxxx	N		
var_8	Char(50)	xxxxx	N		
var_9	Char(50)	xxxxx	N		
var_10	Char(50)	xxxxx	N		
node_numb	Int(4)	9999	Y		
class	Char(50)	xxxxx	Y		

คำอธิบายฟิลด์ของ Table: DM_DATA_RUN

ตารางที่ 4.8 คำอธิบายฟิลด์ของ Table: DM_DATA_RUN

Field	Content	Description
inpt_id	Input ID.	รหัสชุดข้อมูล
seqn_recd	Sequence Record	ลำดับของข้อมูล
node_numb	Node Number	หมายเลขโหนดที่ข้อมูลนั้นอยู่หลังการประมวลแล้ว ซึ่งก็คือค่าลีฟโหนด
var_1	Variable 1	ค่าของข้อมูลในตัวแปรที่ 1
var_2	Variable 2	ค่าของข้อมูลในตัวแปรที่ 2
var_3	Variable 3	ค่าของข้อมูลในตัวแปรที่ 3
var_4	Variable 4	ค่าของข้อมูลในตัวแปรที่ 4
var_5	Variable 5	ค่าของข้อมูลในตัวแปรที่ 5
var_6	Variable 6	ค่าของข้อมูลในตัวแปรที่ 6
var_7	Variable 7	ค่าของข้อมูลในตัวแปรที่ 7
var_8	Variable 8	ค่าของข้อมูลในตัวแปรที่ 8
var_9	Variable 9	ค่าของข้อมูลในตัวแปรที่ 9
var_10	Variable 10	ค่าของข้อมูลในตัวแปรที่ 10
class	Class	เก็บค่าคลาสเลเบิลเพื่อตรวจสอบค่าคลาสที่เหมือนกัน (ถ้าใน Node Number เดียวกันมีค่าคลาสเหมือนกันแสดงว่าโหนดนั้นมีค่าคลาสเลเบิลเพียงคลาสเดียวไม่ต้องทำการแบ่งต่อ)

5. ตาราง DM_TREE เป็นตารางที่ใช้เก็บผลของแบบจำลอง (Tree Model) ที่ได้ทำการวิเคราะห์ข้อมูลแล้วโดยมีฟิลด์ดังนี้

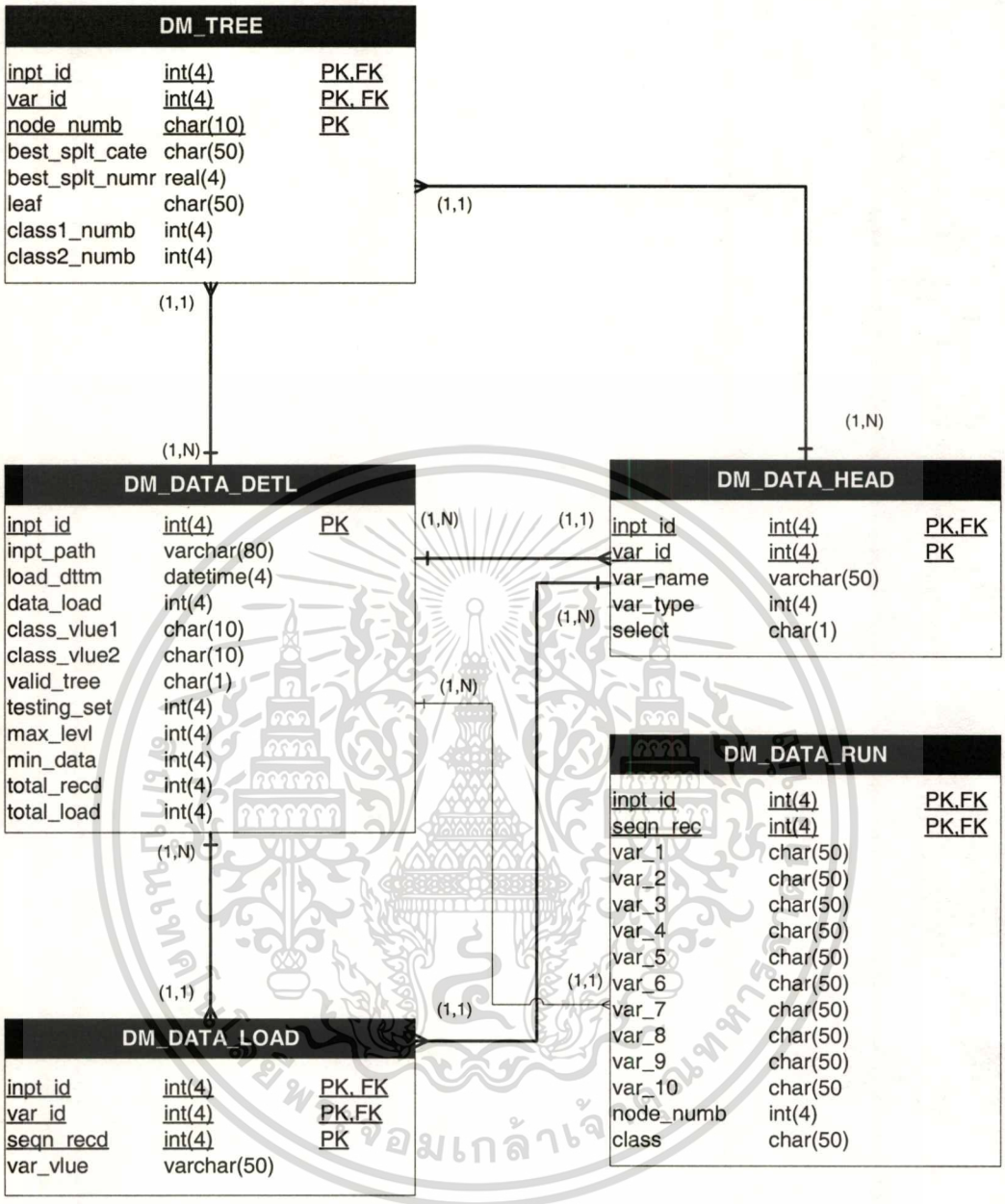
ตารางที่ 4.9 Table: DM_TREE

Field	Type	Format	Required	PK or FK	FK Reference Table
inpt_id	Int(4)	9999	Y	PK,FK	DM_DATA_DETL
var_id	Int(4)	9999	Y	PK,FK	DM_DATA_HEAD
node_numb	char(10)	xxxxx	Y	PK	
best_splt_catategory	char(50)	xxxxxx	N		
best_splt_numeric	real(4)	9999.99	N		
leaf	char(50)	xxxxx	Y		
class1_numb	Int(4)	9999	Y		
class2_numb	Int(4)	9999	Y		

คำอธิบายฟิลด์ของ Table: DM_TREE

ตารางที่ 4.10 คำอธิบายฟิลด์ของ Table: DM_TREE

Field	Content	Description
inpt_id	Input ID.	รหัสชุดข้อมูล
var_id	Variable ID.	รหัสของตัวแปรแต่ละตัวที่นำมาวิเคราะห์
node_numb	Node Number	แสดงหมายเลขโหนดของแบบจำลองต้นไม้ (Tree Model) ที่ผ่านการวิเคราะห์แล้ว ซึ่งจะแสดงให้เห็นว่าผลลัพธ์สุดท้ายของแบบจำลองต้นไม้ (Tree Model) ที่ได้นั้นมีกี่โหนด
best_splt_category	Best Split Category	ค่า Best Split ของข้อมูลที่เป็น category จะแสดงค่าเป็น "-" ในกรณีที่ตัวแปรนั้นเป็นข้อมูลประเภท Numeric
best_splt_numeric	Best Split Numeric	ค่า Best Split ของข้อมูลที่เป็น Numeric จะแสดงค่าเป็น "-" ในกรณีที่ตัวแปรนั้นเป็นข้อมูลประเภท category
leaf	Leaf	เป็นแอตทริบิวต์ที่บอกว่าหมายเลขโหนด (Node Number) นี้มีค่าเป็นลีฟหรือไม่ซึ่งถ้ากรณีที่โหนดไหนยังมีค่าคลาสเลเบิล 2 ค่าปนกัน ค่าในฟิลด์นี้จะแสดงเป็น "-" และถ้าโหนดไหนมีค่าคลาสเลเบิลแค่ค่าเดียวก็จะแสดงเป็นค่าคลาสเลเบิลนั้น
class1_numb	Number of Class1	จำนวนเรคคอร์ดของคลาสที่ 1 ที่อยู่ในโหนดนั้น
class2_numb	Number of Class2	จำนวนเรคคอร์ดของคลาสที่ 2 ที่อยู่ในโหนดนั้น

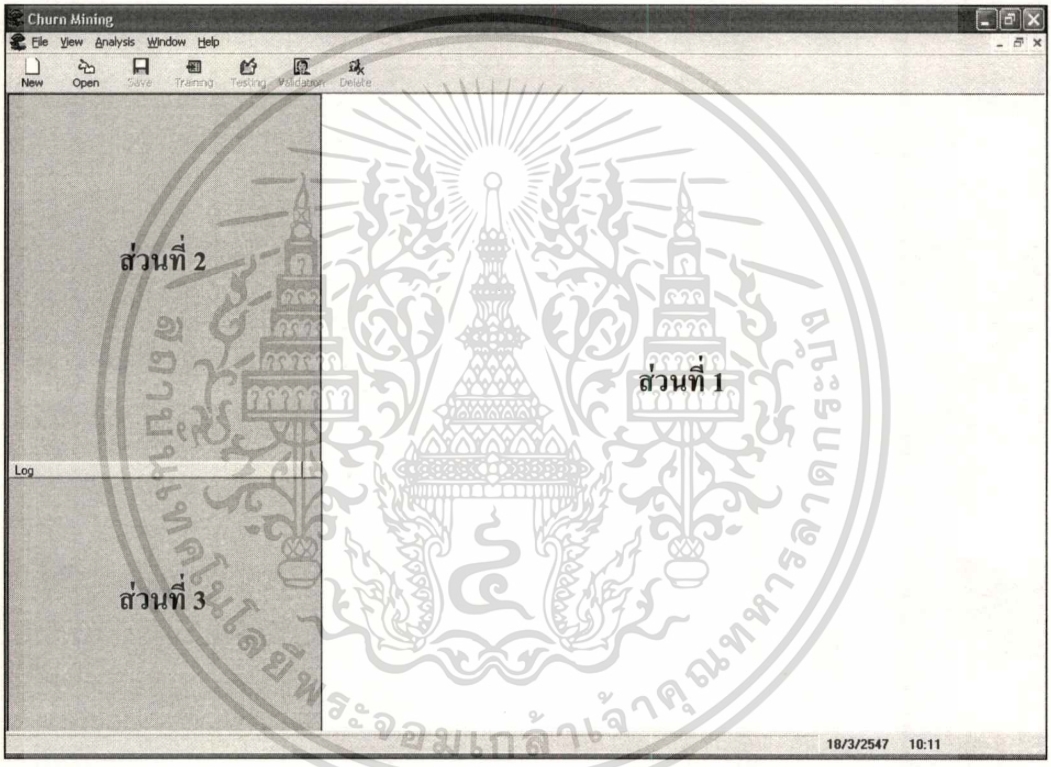


รูปที่ 4.1 Relational Database

4.4.2 การวิเคราะห์และออกแบบส่วนของหน้าจอ

หน้าจอหลักของระบบนี้แบ่งออกเป็น 5 หน้าจอหลักดังนี้

1. หน้าจอหลัก เป็นหน้าจอที่เห็นในการเข้าโปรแกรมครั้งแรกและเป็นหน้าจอที่ใช้ในการแสดงผลการวิเคราะห์ข้อมูลซึ่งในหน้าจอนี้จะแบ่งออกเป็น 3 ส่วนดังนี้
   - ส่วนที่ 1 แสดงแบบจำลองต้นไม้และค่าความเชื่อมั่น
   - ส่วนที่ 2 แสดงแบบจำลองต้นไม้รูปเล็กเพื่อการมองแบบจำลองโดยรวม
   - ส่วนที่ 3 แสดงขั้นตอนการทำงานของโปรแกรมและข้อผิดพลาด



รูปที่ 4.2 หน้าจอหลัก

ในหน้าจอหลักจะพบเมนูและไอคอนต่างๆ โดยมีหลักการทำงานดังต่อไปนี้

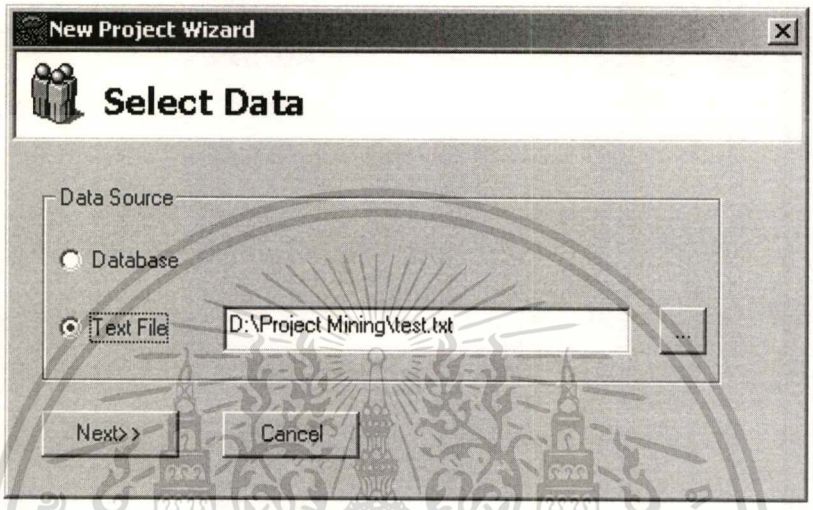
ตารางที่ 4.11 การทำงานของเมนูและไอคอน

Menu	Sub Menu	Icon	Description
File			เป็นเมนูเกี่ยวกับการเปิด บันทึกและปิดโปรเจค
	New Project		เรียกใช้เมื่อต้องการเปิดโปรเจคใหม่ หรือนำข้อมูลชุดใหม่เข้ามาวิเคราะห์
	Open Project		เรียกใช้เมื่อต้องการที่จะเปิดโปรเจคเดิมหรือผลลัพธ์ของข้อมูลเดิมที่ได้ทำการคำนวณไว้แล้ว
	Save Project		เรียกใช้เมื่อต้องการบันทึกผลการวิเคราะห์โปรเจค
	Save Project As	-	เรียกใช้เมื่อต้องการจะเปลี่ยนชื่อผลการวิเคราะห์โปรเจค
	Exit	-	เรียกใช้เมื่อต้องการออกจากระบบ
View			เป็นเมนูเกี่ยวกับมุมมองในการดูข้อมูล
	Training Data		เรียกใช้เมื่อต้องการดูผลลัพธ์ของข้อมูลที่นำมาสร้างแบบจำลองต้นไม้ (Tree Model)
	Testing Data		เรียกใช้เมื่อต้องการดูผลลัพธ์จากข้อมูลที่แบ่งไว้เพื่อใช้ในการทดสอบ
Analysis			เป็นเมนูที่ใช้กำหนดเงื่อนไขในการสร้างแบบจำลองต้นไม้ (Tree Model)
	Validation		เรียกใช้เมื่อต้องการกำหนดเงื่อนไขการสร้างแบบจำลองต้นไม้ (Tree Model)
Window			
Help			เป็นหน้าจอที่ใช้แสดงรายละเอียดเกี่ยวกับโปรแกรม
	About	-	

เมนูหรือไอคอนต่างๆ ที่ไม่แสดงในครั้งแรกที่เปิดโปรแกรมจะสามารถใช้งานได้หลังจากที่ได้ทำการโหลดข้อมูล เลือกประเภทของตัวแปร เลือกตัวแปรที่จะนำมาวิเคราะห์ และกำหนดข้อจำกัดในการสร้างแบบจำลองต้นไม้เรียบร้อยแล้ว (New Project Wizard) ยกเว้นในส่วนของ View ถ้าระบบกำลังแสดงผลของ Training Data ในส่วนเมนูและไอคอนของ Testing Data จะไม่

แสดง และเช่นเดียวกันถ้าระบบกำลังแสดงผลของ Testing Data ในส่วนเมนูและไอคอนของ Training Data ก็จะไม่แสดง

2. หน้าจอการโหลดข้อมูลจากเท็กซ์ไฟล์และฐานข้อมูล

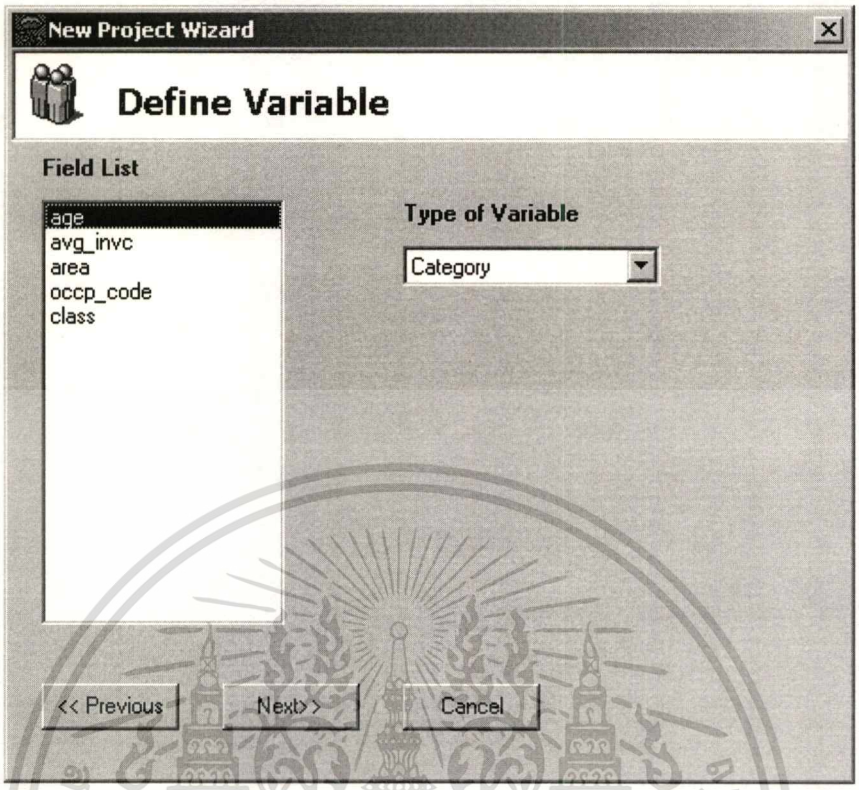


รูปที่ 4.3 หน้าจอการโหลดข้อมูล

ในหน้าจอ Select Data นั้นจะแบ่งออกเป็น 2 ส่วน ดังนี้

- Database หมายถึงกรณีที่ข้อมูลที่ต้องการวิเคราะห์นั้นเก็บอยู่ในฐานข้อมูล MS SQL 2000
- Text File หมายถึง กรณีที่ข้อมูลที่ต้องการวิเคราะห์เก็บอยู่ในเท็กซ์ไฟล์ (Text File) ซึ่งเมื่อผู้ใช้ระบบเลือก Text File ระบบจะแสดงปุ่ม Browse เพื่อให้ผู้ใช้ระบบเลือกเท็กซ์ไฟล์ที่ต้องการจะนำมาวิเคราะห์

3. หน้าจอ Define Variable เป็นหน้าจอที่ใช้กำหนดประเภทของตัวแปรว่าเป็นตัวแปรแบบ Numeric หรือ Category

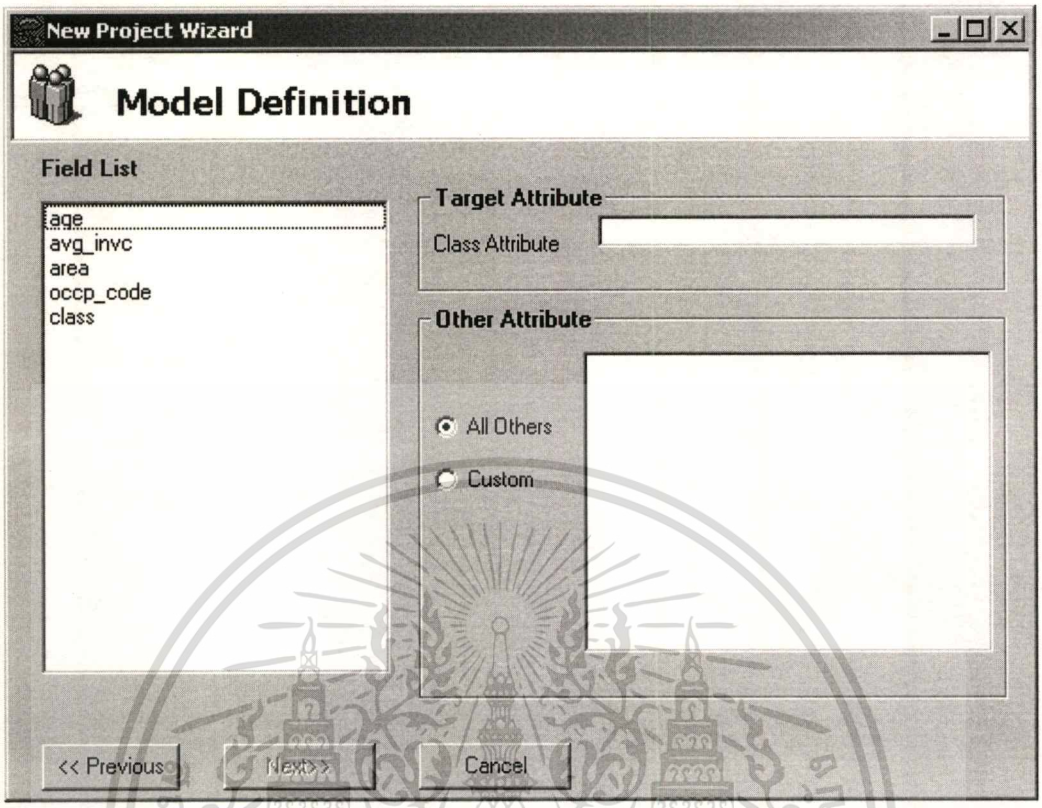


รูปที่ 4.4 หน้าจอการกำหนดประเภทของตัวแปร

ในหน้าจอ Define Variable ประกอบด้วย 2 ส่วนดังนี้

- Field List เป็นส่วนที่แสดงตัวแปรทั้งหมดที่สามารถนำมาวิเคราะห์ได้ ซึ่งระบบจะทำการดึงชื่อตัวแปรขึ้นมาให้ทั้งหมด
- Type of Variable เป็นส่วนที่ใช้ในการกำหนดประเภทของตัวแปรซึ่งระบบจะทำการตั้งค่าเริ่มต้นทุกตัวแปรเป็น Category เพื่อป้องกันในกรณีที่ผู้ใช้ระบบไม่ได้ตั้งค่าตัวแปรคลาสเลเบิล

4. หน้าจอกำหนดตัวแปรที่ใช้ในการวิเคราะห์ (Model Definition) เป็นหน้าจอที่ใช้กำหนด C คลาสเลเบิลและ ตัวแปรที่จะนำมาวิเคราะห์



รูปที่ 4.5 หน้าจอกำหนดตัวแปรที่ใช้ในการวิเคราะห์

หน้าจอ Model Definition นั้นแบ่งเป็น 3 ส่วนดังนี้

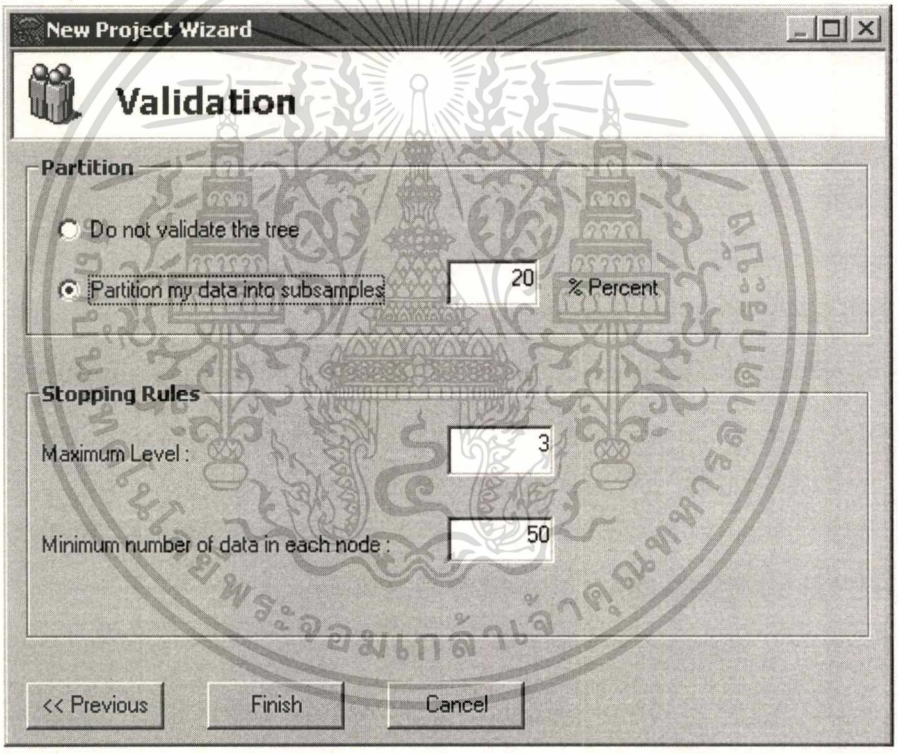
- Field List เป็นส่วนที่แสดงตัวแปรทั้งหมดที่ระบบสามารถนำมาวิเคราะห์ได้
- Target Attribute เป็นส่วนที่ใช้ในการกำหนด Class Label
- Other Attributes เป็นส่วนที่ใช้ในการกำหนดตัวแปรที่นำมาวิเคราะห์ร่วมกับคลาสเลเบิล ซึ่งใน Other Attributes นั้นได้แบ่งออกเป็น 2 ส่วนดังนี้
    - All Others เป็นส่วนที่แสดงว่าในการวิเคราะห์นี้จะใช้ตัวแปรทุกตัวที่เหลืออยู่ใน Field List มาวิเคราะห์ร่วมกับคลาสเลเบิล
    - Custom เป็นส่วนที่แสดงว่าผู้ใช้ระบบจะเลือกตัวแปรเพียงบางตัวมาวิเคราะห์ร่วมกับคลาสเลเบิล

ซึ่งการทำงานในหน้าจอนี้มีข้อตกลงในการใช้งานดังนี้

- ในการเลือกตัวแปรต่างๆ ที่นำมาวิเคราะห์นั้นจะใช้วิธีลากแล้วปล่อย (Drag and Drop)เพื่อสะดวกและง่ายต่อการใช้งาน

- การที่ผู้ใช้ระบบจะทำงานในหน้าจอถัดไปได้นั้นต้องกำหนด Class Label ก่อนเสมอปุ่ม Next จึงจะสามารถทำงานได้
- ในกรณีผู้ใช้ระบบเลือกที่ All Others นั้นไม่จำเป็นที่จะต้องลากตัวแปรมาไว้ที่ช่อง Other Attributes และถ้าผู้ใช้ระบบเลือกตัวแปรเข้ามา ระบบจะทำการเปลี่ยนค่าที่ตั้งไว้เป็น Custom ทันที

5. หน้าจอ Validation เป็นหน้าจอที่ ใช้ในการกำหนดส่วนแบ่งของข้อมูลที่จะนำมาวิเคราะห์และจะนำมาทดสอบ และเป็นหน้าจอที่ใช้ในการกำหนด Stopping Rule ซึ่งเป็นเงื่อนไขในการแตกแบบจำลองต้นไม้ (Tree Model)



รูปที่ 4.6 หน้าจอ Validation

ในหน้าจอ Validation นั้นได้แบ่งออกเป็น 2 ส่วนดังนี้

- Partition เป็นส่วนที่ใช้ในการแบ่งข้อมูลไว้ส่วนหนึ่งเพื่อใช้เป็นข้อมูลที่ใช้ในการทดสอบแบบจำลองต้นไม้ ซึ่งการแบ่งข้อมูลไว้สำหรับทดสอบนั้นจะคิดเป็นเปอร์เซ็นต์ของจำนวนข้อมูลทั้งหมด
- Stopping Rules เป็นส่วนที่ใช้ในการกำหนดกฎในการหยุดแตกแบบจำลองต้นไม้ ซึ่งแบ่งออกเป็น 2 ส่วนดังนี้
    - Maximum Level : ใช้ในการกำหนดระดับของแบบจำลองต้นไม้ที่สามารถแตกได้ ซึ่งระดับ Root จะถือเป็นระดับที่ 1
    - Minimum number of data in each node : ใช้ในการกำหนดข้อมูลต่ำสุดของแต่ละโหนดที่สามารถนำมาแตกแบบจำลองต้นไม้ได้ ถ้ามีจำนวนข้อมูลน้อยกว่าค่าที่กำหนดไว้ระบบก็จะทำการหยุดแตกแบบจำลองต้นไม้

ในบทนี้จะกล่าวถึงแค่หน้าจอการใช้งานและหลักการทำงานในแต่ละหน้าจอเท่านั้น ซึ่งวิธีการใช้งานนั้นจะได้กล่าวไว้อีกทีในบทที่ 5 คู่มือการใช้งานระบบ

4.5 การเขียนโปรแกรมและทดสอบระบบ

เมื่อได้ทำการออกแบบฐานข้อมูลและหน้าจอเรียบร้อยแล้วก็ได้ทำการเขียนโปรแกรมและทดสอบโปรแกรมซึ่งในการเขียนโปรแกรมนั้นมีส่วนที่สำคัญ 3 ส่วนดังต่อไปนี้

- ส่วนโหลดข้อมูล
- ส่วนวิเคราะห์ข้อมูล
- ส่วนแสดงผล

การทดสอบโปรแกรมนั้นได้ทำการทดสอบในส่วนของ Unit Test, Integrate Test และ System Test เพื่อทดสอบว่าระบบนั้นสามารถทำงานได้ตามความต้องการหรือไม่และได้ทำการทดสอบในส่วนของ User Acceptance Test เพื่อดูว่าระบบที่ได้พัฒนานั้นง่ายต่อการใช้งานและผู้ใช้ระบบมีความพอใจหรือไม่

4.6 การทำเอกสารประกอบระบบ

ในการทำเอกสารประกอบระบบนั้นได้จัดทำเป็นคู่มือการใช้ระบบซึ่งได้กล่าวไว้ในบทที่ 5 คู่มือการใช้งานระบบ

4.7 การติดตั้งระบบ

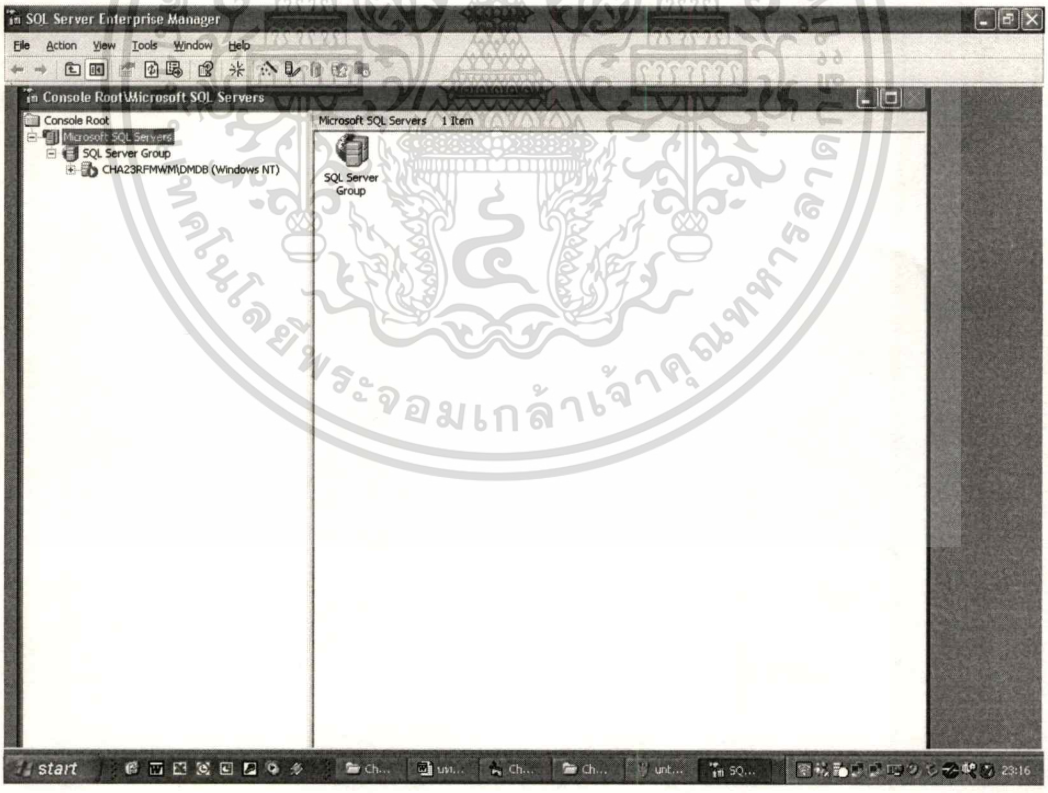
ในการติดตั้งระบบนั้นทางผู้พัฒนาระบบได้ทำเป็นตัว Install ซึ่งคู่มือในการติดตั้งระบบนั้นได้กล่าวไว้ในบทที่ 5 คู่มือการใช้งานระบบ

บทที่ 5
คู่มือการใช้งานระบบ

คู่มือการใช้งานระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีนั้นจะแบ่งออกเป็น 2 ส่วนคือส่วนการติดตั้งโปรแกรมและส่วนการใช้งานโปรแกรมตามรายละเอียดที่จะกล่าวถึงดังนี้

5.1 ส่วนการติดตั้งโปรแกรม

1. ให้ผู้ใช้ทำการติดตั้งโครงสร้างฐานข้อมูลโดยเครื่องที่จะติดตั้งนั้นต้องมีโปรแกรม MS SQL Server 2000 เมื่อได้ทำการติดตั้งโปรแกรมเรียบร้อยแล้วให้ทำการเข้า MS SQL Server -> Enterprise Manager จะพบหน้าจอดังรูปที่ 5.1

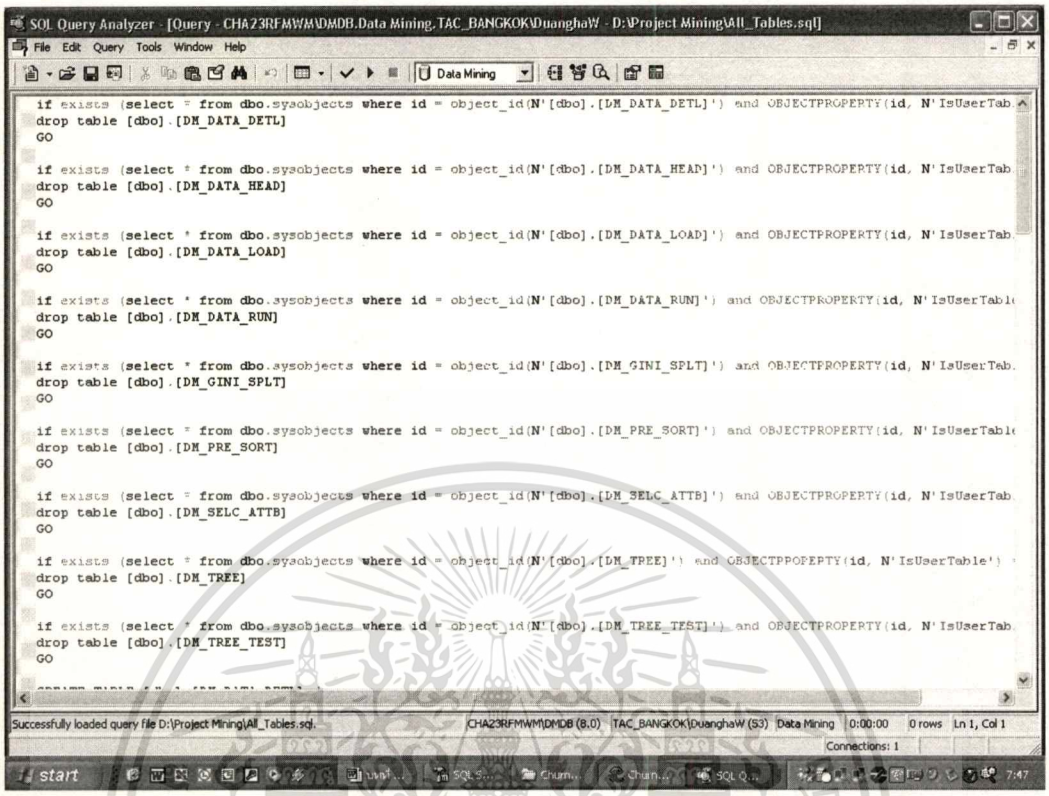


รูปที่ 5.1 หน้าจอ MS SQL Server 2000

2. เข้าที่เมนู Tool ->SQL Query Analyzer จะพบหน้าจอดังรูปที่ 5.2

รูปที่ 5.2 SQL Query Analyzer

3. เมื่อเข้าหน้าจอ SQL Query Analyzer ให้ทำการเปิดไฟล์ All_Tables.sql โดยเข้าที่ File -> Open จะพบหน้าจอดังรูปที่ 5.3



รูปที่ 5.3 All_Tables.sql

4. ทำการกดปุ่ม Run เพื่อทำการติดตั้งโครงสร้างฐานข้อมูล

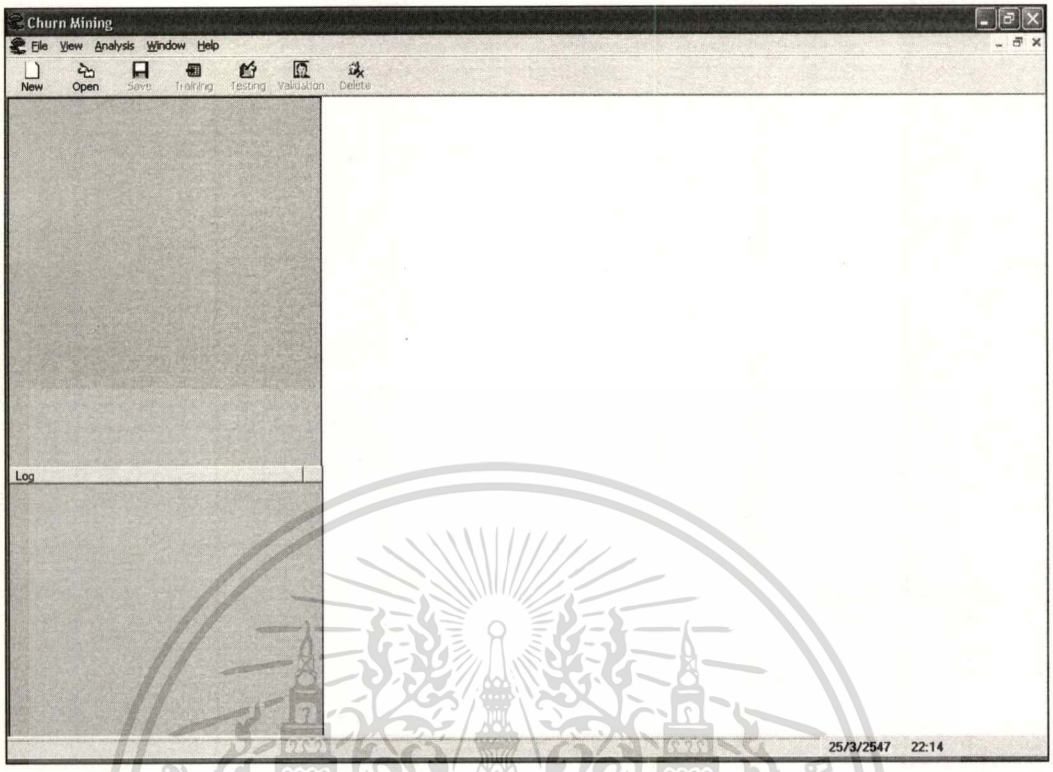
5. ให้ผู้ใช้ทำการดับเบิ้ลคลิกที่ตัว ChurnMining.exe เพื่อเข้าสู่โปรแกรม Churn Mining (เครื่องของผู้ใช้ควรเป็น Windows XP หรือเวอร์ชั่นที่สูงกว่า)

5.2 ส่วนการใช้งานโปรแกรม

ขั้นตอนการใช้งานโปรแกรมมีรายละเอียดดังนี้

5.2.1 การเรียกใช้โปรแกรมและการนำข้อมูลเข้า

ผู้ใช้ระบบสามารถทำการเรียกใช้โปรแกรมได้โดยเลือกที่ Churn Mining.exe จะพบหน้าจอดังรูปที่ 5.4 ซึ่งถือเป็นหน้าจอหลักในการทำงาน



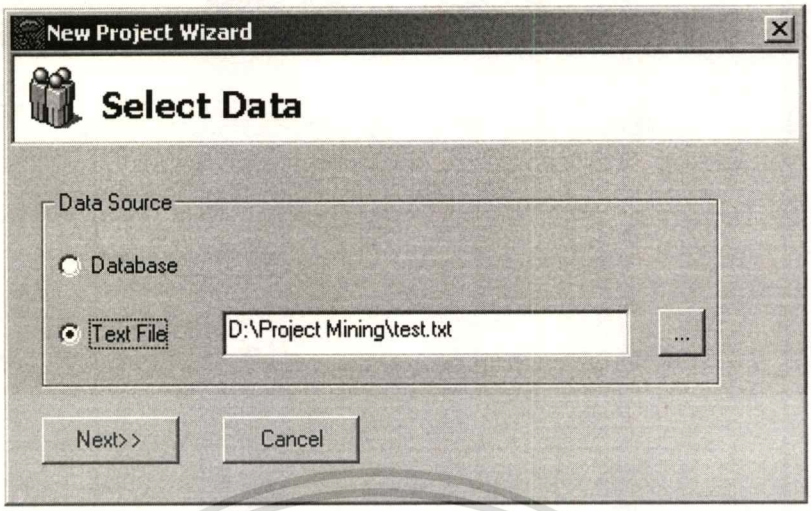
รูปที่ 5.4 หน้าจอหลัก

ซึ่งในส่วนของหน้าจอหลักจะพบเมนูต่างๆ ดังตารางที่ 5.1

ตารางที่ 5.1 คำอธิบายเมนูและไอคอน

Menu	Sub Menu	Icon	Description
File			เป็นเมนูเกี่ยวกับการเปิด บันทึกและปิดระบบ
	New Project		เรียกใช้เมื่อต้องการเปิดโปรเจคใหม่ หรือนำข้อมูลชุดใหม่เข้ามาวิเคราะห์
	Open Project		เรียกใช้เมื่อต้องการที่จะเปิดโปรเจคเดิมหรือผลลัพธ์ของข้อมูลเดิมที่ได้ทำการคำนวณไว้แล้ว
	Save Project		เรียกใช้เมื่อต้องการบันทึกผลการวิเคราะห์โปรเจค
	Save Project As	-	เรียกใช้เมื่อต้องการจะเปลี่ยนชื่อผลการวิเคราะห์โปรเจค
	Exit	-	เรียกใช้เมื่อต้องการออกจากระบบ
View			เป็นเมนูเกี่ยวกับมุมมองในการดูข้อมูล
	Training Data		เรียกใช้เมื่อต้องการดูผลลัพธ์ของข้อมูลที่นำมาสร้างแบบจำลองต้นไม้ (Tree Model)
	Testing Data		เรียกใช้เมื่อต้องการดูผลลัพธ์จากข้อมูลที่แบ่งไว้เพื่อใช้ในการทดสอบ
Analysis			เป็นเมนูที่ใช้กำหนดเงื่อนไขในการสร้างแบบจำลองต้นไม้ (Tree Model)
	Validation		เรียกใช้เมื่อต้องการกำหนดเงื่อนไขการสร้างแบบจำลองต้นไม้ (Tree Model)
Window			เป็นหน้าจอที่ใช้แสดงหน้าต่างที่เปิดใช้
Help			เป็นหน้าจอที่ใช้แสดงรายละเอียดเกี่ยวกับโปรแกรม
	About	-	

เมื่อผู้ใช้ระบบต้องการที่จะวิเคราะห์ข้อมูลให้ทำการเลือกไปที่เมนู File ->New Project จะพบหน้าจอ Select Data ดังรูปที่ 5.5



รูปที่ 5.5 หน้าจอนำเข้าข้อมูล

ซึ่งการนำข้อมูลเข้าไปในระบบนั้นสามารถทำได้ 2 วิธีคือ

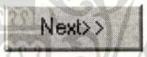
- นำข้อมูลเข้าจากฐานข้อมูล

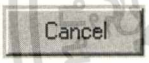
ในการนำเข้าข้อมูลจากฐานข้อมูลจะมีเงื่อนไขดังนี้คือ ฐานข้อมูลจะต้องเป็นข้อมูลที่มาจาก Microsoft SQL Server 2000 เท่านั้น ซึ่งในระบบนี้ได้มีการสร้างฐานข้อมูลไว้ให้ผู้ใช้ระบบสามารถนำข้อมูลเหล่านี้มาวิเคราะห์ได้ ซึ่งเป็นข้อมูลลูกค้าที่ปิดบริการการใช้โทรศัพท์มือถือไปแล้วในระบบแบบใช้ก่อน จ่ายทีหลัง (Post-Paid) ระหว่างวันที่ 01/07/2002 – 31/12/2002 ทั้งหมด 1,500 เรคคอร์ด โดยมีตัวแปรที่สามารถนำมาใช้พิจารณาดังตารางที่ 5.2

ตารางที่ 5.2 ตารางตัวแปรที่นำมาวิเคราะห์

Variable	Description
Customer Age Current	อายุของลูกค้า ณ วันที่ปิดบริการ
Subscriber Age	อายุการใช้งานโทรศัพท์มือถือมีหน่วยเป็นเดือน
Average Invoice	ค่าใช้จ่ายย้อนหลังโดยเฉลี่ย 3 เดือนก่อนลูกค้าปิดบริการ
Credit Limit	วงเงินในการใช้โทรศัพท์มือถือ
Occupation Code	อาชีพของลูกค้า
Package Current Code	โปรโมชั่นที่ลูกค้าใช้ก่อนปิดบริการ
Area Code	จุดจดทะเบียน
Class	เป็นฟิลด์ที่ใช้ในการทำนายมี 2 ค่าคือ Yes กับ No

ซึ่งเมื่อผู้ใช้ระบบต้องการโหลดข้อมูลจากฐานข้อมูลให้เลือกที่ 

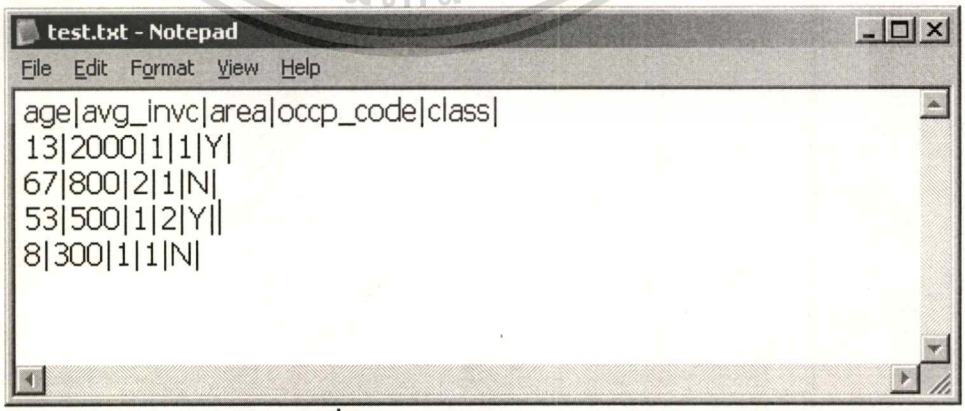
และกดปุ่ม Next 

จะพบหน้าจอ Define Variable หรือกดปุ่ม Cancel 

เพื่อทำการยกเลิกการโหลดข้อมูล ซึ่งจะได้กล่าวถึงในขั้นตอนถัดไป

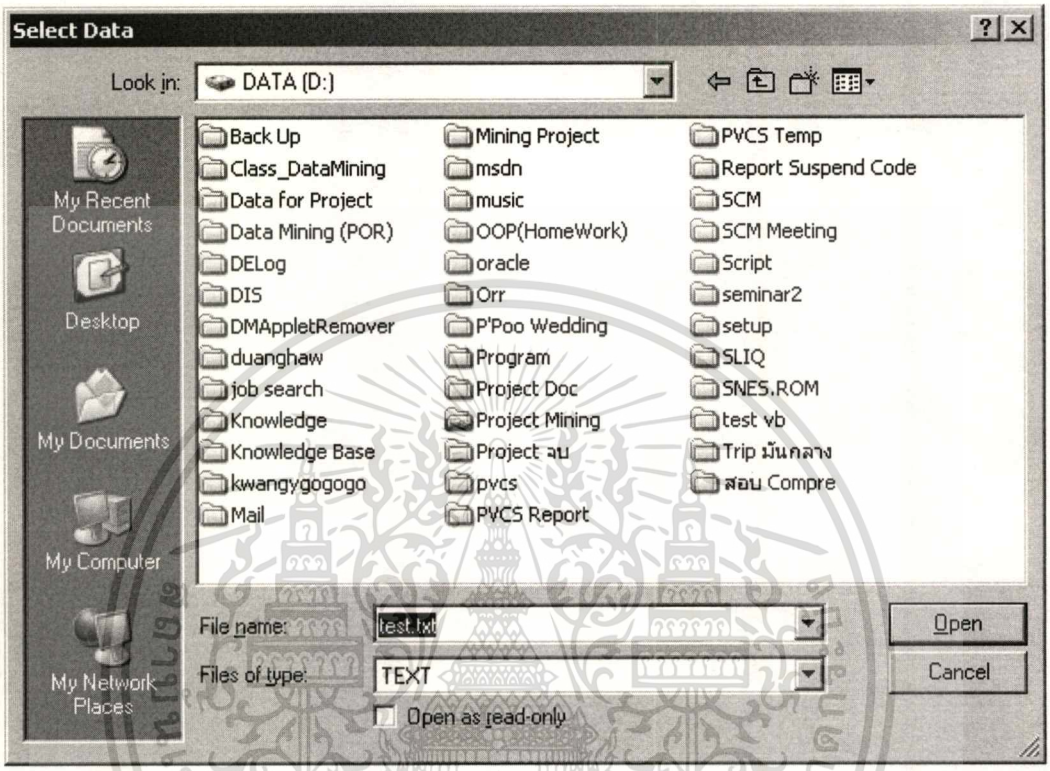
- นำข้อมูลเข้าจากเท็กซ์ไฟล์

ในการโหลดข้อมูลจากเท็กซ์ไฟล์เข้าโปรแกรมนั้น ข้อมูลจะต้องมีรูปแบบของข้อมูลดังนี้ คือ ชื่อตัวแปร (Attribute) จะอยู่ในแถวแรกและข้อมูลแต่ละตัวจะขั้นด้วยไปป์ (|) ดังรูปที่ 5.6



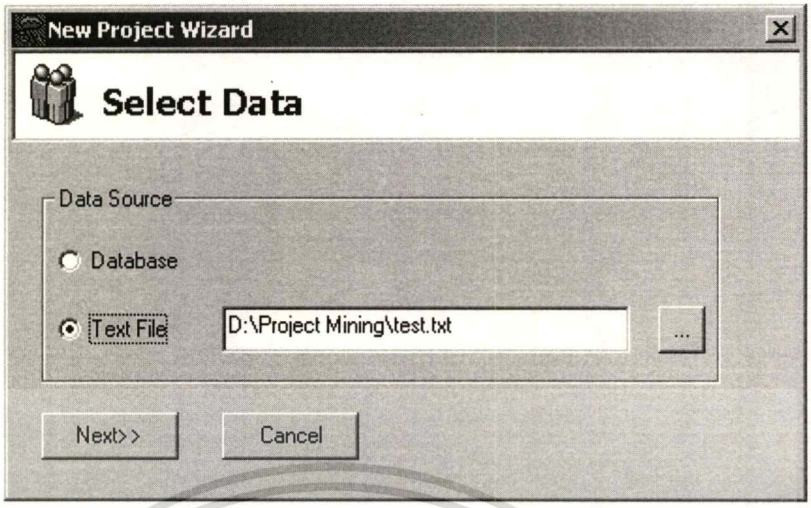
รูปที่ 5.6 ตัวอย่างข้อมูลเท็กซ์ไฟล์

ถ้าผู้ใช้ระบบต้องการโหลดข้อมูลจากเท็กซ์ไฟล์ ให้เลือกที่ Text File และกดปุ่ม Browse จะพบหน้าจอ Select Data ดังรูปที่ 5.7

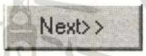


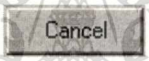
รูปที่ 5.7 หน้าจอ Select Data

ให้ทำการเลือกเท็กซ์ไฟล์ที่จะนำมาวิเคราะห์ ซึ่งเมื่อทำการเลือกข้อมูลเรียบร้อย ระบบจะแสดงไฟล์ข้อมูลดังรูปที่ 5.8



รูปที่ 5.8 หน้าจอแสดงเท็กซ์ไฟล์

เมื่อเลือกเท็กซ์ไฟล์เสร็จแล้วให้กดปุ่ม Next 

จะพบหน้าจอ Define Variable หรือ กดปุ่ม Cancel 

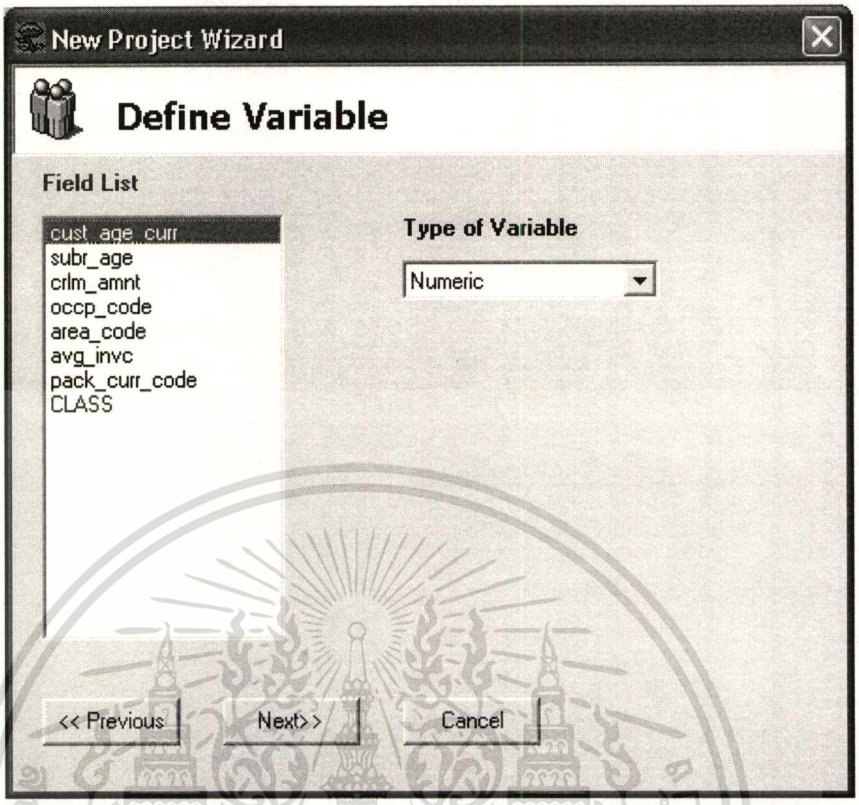
เพื่อทำการยกเลิกการโหลดข้อมูล ซึ่งจะได้กล่าวถึงในขั้นตอนถัดไป

5.2.2 การกำหนดประเภทของตัวแปร

เมื่อกดปุ่ม Next จากหน้าจอ Select Data จะพบหน้าจอ Define Variable ดังรูปที่ 5.9 ซึ่งเป็นหน้าจอที่ใช้ในการกำหนดประเภทของตัวแปรว่าตัวแปรที่นำมาคิดนั้นเป็นตัวแปรประเภท Numeric หรือ Category ซึ่งต้องทำการกำหนดทุกตัวแปร

หน้าจอ Define Variable นั้นแบ่งเป็น 2 ส่วนดังนี้

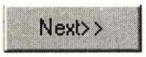
- Field List เป็นส่วนที่แสดงตัวแปรทั้งหมดที่สามารถนำมาวิเคราะห์ได้
- Type of Variable เป็นส่วนที่ใช้กำหนดประเภทของตัวแปร



รูปที่ 5.9 หน้าจอ Define Variable

5.2.2.1 ขั้นตอนการกำหนดประเภทของตัวแปร

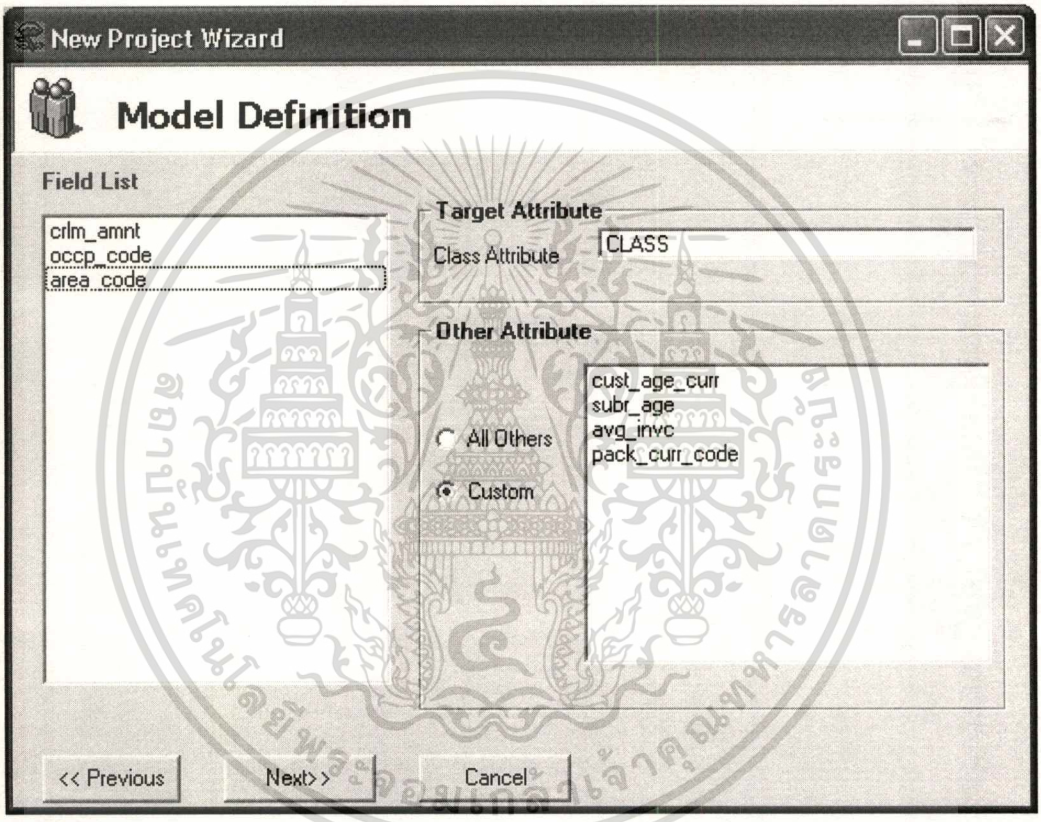
ขั้นตอนการกำหนดประเภทของตัวแปรมีดังนี้

1. ทำการเลือกตัวแปรที่ต้องการกำหนดค่าประเภทในช่อง Field List สังเกตว่าตัวแปรนั้นจะขึ้นเป็นแถบสีน้ำเงิน
2. ทำการเลือกประเภทของตัวแปรในช่อง Type of Variable โดยกดที่ Drop down List ซึ่งตัวแปรใดที่กำหนดเป็นคลาสเลเบิล ต้องกำหนดเป็น Category เท่านั้น
3. กดปุ่ม Next  เพื่อเข้าสู่หน้าจอ Model Definition ดังรูปที่ 5.10
4. กดปุ่ม previous  เพื่อกลับเข้าสู่หน้าจอ Select Data

5.2.3 การกำหนดตัวแปรที่ใช้ในการคำนวณ

หน้าจอ Model Definition เป็นหน้าจอที่ใช้ในการกำหนดค่า Class Attribute และตัวแปรต่างๆ ที่ใช้ในการวิเคราะห์ ซึ่งหน้าจอ Model Definition แบ่งเป็น 3 ส่วนดังนี้

- Field List เป็นส่วนที่แสดงตัวแปรทั้งหมดที่สามารถนำมาวิเคราะห์ได้
- Target Attribute เป็นส่วนที่ให้ผู้ใช้ระบบกำหนดตัวแปรที่จะนำมาใช้เป็น Class Attribute
- Other Attribute เป็นส่วนที่ให้ผู้ใช้ระบบเลือกตัวแปรที่จะนำมาคำนวณ

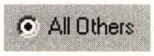
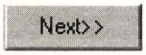
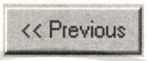


รูปที่ 5.10 หน้าจอ Model Definition

5.2.3.1 ขั้นตอนการกำหนดตัวแปรที่ใช้ในการคำนวณ

ขั้นตอนการกำหนดตัวแปรที่ใช้ในการคำนวณมีดังนี้

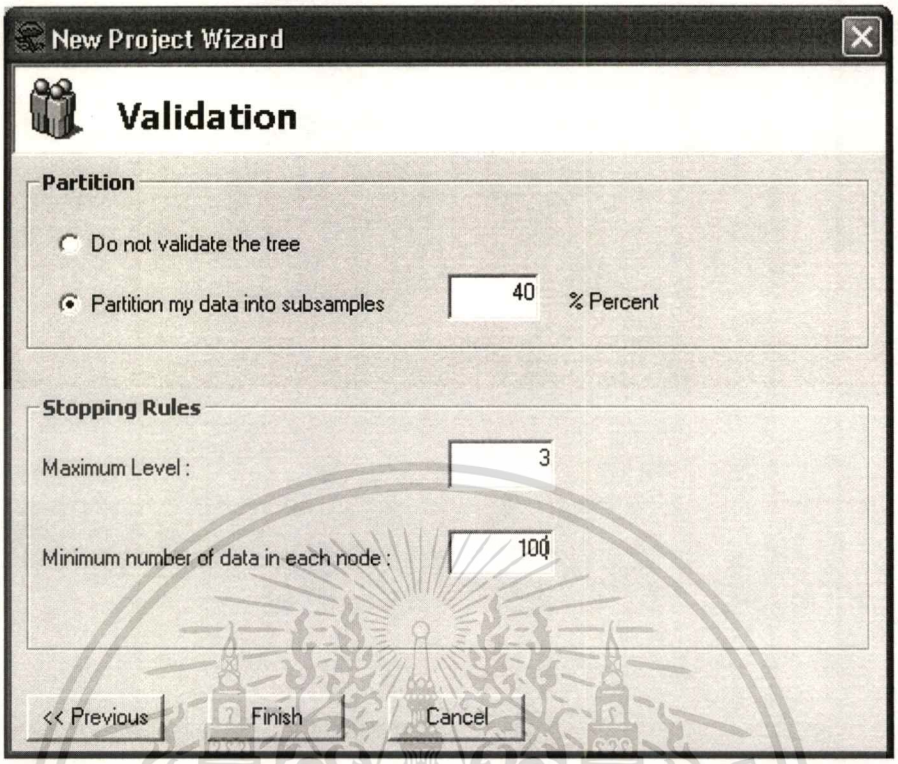
1. ทำการกำหนดตัวแปรที่จะใช้เป็น Class Attribute โดยการลากตัวแปรในส่วน Field List มาใส่ในช่อง Class Attribute

2. ทำการกำหนดตัวแปรที่จะใช้วิเคราะห์ร่วมกับ Class Attribute โดยการลากตัวแปรในส่วน Field List มาใส่ในส่วน Other Attribute โดยถ้าต้องการเลือกตัวแปรทั้งหมดที่ไม่ใช่ Class Attribute มาคำนวณให้ทำการเลือกที่ All Others  ถ้าในกรณีที่ต้องการเลือกแค่บาง Attribute ให้ทำการเลือกที่ Custom  และทำการลากตัวแปรที่ต้องการมาใส่ในส่วน Other Attribute
3. กดปุ่ม Next  เพื่อเข้าสู่หน้าจอ Validation ดังรูปที่ 5.11
4. กดปุ่ม previous  เพื่อกลับเข้าสู่หน้าจอ Define Variable

5.2.4 การกำหนดส่วนแบ่งข้อมูลและกฎในการสร้างแบบจำลองต้นไม้

หน้าจอ Validation เป็นหน้าจอที่ใช้ในการกำหนดส่วนแบ่งของข้อมูล ในการนำไปสร้างแบบจำลอง (Training Data) และข้อมูลที่จะนำมาใช้ในการวิเคราะห์ความถูกต้องของแบบจำลอง (Testing Data) โดยในหน้าจอนี้จะแบ่งออกเป็น 2 ส่วนดังนี้

- Partition เป็นส่วนที่ใช้ในการแบ่งข้อมูล ถ้าไม่ต้องการแบ่งข้อมูล (Testing Data) เพื่อทำการทดสอบความเชื่อมั่นของโมเดลให้เลือกที่ Radio Button : Do not validate the tree ถ้าต้องการแบ่งข้อมูลส่วนหนึ่งไว้วิเคราะห์ความถูกต้องของแบบจำลองให้เลือกที่ Radio Button : Partition my data into sub samples และทำการกำหนดจำนวนเปอร์เซ็นต์ของข้อมูลที่เราต้องการแบ่งไว้
- Stopping Rule เป็นส่วนที่ใช้ในการกำหนดการสร้างแบบจำลองต้นไม้ดังนี้
  - Maximum Level : เป็นการกำหนดระดับสูงสุดในการแตกแบบจำลองต้นไม้ เช่น ถ้าเราใส่ Maximum Level เท่ากับ 3 แสดงว่าแบบจำลองต้นไม้ที่ได้ในการวิเคราะห์นั้นจะไม่เกิน 3 ระดับ ซึ่งรวมรูตโหนดด้วย
  - Minimum number of data in each node : กำหนดจำนวนข้อมูลต่ำสุดในแต่ละโหนด

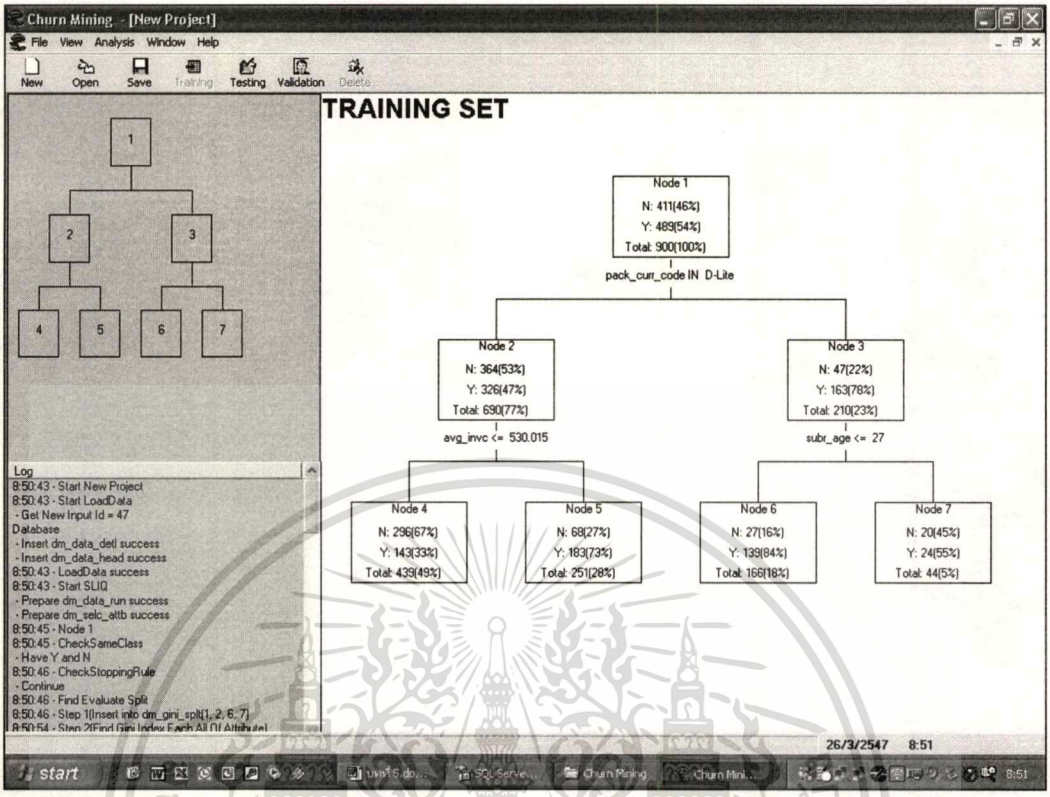


รูปที่ 5.11 หน้าจอ Validation

5.2.5 การวิเคราะห์ข้อมูล

เมื่อทำการใส่ข้อมูลต่างๆ ตามที่กล่าวมาข้างต้นจนครบให้กดปุ่ม Finish

โปรแกรมจะทำการแสดงผลข้อมูลออกมาในรูปแบบจำลองต้นไม้ดังรูปที่ 5.12



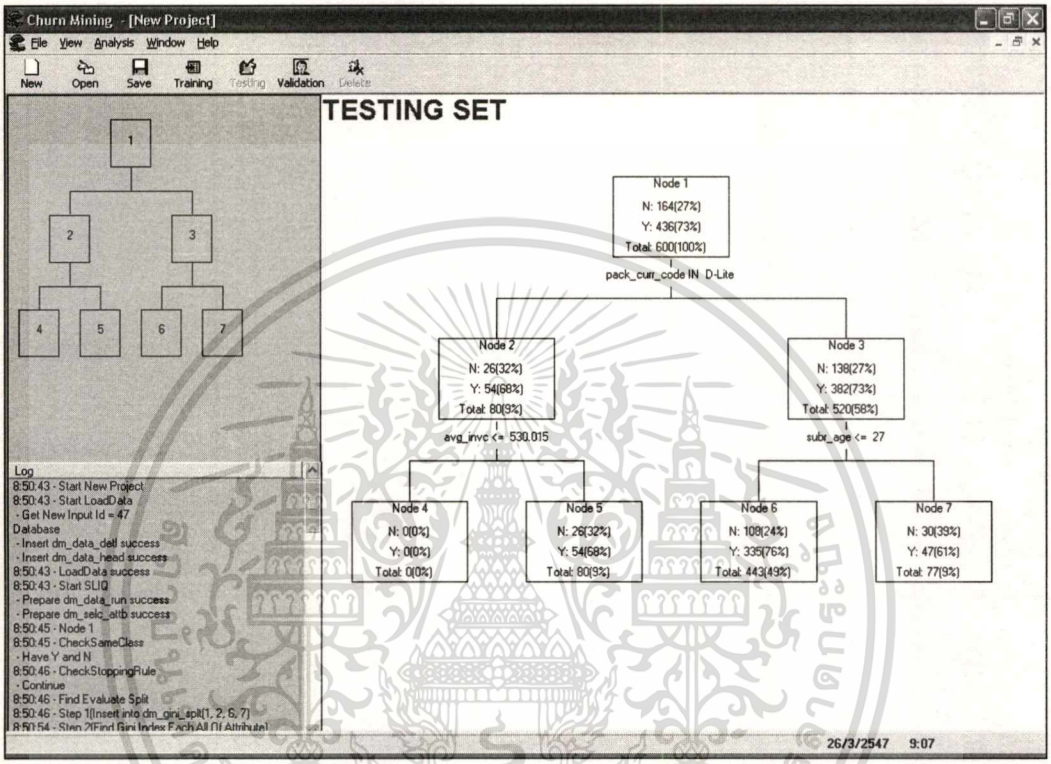
รูปที่ 5.12 แบบจำลองต้นไม้

จากหน้าจอจะแบ่งออกเป็น 3 ส่วนดังนี้

- ส่วนแสดงแบบจำลองต้นไม้รูปเล็ก จะใช้ในกรณีที่แบบจำลองต้นไม้มีขนาดใหญ่เกินหน้าจอที่แสดงไว้ ผู้ใช้สามารถที่จะคลิกไปที่หมายเลขโหนดในแบบจำลองรูปเล็กและระบบก็ทำการเลื่อนภาพไปที่โหนดที่ผู้ใช้ต้องการโดยอัตโนมัติ
- ส่วน Log จะใช้ในการแสดงรายละเอียดขั้นตอนการทำงานของโปรแกรมรวมทั้งบอกถึงข้อผิดพลาดในกรณีระบบไม่สามารถวิเคราะห์ผลได้
- ส่วนแสดงแบบจำลองต้นไม้ ในส่วนนี้จะแสดงรายละเอียดของโหนดในแต่ละโหนดว่ามีข้อมูลอยู่เท่าไร และแสดงค่าความเชื่อมั่น ซึ่งในหน้าจอนี้จะแสดงข้อมูล 2 ส่วนคือในส่วนของ Training Set และ Testing Set

5.2.6 การแสดงข้อมูลในส่วนของ Testing Set

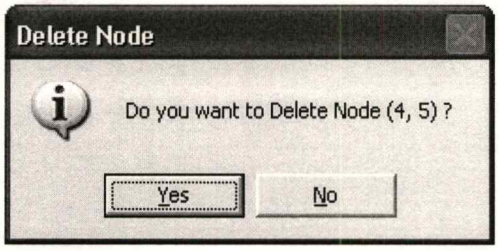
ในการดูข้อมูลในส่วนของ Testing Set นั้นให้ผู้ใช้เข้าไปที่เมนู View -> Testing Data หรือจะคลิกที่ตรงไอคอน Testing  ระบบจะนำข้อมูลในส่วนที่เป็น Testing มาวิเคราะห์ให้โดยใช้เงื่อนไขแบบจำลองต้นไม้เดิมซึ่งจะได้ผลดังรูปที่ 5.13



รูปที่ 5.13 Testing data

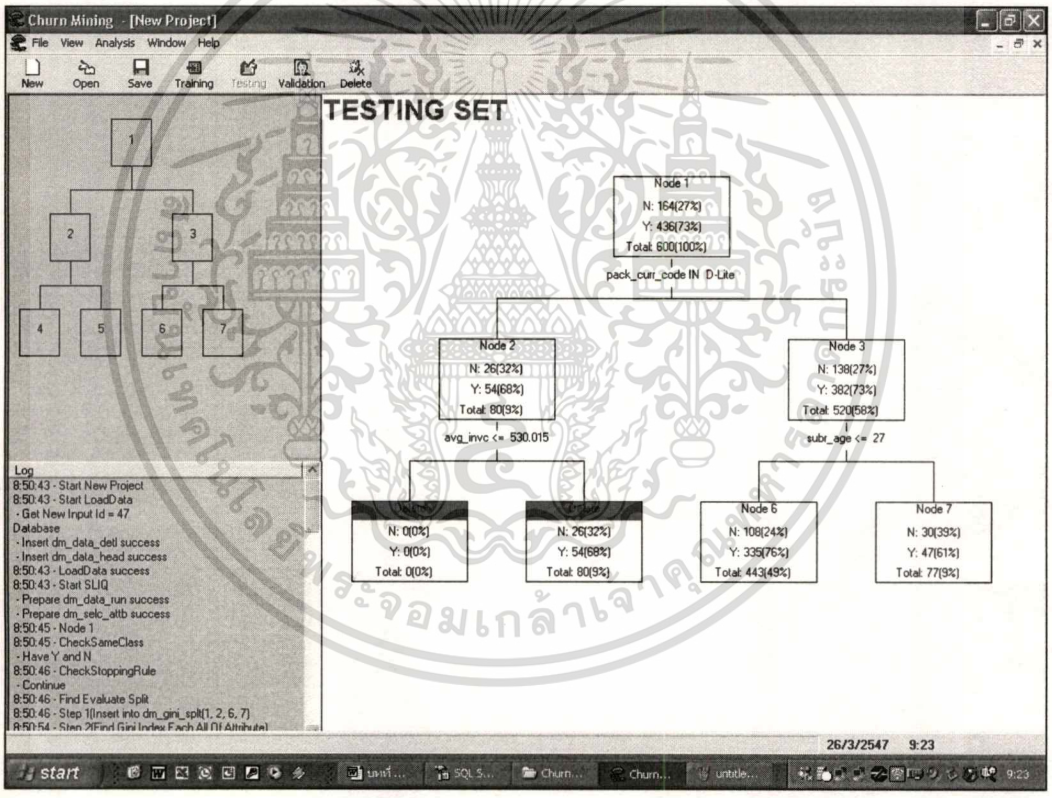
5.2.7 การปรับแต่งแบบจำลองต้นไม้

เมื่อผู้ใช้ระบบทำการวิเคราะห์แล้วว่าโหนดบางโหนดนั้นไม่มีผลต่อการวิเคราะห์ ผู้ใช้ระบบสามารถทำการปรับแต่งแบบจำลองต้นไม้เองได้ โดยให้ทำการคลิกไปที่หมายเลขโหนดในแบบจำลองต้นไม้จะพบกล่องข้อความดังรูปที่ 5.14



รูปที่ 5.14 กล่องข้อความ Delete node

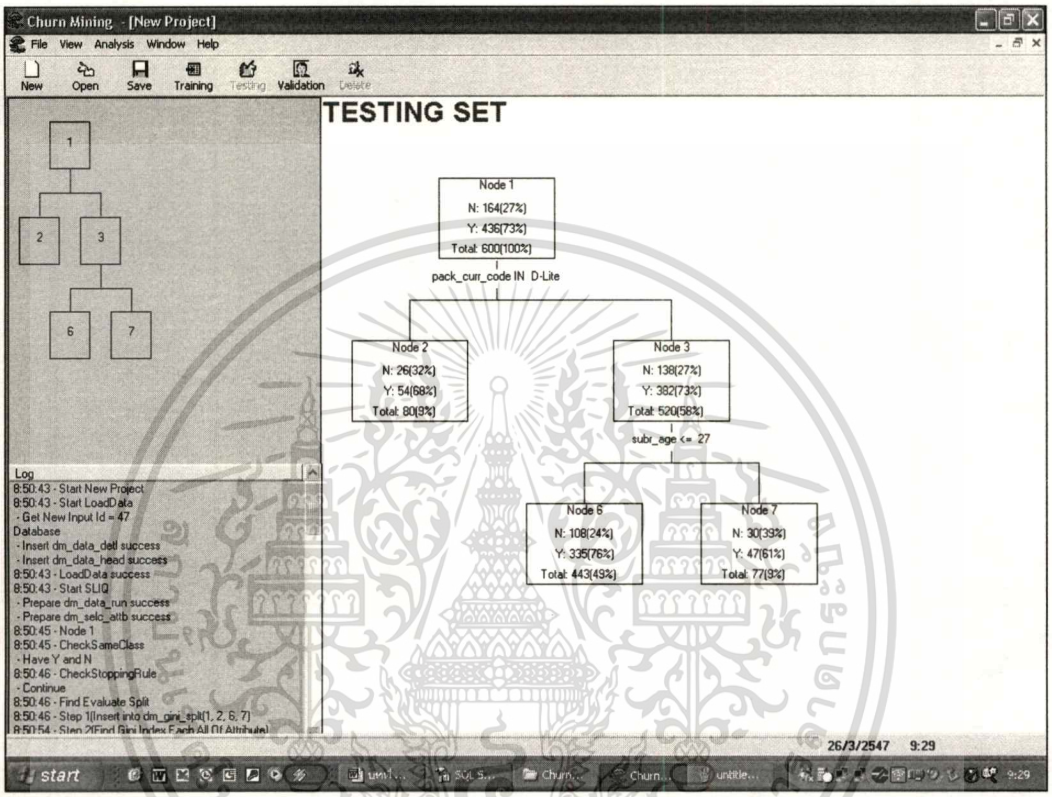
ถ้าต้องการลบโหนดให้กดปุ่ม Yes ถ้าไม่ต้องการลบให้กดปุ่ม No ถ้าในกรณีกดปุ่ม Yes จะพบหน้าจอดังรูป 5.15



รูปที่ 5.15 หน้าจอ Delete Node

ระบบจะแสดงโหนดที่ผู้ใช้ต้องการจะลบโดยแสดงคำว่า Delete ในโหนดนั้นซึ่ง ถ้าผู้ใช้ยืนยันที่จะลบ ให้ทำการคลิกที่ไอคอน Delete 

ระบบจะทำการลบโหนดนั้นทิ้งไปซึ่งผลจะแสดงดังรูป 5.16



รูปที่ 5.16 หน้าจอแสดงการลบโหนด

5.2.8 การออกจากระบบ

ในการออกจากระบบนั้นสามารถทำได้ 2 ทางดังนี้

- ไปที่เมนู File -> Exit
- กดปุ่ม Close ที่หน้าจอ

บทที่ 6
บทสรุปและข้อเสนอแนะ

จากการที่ได้ทำการศึกษาและพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีนั้น ทำให้ได้ข้อสรุปดังต่อไปนี้

6.1 สรุปผลการพัฒนา

จากการพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีพบว่าระบบสามารถทำงานได้ตามวัตถุประสงค์ คือ

1. ระบบสามารถช่วยในการวิเคราะห์ดิซิชั่นทรีได้อย่างถูกต้อง ตรงตามแนวคิดของอัลกอริทึม SLIQ
2. ระบบสามารถแสดงค่าความเชื่อมั่นเพื่อเป็นส่วนช่วยในการตัดสินใจให้กับผู้ใช้ระบบได้
3. ระบบสามารถนำไปใช้ในการวิเคราะห์ข้อมูลเรื่องใดก็ได้ ไม่จำเป็นต้องใช้ในการวิเคราะห์เรื่อง Churn Management เท่านั้น แต่ต้องอยู่ภายใต้แนวคิดดิซิชั่นทรีและอัลกอริทึม SLIQ
4. ระบบสามารถแสดงผลการวิเคราะห์ให้อยู่ในรูปแบบที่ง่ายต่อความเข้าใจของผู้ใช้ระบบ โดยแสดงเป็นรูปแบบจำลองต้นไม้ (Tree Model)

6.2 ประโยชน์ที่ได้รับ

จากการพัฒนาระบบวิเคราะห์สาเหตุการเปลี่ยนแปลงผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้าโดยใช้ดิซิชั่นทรีนั้นทำให้ผู้พัฒนาระบบได้รับประโยชน์ดังนี้

- ทำให้ผู้พัฒนาระบบมีความรู้ความเข้าในทฤษฎีดาต้าไมนิ่งเพิ่มมากขึ้น เนื่องจากได้นำความรู้นั้นมาพัฒนาระบบและทำให้ผู้พัฒนาระบบเห็นว่าทฤษฎีดาต้าไมนิ่งนั้นมีประโยชน์และสามารถนำไปช่วยในการตัดสินใจได้จริง
- ทำให้ผู้พัฒนาระบบได้ศึกษาแนวคิดของอัลกอริทึม SLIQ
- ทำให้ผู้พัฒนาระบบทราบถึงวิธีการบริหารโครงงาน ทำให้สามารถนำไปประยุกต์ใช้ในการทำงานได้เป็นอย่างดี

- ทำให้ผู้พัฒนาระบบได้เรียนรู้การพัฒนาโปรแกรมด้วยภาษา Visual Basic มากยิ่งขึ้น
- นอกจากนี้ระบบที่ได้พัฒนาเสร็จสิ้นแล้ว สามารถนำไปใช้ในในองค์กรเพื่อช่วยในการวิเคราะห์ข้อมูลลูกค้าได้ ซึ่งระบบนี้สามารถนำไปวิเคราะห์กับเรื่องอะไรก็ได้ไม่จำเป็นต้องเกี่ยวกับ Churn Management เท่านั้น

6.3 ข้อเสนอแนะ

1. ในการวิเคราะห์ข้อมูลอะไรก็ตามสิ่งที่สำคัญที่สุดก็คือข้อมูลที่ผู้ใช้ระบบนำมาใช้ ดังนั้นข้อมูลต่าง ๆ นั้นต้องผ่านขั้นตอนเตรียมข้อมูลมาอย่างถูกต้อง เพื่อที่จะส่งผลให้ผลการวิเคราะห์นั้นถูกต้องด้วย ซึ่งในระบบนี้ยังขาดขั้นตอนในการกรองข้อมูล
2. ข้อมูลที่ผู้พัฒนาระบบนำมาใช้ในการวิเคราะห์นั้นถือว่ามีจำนวนน้อยเกินไปเมื่อเทียบกับข้อมูลจริงที่อยู่ในองค์กร แต่เนื่องจากผู้พัฒนาระบบสามารถนำข้อมูลออกมาได้เพียงเท่านี้ จึงอาจทำให้ผลของการวิเคราะห์ข้อมูลจากฐานข้อมูลที่เตรียมไว้นั้นไม่ถูกต้องเท่าที่ควร
3. ระบบที่เกี่ยวกับการวิเคราะห์ข้อมูลนั้นควรจะสามารถมองมุมมองของรูปแบบจำลองต้นไม้ (Tree Model) ได้หลายมุมมองและควรที่จะคำนวณค่าทางสถิติและแสดงรายงานต่างๆ เพื่อใช้ประกอบการตัดสินใจ ซึ่งยังขาดการพัฒนาในจุดนี้
4. ระบบที่ใช้ในการวิเคราะห์ ดาต้าไมนิ่งนั้นมีอัลกอริทึมอยู่หลายอัลกอริทึมที่สามารถนำมาใช้ในการวิเคราะห์ แต่ในระบบนี้ใช้แค่อัลกอริทึม SLIQ เท่านั้นซึ่งถ้าเป็นไปได้ก็ควรจะมีอัลกอริทึมที่หลากหลายมากกว่านี้

บรรณานุกรม

Berson Alex et al. 1999. **"Building Data Mining Applications for CRM"**

Cabena Peter et al. **"Discovery Data Mining From Concept to Implement"**. :Prentice Hall PTR,Upper Saddle River,New Jersey 07458.

Han, Jiawei and Kanber, Micheline. 2001. **"Data Mining Concepts and Techniques."** San Fancisco :Morgan Kaufmann Publishers.

Manish Mehta et al. 2001.**SLIQ: A Fast Scalable Classifier for Data Mining** [Online].Availble: http://www.cs.yorku.ca/~jarek/courses/6421/sliq.pdf

Rajeev Rastogi et al. 1998: **PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning** [Online]. Availble : http://citeseer.nj.nec.com/correct/386494

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวดวงหทัย วงศ์สวัสดิ์สุริยะ
สถานที่เกิด	จังหวัดกรุงเทพ
ระดับประถมศึกษา	โรงเรียนเซนต์โยเซฟ บางนา
ระดับมัธยมศึกษาตอนต้น	โรงเรียนเตรียมอุดมศึกษาพัฒนาการ
ระดับมัธยมศึกษาตอนปลาย	โรงเรียนเตรียมอุดมศึกษาพัฒนาการ
ระดับอุดมศึกษา	คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
วุฒิการศึกษาระดับปริญญาตรี	สถิติศาสตร์บัณฑิต (สตบ)
ประสบการณ์การทำงาน	ศูนย์การศึกษาต่อเนื่องแห่งจุฬาลงกรณ์มหาวิทยาลัย บมจ. โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น