# การใช้ดาต้าไมนิ่งเพื่อการทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่

# DATA MINING FOR CHURN PREDICTION ANALYSIS OF MOBILE OPERATOR

โดย

สุภาภรณ์ ศิริติกุล

SUPAPORN SIRITIKUL



อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช



รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการศึกษากรณีพิเศษ

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 1 ปีการศึกษา 2549

# DATA MINING FOR CHURN PREDICTION ANALYSIS OF MOBILE OPERATOR

**SUPAPORN SIRITIKUL**

**A SPECIAL STUDY 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** | การใช้ดาต้าไมนิ่งเพื่อการทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ |
| **นักศึกษา** | นางสาวสุภาภรณ์ ศิริติกุล |
| **รหัสนักศึกษา** | 47066713 |
| **ปริญญา** | วิทยาศาสตรมหาบัณฑิต |
| **สาขาวิชา** | เทคโนโลยีสารสนเทศ |
| **แขนงวิชา** | การจัดการเทคโนโลยีสารสนเทศ |
| **ปีการศึกษา** | 2549 |
| **อาจารย์ที่ปรึกษา** | รศ.ดร.วรพจน์ กรีสุระเดช |

## บทคัดย่อ

ในปัจจุบันการดำเนินธุรกิจมีการแข่งขันกันอย่างสูง ส่งผลให้ผู้ประกอบการและองค์กรธุรกิจต่างมุ่งหากลยุทธ์ใหม่ๆเพื่อเพิ่มศักยภาพองค์กรของตนเองให้สามารถแข่งขันกับองค์กรอื่นๆได้ ซึ่งในโครงการศึกษาฉบับนี้ได้นำเสนอเกี่ยวกับการนำเทคนิคของดาต้าไมนิ่งมาประยุกต์ใช้ เพื่อทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ โดยใช้เทคนิคของ Predictive Model โดยใช้โปรแกรมสำเร็จรูปทางด้านดาต้าไมนิ่งของบริษัท SPSS ที่มีชื่อว่า Clementine มาเป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูลเกี่ยวกับพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้า และนำผลลัพธ์ที่ได้จากการทำนาย ซึ่งอยู่ในรูปแบบของโมเดลไปใช้ประโยชน์ เพื่อเป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด เพื่อตอบสนองความต้องการของลูกค้า ทำให้ลูกค้าเกิดความพึงพอใจสูงสุด ทำให้องค์กรสามารถรักษาฐานลูกค้าให้ยังคงอยู่กับองค์กรต่อไป

| | |
|---|---|
| **Title** | Data Mining for Churn Prediction Analysis of Mobile Operator |
| **Student** | Miss Supaporn Siritikul |
| **Student ID.** | 47066713 |
| **Degree** | Master of Science |
| **Programme** | Information Technology Management |
| **Academic Year** | 2006 |
| **Advisor** | Assoc. Prof. Dr. Worapoj Kreesuradej |


## ABSTRACT

Presently, there is high competition in business. Each company try to use new strategy to increase their potential that can be compete with other company. This project presented about using data mining technique for churn prediction of mobile operator that use Predictive Model technique and use Clementine Program as tool for analyze usage behavior of each customers and use information from this result to plan marketing strategies for built customer satisfaction that can improve customer retention and customer loyalty with organization.

# กิตติกรรมประกาศ

โครงการศึกษากรณีพิเศษครั้งนี้สำเร็จลงได้ ข้าพเจ้าต้องกราบขอบพระคุณ รศ.ดร. วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ที่ปรึกษาโครงการเป็นอย่างสูง ที่ได้ให้ความอนุเคราะห์ต่อข้าพเจ้าเกี่ยวกับความรู้และคำแนะนำที่ดีและเป็นประโยชน์ต่อการทำโครงการ จนทำให้โครงการศึกษาในครั้งนี้สำเร็จลุล่วงไปด้วยดี

ขอกราบขอบพระคุณคณาจารย์ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า ทำให้ข้าพเจ้าสามารถนำเอาความรู้ที่ได้จากการศึกษามาประยุกต์ใช้ในโครงการศึกษานี้

ขอขอบพระคุณบริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) ที่ให้การสนับสนุนในส่วนของข้อมูล ที่ใช้เป็นตัวอย่างในการศึกษาของโครงการนี้

ขอขอบคุณเพื่อนๆ พี่ๆ และน้องๆ แขนงวิชาการจัดการเทคโนโลยีสารสนเทศ (ITM) รุ่น16 คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆ คนที่คอยให้คำแนะนำและความช่วยเหลือต่างๆ แก่ข้าพเจ้า

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา ของข้าพเจ้า ผู้ซึ่งเป็นกำลังใจที่ดีที่สุดที่ทำให้ข้าพเจ้ามีวันนี้ด้วยความภาคภูมิใจ

คุณค่าและประโยชน์อันพึงมาจากรายงานฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

สุภาภรณ์ ศิริติกุล

# สารบัญ

หน้า

# สารบัญ (ต่อ)

หน้า

# สารบัญตาราง

# สารบัญรูป

# สารบัญรูป (ต่อ)

รูปที่ หน้า

# บทที่1

# บทนำ

ในบทนี้จะกล่าวถึงความเป็นมาของปัญหาในการนำเอาเทคนิคของดาต้าไมนิ่ง เข้ามาช่วยในการค้นหาสารสนเทศที่เป็นประโยชน์ต่อองค์กร ซึ่งซ่อนอยู่ในฐานข้อมูลขนาดใหญ่, วัตถุประสงค์ของโครงการ, ขอบเขตของการศึกษา และขั้นตอนการดำเนินงานของโครงการ รวมไปถึงประโยชน์ที่จะได้รับจากการพัฒนาโครงการนี้

## 1.1 ความเป็นมาและความสำคัญของปัญหา

การดำเนินธุรกิจในปัจจุบันไม่ว่าจะเป็นธุรกิจ หรืออุตสาหกรรมใดก็ตาม ต่างก็มีการแข่งขันกันอย่างสูง ทั้งในด้านสินค้าและบริการ ทำให้นักการตลาดและผู้บริหารของแต่ละธุรกิจจำเป็นต้องกำหนดนโยบายและแผนกลยุทธ์ของตนให้มีประสิทธิภาพมากที่สุด โดยอาศัยทรัพยากรที่สำคัญที่องค์กรมีอยู่ นั่นก็คือ ฐานข้อมูลลูกค้า

ธุรกิจโทรคมนาคมในประเทศไทย โดยเฉพาะธุรกิจที่เกี่ยวกับผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ นับเป็นอีกธุรกิจหนึ่งที่มีการแข่งขันกันอย่างรุนแรง โดยผู้ให้บริการแต่ละรายต่างแข่งขันกันสร้างสรรค์สินค้า และบริการให้มีความแตกต่างจากคู่แข่ง เพื่อที่จะสามารถตอบสนองและเข้าถึงความต้องการของลูกค้าให้ได้มากที่สุด ตรงกลุ่มเป้าหมายที่สุด และอยู่ในช่วงเวลาและโอกาสที่เหมาะสมที่สุด จึงจะเป็นผู้ครองส่วนแบ่งทางการตลาดในธุรกิจนั้น ยิ่งไปกว่านั้นการใช้ข้อมูลที่มีอยู่ให้เกิดประโยชน์สูงสุด ยังเท่ากับเป็นการลดต้นทุนทางการแข่งขัน และสามารถสร้างผลกำไรให้กับธุรกิจได้อีกด้วย

โดยในโครงการศึกษาฉบับนี้ ได้นำเสนอแนวทางการวิเ คราะห์ข้อมูลในเชิงธุรกิจ โดยใช้วิธีการสืบค้นข้อมูลจากฐานข้อมูล (Knowledge Discovery in Database: KDD) หรือที่เรียกว่าการขุดค้นข้อมูล (Data Mining) เพื่อใช้วิเคราะห์พฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้า และทำนายได้ว่าลูกค้ารายใดที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ในอนาคต โดยนำผลลัพธ์ที่ได้จากการทำนาย ซึ่งเป็นสารสนเทศที่มีประโยชน์ สามารถนำไปใช้เป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด เพื่อตอบสนองความต้องการของลูกค้า ทำให้ลูกค้าเกิดความพึงพอใจสูงสุด และยังเป็นการป้องกันการยกเลิกของลูกค้าปัจจุบัน และลดการสูญเสียลูกค้าให้กับบริษัทคู่แข่ง ทำให้การรักษาฐานลูกค้า (Customer Retention) ให้อยู่กับองค์กร มีประสิทธิภาพมากยิ่งขึ้น

ส่งผลให้ยอดการยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ลดลง ซึ่งเป็นการเพิ่มรายได้ให้แก่บริษัท เพราะทำให้บริษัทไม่สูญเสียรายได้จากลูกค้ากลุ่มนี้

## 1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

การนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ วัตถุประสงค์เพื่อให้องค์กรสามารถนำเอาสารสนเทศที่ได้นี้มาใช้ประกอบเพื่อเป็นแนวทางในการปรับปรุงกลยุทธ์ทางการตลาด เพื่อเพิ่มประสิทธิภาพในการรักษาฐานลูกค้าในปัจจุบัน

## 1.3 ขอบเขตการศึกษา

โครงการนี้เป็นการศึกษาเทคนิคของดาต้าไมนิ่ง เพื่อมาประยุกต์ใช้กับปัญหาทางธุรกิจ โดยอาศัยหลักการของ Predictive Model ในการพยากรณ์ข้อมูลในฐานข้อมูลลูกค้าเพื่อการทำนายว่าลูกค้ารายใดที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ โดยในโครงการนี้ใช้ข้อมูลลูกค้า ซึ่งเป็นผู้ใช้บริการโทรศัพท์เคลื่อนที่แบบจ่ายรายเดือน (Postpaid) ของบริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) ซึ่งเป็นผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ระบบ DTAC มาเป็นฐานข้อมูลในการศึกษา ซึ่งจะนำโปรแกรมสำเร็จรูปทางด้านดาต้าไมนิ่งที่มีชื่อว่า Clementine มาเป็นเครื่องมือช่วยในการวิเคราะห์ และค้นหาความสัมพันธ์ของข้อมูล

## 1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1) ศึกษาแนวคิดและทฤษฎีเบื้องต้นที่เกี่ยวข้องกับเทคนิคดาต้าไมนิ่ง เพื่อนำมาประยุกต์ใช้ในการพัฒนาโครงการ
2) ศึกษาทฤษฎี Predictive Model และอัลกอริทึมต่างๆ ที่เกี่ยวข้อง เพื่อนำมาประยุกต์ใช้ในการศึกษาโครงการ
3) เก็บรวบรวมข้อมูลที่เกี่ยวข้องในการทำโครงการ
4) วิเคราะห์และประมวลผลข้อมูลเพื่อทำนายผลลัพธ์ของลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่

5) สรุปผลการศึกษาของโครงการ

## 1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการที่ได้ศึกษาแนวคิดและทฤษฎีของดาต้าไมนิ่ง เพื่อมาทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ คาดว่าจะได้รับประโยชน์ ดังนี้

1) ทำให้มีความรู้ความเข้าใจถึงหลักการ และกระบวนการทำงานของดาต้าไมนิ่ง
2) ทำให้เข้าใจระบบการทำงานของ Predictive Model และอัลกอริทึมต่างๆที่เกี่ยวข้องมากยิ่งขึ้น
3) สามารถนำความรู้ที่ได้ไปประยุกต์ใช้กับปัญหาจริงทางธุรกิจขององค์กรได้ ทำให้สามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ในอนาคตได้อย่างถูกต้องแม่นยำมากขึ้น และสามารถนำสารสนเทศที่ได้ไปปรับปรุงกลยุทธ์ทางการตลาดเพื่อตอบสนองความต้องการของลูกค้า ทำให้ลูกค้าเกิดความพึงพอใจสูงสุด ทำให้สามารถรักษาฐานลูกค้าให้ยังคงอยู่กับองค์กรต่อไป

# บทที่ 2
# ทฤษฎีและหลักการของดาต้าไมนิ่ง

## 2.1 ความหมายของดาต้าไมนิ่ง

ในอดีตการจะค้นหาข้อมูลที่มีประโยชน์จากฐานข้อมูลนั้นเป็นเรื่องยาก ยิ่งถ้าหากเป็นฐานข้อมูลที่มีขนาดใหญ่มากๆ ก็จะยิ่งต้องใช้เวลาในการค้นหานานมาก จึงทำให้นักพัฒนาระบบต่างคิดค้นวิธีการที่จะทำให้สามารถค้นหาข้อมูลสารสนเทศที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ตลอดจนความสัมพันธ์กันของปัจจัยต่างๆ เพื่อนำมาใช้ประโยชน์ในการวิเคราะห์ การพยากรณ์ที่แม่นยำถูกต้อง ซึ่งสามารถใช้ประโยชน์ในการกำหนดแนวทางหรือแผนในการปฏิบัติงานขององค์กรนั้นให้มีประสิทธิภาพมากที่สุด

ดังนั้น การที่เราจะค้นหาข้อมูลที่เป็นสารสนเทศที่เราต้องการจากแหล่งข้อมูลดิบที่มีมากมายมหาศาลนั้น เราจำเป็นต้องมีเครื่องมือที่จะช่วยในการค้นหาสารสนเทศเหล่านั้น ซึ่งหนึ่งในนั้นก็คือ เทคนิคของดาต้าไมนิ่ง

โดยนิยามของดาต้าไมนิ่ง (Data Mining) หมายถึง กระบวนการในการค้นหาเอาข้อมูลสารสนเทศที่ซ่อนอยู่ภายใต้ฐานข้อมูลที่มีอยู่จำนวนมากมาย ซึ่งเก็บอยู่ในระบบฐานข้อมูลขององค์กรออกมา โดยใช้กระบวนการต่างๆในการค้นหาข้อมูลออกมาจากฐานข้อมูล แล้วนำมาตั้งเป็นสมมติฐาน หลังจากนั้นก็นำข้อมูลที่ต้องการทราบ มาทำการทดสอบสมมติฐานที่สร้างไว้ ซึ่งสารสนเทศที่ได้ออกมานั้น ต้องมีลักษณะดังนี้คือ

1) เป็นข้อมูลที่ไม่เคยรู้ล่วงหน้ามาก่อน (Unknown) หมายถึง ข้อมูลสารสนเทศที่ได้รับนั้น ต้องไม่เคยค้นพบมาก่อนหน้า และไม่สามารถคาดเดาได้ว่าผลที่ไดรับจะออกมาในลักษณะใด
2) ต้องเป็นข้อมูลที่มีความถูกต้อง (Valid) หมายถึง สารสนเทศที่ได้รับต้องเป็นสารสนเทศที่มีความถูกต้อง เนื่องจากต้องนำไปใช้ประกอบกับข้อมูลส่วนอื่นๆ ดังนั้นต้องมีความถูกต้อง น่าเชื่อถือ
3) สามารถนำไปใช้ประโยชน์ได้ (Actionable) คือ ต้องสามารถนำเอาข้อมูลและสารสนเทศที่ค้นพบออกมา ไปใช้ประโยชน์ในด้านอื่นๆได้ เช่น นำมาช่วยตัดสินใจในการวางแผนการตลาด เพื่อสร้างความได้เปรียบทางการแข่งขันในเชิงธุรกิจ เป็นต้น

ดังนั้นการทำดาต้าไมนิ่งจึงเปรียบเสมือนการขุดหาแร่จากเหมืองแร่ที่มีขนาดใหญ่ กว่าที่จะได้แร่ที่มีค่าอย่างที่ต้องการนั้นต้องผ่านกระบวนการมากมายหลายขั้นตอน ในการขุดค้น กลั่นกรอง เพื่อที่จะได้แร่ที่มีค่าออกมา นั่นจึงเป็นที่มาของคำว่า ดาต้าไมนิ่ง หรือการทำเหมืองข้อมูล

## 2.2 วิวัฒนาการของเทคโนโลยีฐานข้อมูล

วิวัฒนาการของเทคโนโลยีด้านฐานข้อมูลนั้น ได้มีการพัฒนามาทุกยุคทุกสมัย ตั้งแต่ในอดีตจนถึงปัจจุบัน ซึ่งเป็นเทคโนโลยีที่มีความสำคัญมาก เนื่องจากข้อมูลเป็นสิ่งสำคัญในการนำมาใช้ประโยชน์ในด้านต่างๆ อีกทั้งแนวโน้มของการเพิ่มขึ้นของข้อมูล ก็มีแนวโน้มที่เพิ่มขึ้นสูงมาก จึงได้มีการพัฒนาและปรับปรุงวิธีการต่างๆ เพื่อที่จะสามารถเก็บรวบรวม และประมวลผลข้อมูลที่มีอยู่อย่างมหาศาลได้อย่างมีประสิทธิภาพ โดยสามารถสรุปวิวัฒนาการของการพัฒนาเทคโนโลยีด้านฐานข้อมูลได้เป็นช่วงเวลา ดังนี้

ช่วงปี ค.ศ. 1960 เทคโนโลยีฐานข้อมูลได้เริ่มพัฒนามาจากระบบ File processing พื้นฐาน จากนั้นจึงมีการค้นคว้าและพัฒนาระบบฐานข้อมูลมาเรื่อย ๆ เป็นระบบการเก็บข้อมูล, การสร้างฐานข้อมูล (Database), ระบบ IMS และระบบเครือข่าย DBMS

ช่วงปี ค.ศ. 1970 ได้นำไปสู่การพัฒนาระบบการเก็บข้อมูลในรูปแบบตาราง (Relational Database System) โดยมีการสร้างเครื่องมือต่างๆ ที่ช่วยอำนวยความสะดวกในการจัดการกับข้อมูล อีกทั้งยังมีการคิดค้นภาษาที่ใช้ในการเรียกดูข้อมูล (Query Language) เพื่ออำนวยความสะดวกในการเข้าถึงข้อมูลในฐานข้อมูล

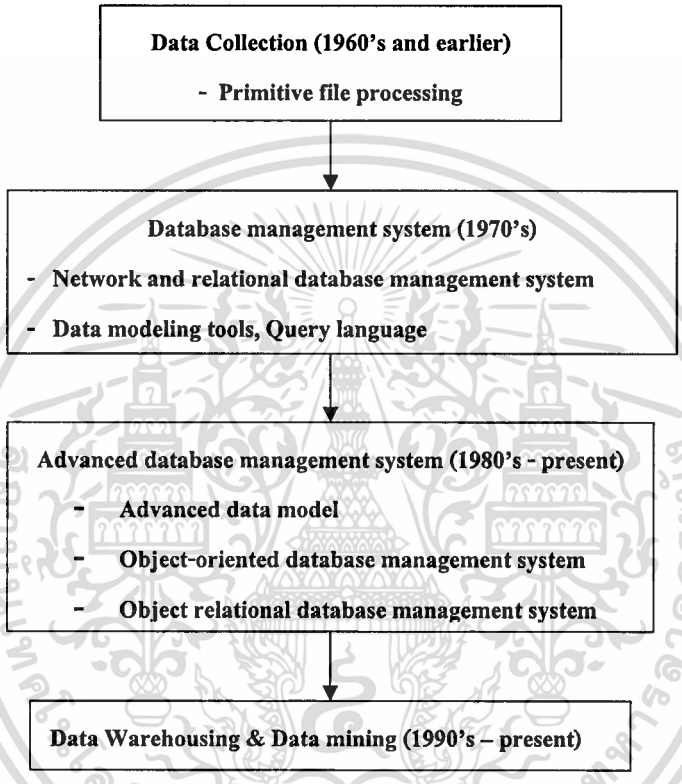
ช่วงปี ค.ศ. 1980 เทคโนโลยีฐานข้อมูลได้เริ่มมีการปรับปรุงและพัฒนาระบบจัดการฐานข้อมูลที่มีศักยภาพมากขึ้น ทำให้สามารถจัดเก็บข้อมูลจำนวนมากที่มีความซับซ้อนได้อย่างมีประสิทธิภาพเพิ่มขึ้น เกิดระบบการจัดการฐานข้อมูลที่มีประสิทธิภาพ เช่น Object-Oriented Database Management System, Object Relational Database Management System เป็นต้น

ช่วงปี ค.ศ. 1990 จนถึงยุคปัจจุบัน สามารถจัดเก็บข้อมูลได้ในหลายรูปแบบ แตกต่างกันทั้งระบบปฏิบัติการ หรือการจัดเก็บฐานข้อมูล ซึ่งเป็นการนำข้อมูลทั้งหมดมารวมและจัดเก็บไว้ในรูปแบบเดียวกันเรียกว่า ดาต้าแวร์เฮาส์ (Data Warehouse) เพื่อเพิ่มความสะดวกในการบริหารจัดการข้อมูล ซึ่งเทคโนโลยี Data Warehouse จะรวมไปถึงการทำ Data Cleansing, Data Integration และ On-Line Analytical Processing (OLAP) ซึ่งเป็นเทคนิคในการวิเคราะห์ข้อมูลในหลายๆ มิตินั้น ได้เกิดขึ้นมาตามลำดับ

การละเลยข้อมูลควบคู่ไปกับการขาดเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลที่มีศักยภาพนำไปสู่สถานการณ์ที่ว่า "ข้อมูลมากแต่ความรู้น้อย" (data rich but information poor) การเติบโตขึ้นอย่างรวดเร็วของข้อมูลจำนวนมาก ที่สะสมไว้ในฐานข้อมูลขนาดใหญ่มาก ซึ่งเกินกว่าที่กำลังคนจะ

สามารถจัดการได้ เป็นผลทำให้มีความจำเป็นที่ต้องมีเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลและหาความเป็นไปได้ของข้อมูลที่เป็นประโยชน์ที่อยู่ในฐานข้อมูลออกมา ซึ่งก็คือ ดาต้าไมนิ่ง

จากที่ได้กล่าวถึงประวัติความเป็นมาและวิวัฒนาการของเทคโนโลยีฐานข้อมูลตั้งแต่ในยุคอดีตจนถึงปัจจุบัน จะสามารถแสดงได้ดังรูปที่ 2.1




**รูปที่ 2.1** แสดงวิวัฒนาการของเทคโนโลยีฐานข้อมูล

## 2.3 เหตุผลของการทำดาต้าไมนิ่ง

1. ข้อมูลที่ถูกเก็บไว้ในฐานข้อมูล หากเก็บไว้เฉย ๆ ก็จะไม่เกิดประโยชน์ดังนั้นจึงต้องมีการสกัดเอาสารสนเทศที่มีประโยชน์ออกมาจากฐานข้อมูล และนำไปใช้

2. ในอดีตนั้นเราใช้คนเป็นผู้สืบค้นข้อมูลต่างๆ จากฐานข้อมูล ซึ่งผู้สืบค้นจะทำการสร้างเงื่อนไขขึ้นมาตามภูมิปัญญาของผู้สืบค้นเอง อาจทำให้ได้ข้อมูลไม่ครบถ้วน

3. ในปัจจุบันการวิเคราะห์ข้อมูลจากฐานข้อมูลเดียวอาจไม่ให้ความรู้เพียงพอและลึกซึ้งสำหรับการดำเนินงานภายใต้สถานการณ์ที่มีการแข่งขันสูงและมีการเปลี่ยนแปลงที่รวดเร็วจึง

จำเป็นที่จะต้องรวบรวมข้อมูลจากฐานข้อมูลหลายๆ ฐานข้อมูลเข้าด้วยกันซึ่งเรียกว่าคลังข้อมูล (Data Warehouse)

ดังนั้นเราจึงจำเป็นต้องใช้เทคนิคของดาต้าไมนิ่ง ในการดึงข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่ เพื่อที่จะนำข้อมูลเหล่านั้นมาใช้งานให้เกิดประโยชน์สูงที่สุด

## 2.4 ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ได้รับความนิยม

ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ได้รับความนิยม คือ

1. จำนวนและขนาดข้อมูลขนาดใหญ่ถูกผลิต และขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทางอินเทอร์เน็ต,ดาวเทียม และแหล่งผลิตข้อมูล อื่น ๆ เช่น เครื่องอ่านบาร์โค้ด, เครดิตการ์ด,อีคอมเมิร์ซ

2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ (Decision Support System) เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจ ส่วนมากข้อมูลจะถูกจัดเก็บแยกมาจาก ระบบปฏิบัติการ (Operational System) โดยจัดอยู่ในรูปของคลังหรือเหมืองข้อมูล ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้

3. ระบบคอมพิวเตอร์สมรรถนะสูงมีราคาต่ำลง และเนื่องจากเทคนิคดาต้าไมนิ่ง ประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนและความต้องการการคำนวณสูง จึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์สมรรถนะสูง ปัจจุบันระบบคอมพิวเตอร์สมรรถนะสูงมีราคาต่ำลง พร้อมด้วยเริ่มมีเทคโนโลยีที่นำเครื่องไมโครคอมพิวเตอร์จำนวนมาก มาต่อเชื่อมกันโดยเครือข่ายความเร็วสูง ทำให้ได้ระบบคอมพิวเตอร์ สมรรถนะสูงในราคาต่ำ

4. การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า เนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสหกรรมและการค้า มีการผลิตข้อมูลไว้อย่างมากมายแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์ จึงเป็นการจำเป็นอย่างยิ่งที่ต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูล ความรู้ที่ได้รับสามารถนำไปวิเคราะห์เพื่อการตัดสินใจ ในการบริหารจัดการในระบบต่าง ๆ ซึ่งจะเห็นได้ว่าความรู้เหล่านี้ถือว่าเป็นผลิตผลอีกชิ้นหนึ่งเลยทีเดียว

## 2.5 ประเภทข้อมูลที่สามารถนำมาทำดาต้าไมนิ่ง

1. Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย entity-relationship (ER model)

2. Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกัน และรวบรวมไว้ในที่ๆ เดียวกัน

3. Transactional Database ประกอบด้วยข้อมูลที่แต่ละทรานเซกชันแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่น ๆ เช่น ข้อมูลแบบ object-oriented , ข้อมูลที่เป็น text file , ข้อมูลมัลติมีเดีย , ข้อมูลในรูปของ web

## 2.6 ลักษณะเฉพาะของข้อมูลที่สามารถทำดาต้าไมนิ่ง

1. ข้อมูลขนาดใหญ่ เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่า หรือโดยการใช้ Database Management System ( DBMS ) ในการจัดการฐานข้อมูล

2. ข้อมูลที่มาจากหลายแหล่ง โดยอาจรวบรวมมาจากหลายระบบปฏิบัติการหรือหลาย DBMS เช่น Oracle , DB2 , MS SQL , MS Access เป็นต้น

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลง ตลอดช่วงเวลาที่ทำการไมนิ่งข้อมูล หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำไมนิ่ง แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำไมนิ่ง สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำดาต้าไมนิ่งใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำไมนิ่งได้เช่นกัน แต่ต้องใช้เทคนิคการทำดาต้าไมนิ่งขั้นสูง

## 2.7 ขั้นตอนการทำดาต้าไมนิ่ง (Process of Data Mining)

ขั้นตอนในการทำดาต้าไมนิ่งหรือเรียกอีกอย่างหนึ่งว่า Knowledge Discovery in Database (KDD) มีขั้นตอนในการทำงานที่สำคัญ ดังต่อไปนี้

1. การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)
2. การเตรียมข้อมูล (Data Preparation)
3. การแปลงรูปแบบข้อมูล (Data Transformation)
4. การทำดาต้าไมนิ่ง (Data Mining)
5. การวิเคราะห์ผลลัพธ์ (Analysis of Results)
6. การนำความรู้ที่ได้ไปใช้งาน (Assimilation of Knowledge)

**2.7.1 การกำหนดวัตถุประสงค์ทางธุรกิจ**

เป็นตัวจักรที่สำคัญในการทำดาต้าไมนิ่ง เนื่องจากเป็นการกำหนดขอบเขตเป้าหมาย ของการทำดาต้าไมนิ่ง ซึ่งจะมีผลต่อทุกๆขั้นตอนของการทำดาต้าไมนิ่ง โดยนักวิเคราะห์ธุรกิจ (Business Analyst) จะต้องระบุ ปัญหาที่เกิดขึ้นในการทำธุรกิจให้ครอบคลุมและชัดเจนรวมทั้งวัตถุประสงค์ ซึ่งขั้นตอนนี้จะสามารถมองถึงอัลกอริทึมและฐานข้อมูลที่จะใช้งานเบื้องต้น เป็นการนำไปสู่การสร้างแบบจำลองที่เหมาะสม ขึ้นอยู่กับเป้าหมายทางธุรกิจ

**2.7.2 การเตรียมข้อมูล**

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้ระยะเวลาประมาณ 60%ของการทำดาต้าไมนิ่ง นับเป็นขั้นตอนที่ใช้เวลานานที่สุด ประกอบด้วย

**2.7.2.1 การคัดเลือกข้อมูล (Data Selection)**

จุดมุ่งหมายของการคัดเลือกข้อมูล คือ การระบุถึงแหล่งข้อมูลที่นำมาใช้ที่จำเป็นต่อการนำมาวิเคราะห์ข้อมูลเบื้องต้น รวมถึงจะต้องมีความเข้าใจเกี่ยวกับลักษณะและตัวแปรของข้อมูลที่จะนำมาทำดาต้าไมนิ่งด้วย ซึ่งตัวแปรของข้อมูลแบ่งออกเป็น 2 ประเภท ดังนี้

1. ข้อมูลที่แบ่งเป็นกลุ่ม (Categorical data) มี 2 ประเภท คือ
   1) Nominal คือ ข้อมูลแบบที่ไม่คำนึงถึงลำดับ หรือลำดับไม่มีความสำคัญ เช่น สถานภาพ (โสด,แต่งงาน,หย่าร้าง) เพศ (ชาย,หญิง) ระดับการศึกษา (มัธยมศึกษา, ปริญญาตรี,ปริญญาโท,ปริญญาเอก)
   2) Ordinal คือ ข้อมูลแบบที่คำนึงถึงลำดับ หรือลำดับมีความสำคัญ เช่น การจัดระดับเครดิตของลูกค้า (ดี,ปานกลาง,แย่)
2. ข้อมูลแบบที่เป็นตัวเลข (Quantitative data) มี 2 ประเภท คือ
   1) Continuous ข้อมูลที่เป็นจำนวนจริง เช่น รายได้,รายจ่าย, ผลกำไร,ค่าเฉลี่ย เป็นต้น
   2) Discrete ข้อมูลที่มีค่าไม่ต่อเนื่อง เป็นจำนวนเต็ม เช่น จำนวนพนักงาน , จำนวนลูกค้า เป็นต้น

**2.7.2.2 การประมวลข้อมูลก่อน (Data Preprocessing)**

จุดมุ่งหมายของขั้นตอนนี้เป็นการนำเอาข้อมูลที่จะใช้ในการทำดาต้าไมนิ่งมาทำให้เป็นข้อมูลที่มีคุณภาพดีก่อนที่จะนำไปใช้งานต่อไป โดยเป็นการตรวจสอบว่าข้อมูลที่ได้เลือกไว้ในขั้นตอนการคัดเลือกข้อมูลนั้น มีความเหมาะสมหรือไม่ เช่น ข้อมูลแบบ Categorical ใช้วัดการกระจายของข้อมูล เพื่อทำความเข้าใจข้อมูลได้ดียิ่งขึ้นโดยอาศัยเครื่องมือทางด้านการสร้างภาพนามธรรม (Visualize) แสดงข้อมูล เช่น กราฟแท่ง ส่วนข้อมูลแบบ Quantitative ที่เป็นตัวเลข วัด

โดยการหาค่าสูงสุด ต่ำสุด ค่าเฉลี่ย ค่ามัธยฐาน และตัววัดทางสถิติอื่นๆ ซึ่งการประมวลข้อมูลก่อนนี้ ประกอบด้วย

1) การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนที่ทำให้ ข้อมูลมีความสมบูรณ์ ถูกต้อง และสอดคล้องกัน เป็นการเพิ่มค่าที่ขาดหายไป (Missing Values) การระบุ Noisy Data ค่าความผิดพลาดหรือความแปรปรวน ที่เกิดขึ้นจากการเก็บรวบรวมข้อมูล การป้อนข้อมูลเข้าสู่ระบบ และการรับส่งข้อมูล ความไม่สอดคล้องกันจากการตั้งชื่อแล้วจึงทำการปรับปรุงค่าข้อมูลให้มีความสอดคล้องกัน เช่น ข้อมูลในฟิลด์ที่ขาดหายไป อาจจะแทนค่าข้อมูลที่ขาดหายไปด้วย Unknown หรือถ้าหากข้อมูลขาดหายไปเป็นจำนวนมากและข้อมูลนั้นไม่สำคัญมากนักอาจจะทำการตัดฟิลด์นั้นทิ้งไป
2) การรวมข้อมูล (Data Integration) เป็นขั้นตอนที่รวบรวมข้อมูลมาจากหลายๆแหล่ง แล้วทำการตรวจหา และขจัดความขัดแย้งและความซ้ำซ้อนของข้อมูล

**2.7.2.3 การแปลงรูปแบบข้อมูล (Data Transformation)**

เป็นขั้นตอนที่ทำการรวบรวมข้อมูลหรือเปลี่ยนแปลงข้อมูล เพื่อให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมที่ใช้ในการทำดาต้าไมนิ่งของงาน ซึ่งความหมาะสมของข้อมูลก็ขึ้นอยู่กับโมเดลที่เราจะใช้งาน ตัวอย่างของโมเดลที่จะใช้งานไม่สามารถทำการคำนวณข้อมูลที่เป็นตัวอักษรได้ ก็จะต้องแปลงตัวอักษรไปเป็นตัวเลขก่อน เช่น ระดับการศึกษา ปริญญาตรี ปริญญาโท และปริญญาเอก ไปเป็นตัวเลข 1, 2, 3 เพื่อให้สอดคล้องกับโมเดลที่จะใช้งาน

**2.7.3 การทำดาต้าไมนิ่ง**

เป็นขั้นตอนในการประมวลผลข้อมูลตามวิธีและอัลกอริทึมที่ได้เลือกไว้ ให้มีความเหมาะสมกับการใช้งาน ซึ่งอาจจะต้องใช้วิธีการและเทคนิคต่างๆ มารวมกัน เพื่อให้ได้ผลลัพธ์ที่ดี ซึ่งการดำเนินการ (Operation) ที่นิยมใช้โดยทั่วไป มีหลายแบบ เช่น Database Segmentation, Predictive Modeling, Link Analysis เป็นต้น แต่ละ Data Mining Operation จะมีอัลกอริทึมให้เลือกใช้ เช่น การทำ Database Segmentation อาจใช้ K-Mean Algorithms หรืออาจใช้ Unsupervised Learning Neural Networks เช่น โมเดล Kohonen Neural Net ถ้าเป็นการทำ Predictive Modeling อาจใช้ CART (Classification And Regression Tree) หรืออาจใช้ Supervised Learning Neural Network เช่น Backpropagation Neural Net ถ้าเป็นการทำ Link Analysis ซึ่งมีการทำอยู่ 2 ลักษณะ คือ Association Rule Discovery และ Sequential Pattern Discovery อาจใช้ Apriori Algorithms

**2.7.4 การวิเคราะห์ผลลัพธ์**

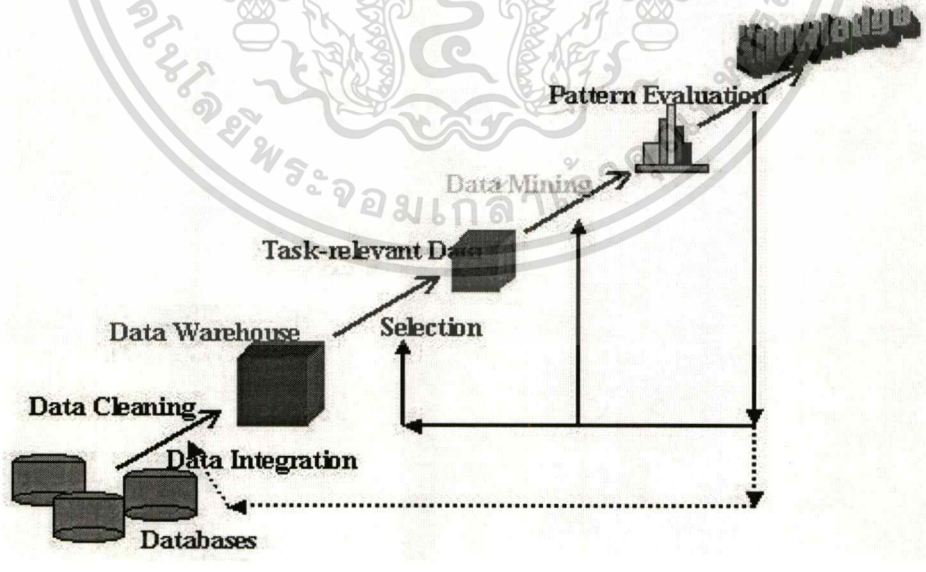
การวิเคราะห์ผลลัพธ์เป็นขั้นตอนที่ทำการวิเคราะห์ผลลัพธ์ และแปลความหมายที่ได้จากการทำดาต้าไมนิ่งในขั้นตอนที่ผ่านมา โดยต้องอาศัยทักษะจากประสบการณ์ ความรู้ความชำนาญในเรื่องที่เกี่ยวข้อง รวมถึงต้องเป็นไปตามการดำเนินการและเทคนิคที่เลือก

**2.7.5 การนำความรู้ที่ได้ไปใช้งาน**

การนำความรู้ที่ได้ไปใช้งาน เป็นขั้นตอนในการเลือกและรวบรวมความรู้ที่ได้จากการวิเคราะห์ผลลัพธ์ นำไปประยุกต์ใช้กับองค์กรจริงๆ เนื่องจากผลลัพธ์ที่ได้อาจมีได้หลายรูปแบบ ซึ่งพบว่าบางผลลัพธ์อาจไม่เป็นประโยชน์กับองค์กร ทำให้ต้องมีการวัดความน่าสนใจของผลลัพธ์ โดยสามารถวัดได้จาก

1) เป็นสารสนเทศที่ไม่เคยรู้มาก่อน (Unknown Information)
2) สารสนเทศที่ได้รับต้องมีความสมเหตุสมผล (Valid) และเชื่อถือได้ (Reliability)
3) สารสนเทศที่ได้จะต้องสามารถนำไปใช้ให้เกิดประโยชน์กับองค์กรได้จริง

ดังนั้น จากการศึกษาถึงขั้นตอนต่างๆ ในการทำดาต้าไมนิ่ง ทำให้เราทราบถึงลักษณะการทำงานในแต่ละขั้นตอน ซึ่งขั้นตอนต่างๆของการทำดาต้าไมนิ่ง สามารถแสดงเป็นรูปภาพได้ดังรูปที่ 2.2




**รูปที่ 2.2** แสดงขั้นตอนต่างๆของการทำดาต้าไมนิ่ง

## 2.8 เทคนิคของดาต้าไมนิ่ง

เทคนิคของการทำดาต้าไมนิ่ง ที่นิยมใช้โดยทั่วไป มีดังนี้

1. Predictive Modeling เป็นการนำข้อมูลมาใช้ในการสร้างโมเดล เพื่อนำไปใช้ในการทำนายค่า แบ่งออกได้เป็น 2 เทคนิค คือ

   1) Classification เป็นการทำนายข้อมูลในอนาคตว่าข้อมูลที่ต้องการพิจารณาควรจะอยู่ในกลุ่มใด โดยมีการแบ่งประเภทกลุ่มไว้อยู่ก่อนแล้ว

   2) Value Prediction เป็นโมเดลที่ใช้ในการทำนายแนวโน้มข้อมูลที่เป็นตัวเลขในอนาคต เช่น การพยากรณ์อากาศ การทำนายราคาหุ้น

2. Database Segmentation เป็นการแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ โดยที่ข้อมูลภายในแต่ละกลุ่ม มีลักษณะที่เหมือนกันหรือใกล้เคียงกัน โดยที่เรายังไม่เคยรู้มาก่อน เช่น ใช้ในการแบ่งกลุ่มของลูกค้าว่ามีจำนวนกี่กลุ่ม

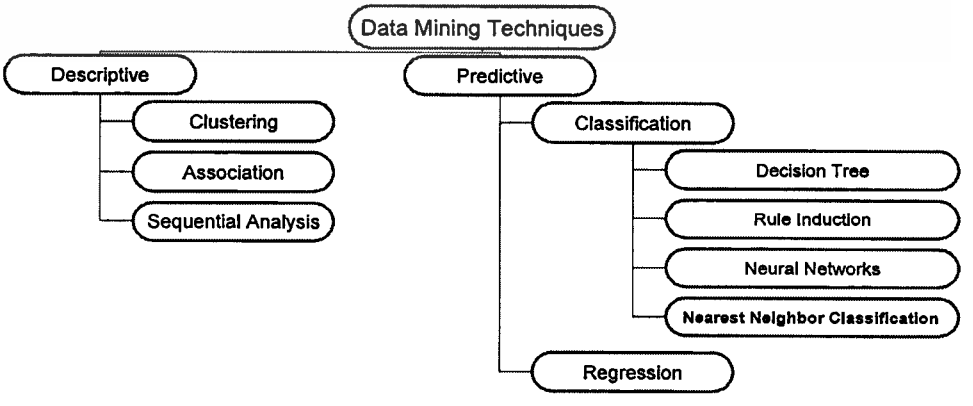
3. Link Analysis เป็นการหาความสัมพันธ์ของข้อมูลในแต่ละเรคคอร์ด หรือกลุ่มของเรคคอร์ดในฐานข้อมูล เช่น การหาความสัมพันธ์ของสินค้าว่าลูกค้ามักจะซื้อสินค้าอะไรควบคู่กันในการซื้อครั้งหนึ่ง (Association Rule) หรือการศึกษาการซื้อสินค้าในระยะยาว (Sequential Pattern Discovery)

4. Deviation Detection เป็นเทคนิคที่ใช้แสดงข้อมูลที่มีลักษณะผิดปกติไปจากข้อมูลทั่วไป แบ่งเป็น 2 ประเภท ดังนี้

   1) Visualization เป็นเทคนิคที่ใช้ในการแสดงข้อมูลในรูปแบบต่างๆ เช่น แผนที่ รูปภาพ กราฟสามมิติ ซึ่งมีประสิทธิภาพในการสื่อสารค่อนข้างมาก

   2) Statistics เป็นการใช้วิธีทางสถิติเข้ามาช่วยตรวจจับข้อมูล

เทคนิคต่างๆของการทำดาต้าไมนิ่ง สามารถสรุปได้ดังรูปที่2.3




**รูปที่ 2.3** แสดงเทคนิคต่างๆดาต้าไมนิ่ง

สำหรับโครงการนี้ จะนำเสนอกระบวนการของการพยากรณ์ (Predictive Model) แบบ Classification ในการทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์มือถือ ซึ่งจะใช้เทคนิคของ Tree Decision ในการวิเคราะห์ เนื่องจากเทคนิค Tree Decision จะสามารถเรียนรู้ถึงรูปแบบของข้อมูล โดยสามารถแสดงให้เห็นภาพที่ชัดเจนออกมาในรูปของแผนภูมิต้นไม้ ที่จะค่อยๆแตกออกมาเป็นกิ่งก้าน ทำให้สามารถแปลความหมายของความสัมพันธ์ของข้อมูลให้เข้าใจได้ง่าย การนำไปใช้เพื่อให้บรรลุวัตถุประสงค์ของธุรกิจก็จะทำได้อย่างมีประสิทธิภาพ มีความถูกต้องมากขึ้น

# บทที่ 3
# เทคนิคการพยากรณ์ (Predictive Model)

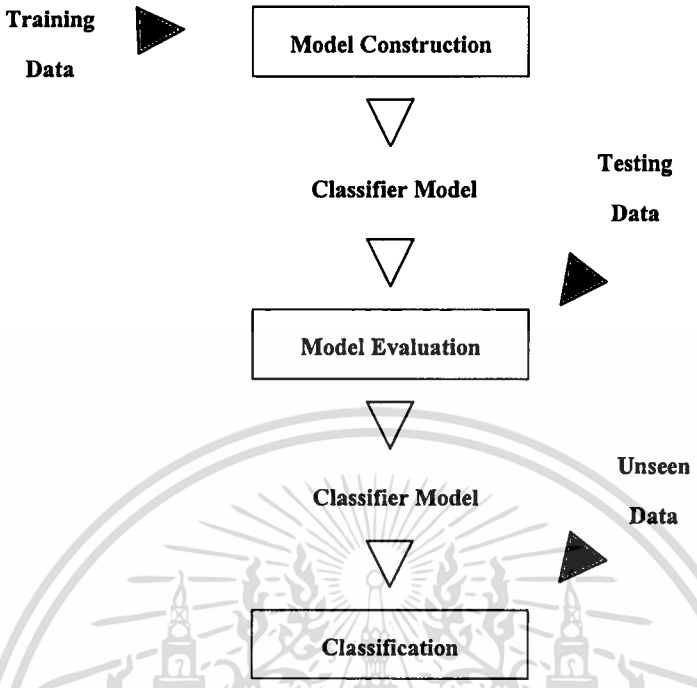
## 3.1 การสร้างแบบจำลองพยากรณ์

จากบทความข้างต้นที่พูดถึงเทคนิคการทำดาต้าไมนิ่ง ที่มีหลากหลายรูปแบบ โดยในบทนี้เราจะมาเน้นในเรื่องของการพยากรณ์ผลลัพธ์โดยใช้เทคนิคของดาต้าไมนิ่ง

Predictive Data Mining หมายถึง การค้นหารูปแบบความสัมพันธ์ของฐานข้อมูลที่มีอยู่เป็นจำนวนมาก เพื่อนำมาใช้ในการพยากรณ์ผลลัพธ์ หรือนำมาใช้ในการคำนวณ เพื่อช่วยในการตัดสินใจในอนาคตได้อย่างถูกต้องและแม่นยำ ซึ่งเทคนิคในการ Predictive จะเป็นการศึกษา Decision Criteria ของข้อมูลที่เก็บไว้ในอดีต เหมือนกับการเรียนรู้จากประสบการณ์ของมนุษย์ที่จะจดจำเหตุการณ์ที่เคยเกิดขึ้นว่ามีลักษณะแบบใด เพื่อตัดสินใจ ถ้าหากเกิดเหตุการณ์ในลักษณะนี้ขึ้นอีก ก็จะสามารถตัดสินใจได้รวดเร็วขึ้น โดยเทคนิคนี้ จะทำนายถึงความเป็นไปได้ ซึ่งจะใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ คือเราจะใช้เทคนิคนี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อตัดสินใจเลือกลักษณะข้อมูลที่ต้องการ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้แล้วจึงนำไปทำนายผลลัพธ์ในกลุ่มข้อมูลที่ต้องการทราบ ซึ่งวิธีนี้เรียกว่า Supervised Learning ดังนั้นข้อมูลที่มีอยู่ต้องสมบูรณ์ จึงจะทำให้ผลลัพธ์ออกมาถูกต้อง เพราะเราต้องนำข้อมูลในอดีตมาสร้างแบบจำลอง โดยในการทำงานจะแบ่งออกเป็น 2 ขั้นตอน คือ

1) ระยะการฝึกอบรม (Training Phase) คือ ขั้นตอนการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ความสัมพันธ์ของข้อมูลในอดีตที่เก็บไว้ในฐานข้อมูล ซึ่งจะใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมดในการสร้างแบบจำลอง
2) ระยะการทดสอบ (Testing Phase) คือ ขั้นตอนที่ใช้ทำการทดสอบแบบจำลองที่สร้างขึ้นมาจาก Training Phase ว่ามีความถูกต้องและมีความน่าเชื่อถือมากน้อยเพียงใด โดยจะนำข้อมูลส่วนที่เหลืออีก 20% จากการแบ่งไว้มาใช้ทดสอบแบบจำลองที่สร้างขึ้น

โดย Predictive Model ที่นำมาใช้ในโครงการนี้คือ เทคนิคการทำ Classification จะเป็นกระบวนการสร้างโมเดลที่จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก, ดี, ปานกลาง และไม่ดี โดยพิจารณาจากประวัติ และผลการเรียน หรือการแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ เป็นต้น โดยพิจารณาจากข้อมูลที่มีอยู่ ซึ่งวิธีที่นิยมใช้กันมากก็คือ Tree Induction และ Neural Induction โดยกระบวนการ Classification นี้แบ่งขั้นตอนการทำงาน เป็น 3 ขั้นตอน ดังรูปที่ 3.1




**รูปที่ 3.1** แสดงกระบวนการของ Classification

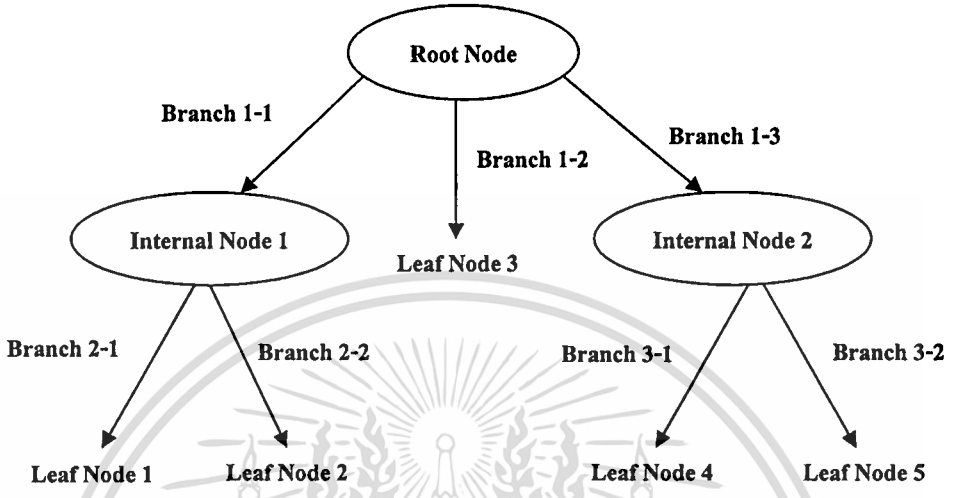
## 3.2 โครงสร้างแบบต้นไม้ (Decision Tree)

ในโครงการนี้ จะเลือกใช้เทคนิคการ Classification แบบวิธี Decision Tree ในการทำการวิเคราะห์ เนื่องจากผลลัพธ์ที่ได้จากการใช้วิธีนี้ จะสามารถตีความหมายของผลการพยากรณ์ได้ง่าย และยังสามารถทำความเข้าใจกระบวนการใช้งานได้ง่าย อีกทั้งยังสามารถหาสาเหตุที่มาที่ไปของผลลัพธ์ได้ในรูป If-Then Rules โดยหลักการของ Tree Decision คือการแตกผลลัพธ์ของตัวแปรที่เรานำมาใช้ในการประมวลผลออกเป็นลำดับชั้น ลักษณะเหมือนแผนภูมิโครงสร้างขององค์กร โดยที่แต่ละโหนด (Node) จะแสดงถึง Attribute ของข้อมูล แต่ละกิ่งแสดงถึงผลในการประมวลผล และ Leaf Node แสดงถึงผลลัพธ์ที่เราต้องการทราบ ซึ่งได้กำหนดไว้แล้ว

ซึ่งเทคนิคของ Tree Decision นั้น จะสามารถรองรับข้อมูลที่มีลักษณะของข้อมูลได้หลายลักษณะ ดังนี้

1) Nominal เป็นลักษณะข้อมูลที่เป็นข้อมูลตัวเลข เช่น 1, 2, 3, 4, 5 เป็นต้น
2) Ordinal เป็นข้อมูลที่มีลักษณะของข้อมูลที่สามารถแยกออกเป็นประเภทต่างๆได้ เช่น ปริญญาตรี, ปริญญาโท, ปริญญาเอก เป็นต้น
3) Interval เป็นข้อมูลที่มีลักษณะเป็นค่าต่อเนื่อง หรือค่าเฉลี่ย เช่น อุณหภูมิ, รายได้ เป็นต้น

โดยจะแสดงผลในรูปแผนภูมิที่จะแสดงการแตกผลลัพธ์ของตัวแปรแต่ละตัวออกมาเรื่อยๆ จนสุดท้ายจะได้คำตอบที่ต้องการ เหมือนลักษณะของกิ่งของต้นไม้ที่แตกออกมา ดังรูปที่ 3.2 เป็นรูปตัวอย่างของ Decision Tree




**รูปที่ 3.2** แสดงตัวอย่างการแสดงผลของ Decision Tree

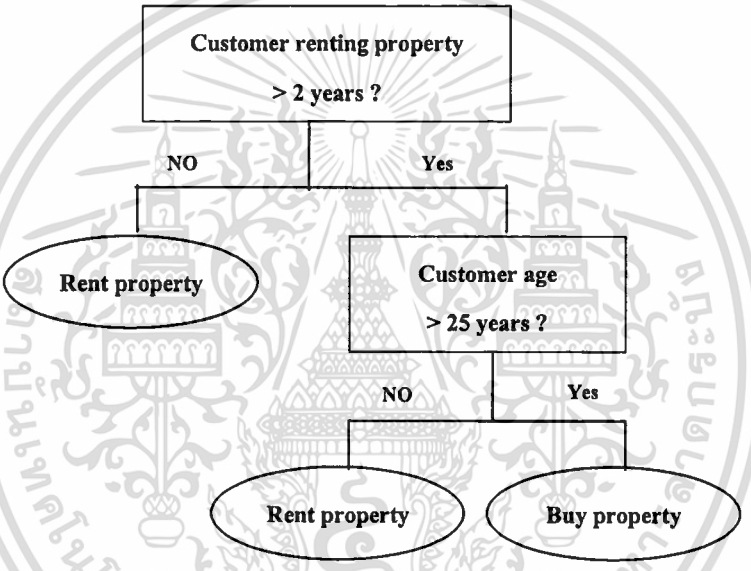
จากรูปจะประกอบไปด้วย Node ต่างๆ จุดที่เริ่มต้นของแผนภูมิจะเรียกว่า Root Node หลังจากนั้นจะแตกข้อมูลออกเป็นกิ่งต่างๆ (Branch) ตามทางเลือกของ Node ต่างๆ ซึ่งกระบวนการนี้จะดำเนินการต่อไปเรื่อยๆ จนกระทั่งได้ผลลัพธ์สุดท้ายของตัวแปรที่เป็นเป้าหมาย (Target Attribute) เรียกว่า Leaf Node ซึ่งจะเก็บค่าของคำตอบไว้ที่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

ยกตัวอย่าง สมมติว่าบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะ ดูแลอาคารต่าง ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา (Branch) ข้อมูลพนักงาน (Staff) ข้อมูลทรัพย์สิน (Property) และข้อมูลลูกค้า (Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูล

เหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer rental) รายการให้เช่า (Rentals) รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า

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

ดังรูปที่ 3.3 แสดงให้เห็นถึง Decision Tree สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า




**รูปที่ 3.3** ตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน

ดังนั้นในการใช้เทคนิค Decision Trees จึงต้องขึ้นอยู่กับลักษณะของข้อมูลที่นำมาใช้ในการประมวลผลด้วย ซึ่งเทคนิคนี้จะประกอบไปด้วย Algorithm หลายประเภท ยกตัวอย่างเช่น CHAID (Chi-Square Automatic Interaction Detection), CART (Classification and Regression Trees), ID3 (Iterative Dichotomiser3), QUEST (Quick, Unbiased, Efficient Statistical Tree), C4.5 หรือ C5.0 เป็นต้น

## 3.3 อัลกอริทึมที่นิยมใช้ในการสร้าง Decision Tree

ในแต่ละอัลกอริทึม จะมีวิธีที่แตกต่างกันในการหา Attribute และการตัด Tree ที่เกิด Over fitting ในที่นี้จะขอยกตัวอย่างประกอบเพื่ออธิบายการทำงานของแต่ละอัลกอริทึม โดยตัวอย่างที่ยกมานั้นเป็นข้อมูลเกี่ยวกับลูกค้า เพื่อใช้ในการทำนายการซื้อสินค้า ในที่นี้สมมติให้เป็น Product Y

**ตารางที่ 3.1** แสดงชุดข้อมูลเกี่ยวกับการซื้อสินค้า Product Y

| Age | Income | Working | Credit Rating | Product Y |
|---|---|---|---|---|
| ≤ 30 | High | No | Fair | No |
| ≤ 30 | High | No | Excellent | No |
| 31...40 | High | No | Fair | Yes |
| > 40 | Medium | No | Fair | Yes |
| >40 | Low | Yes | Fair | Yes |
| >40 | Low | Yes | Excellent | No |
| 31...40 | Low | Yes | Excellent | Yes |
| ≤ 30 | Medium | No | Fair | No |
| ≤ 30 | Low | Yes | Fair | Yes |
| >40 | Medium | Yes | Fair | Yes |
| ≤ 30 | Medium | Yes | Excellent | Yes |
| 31...40 | Medium | No | Excellent | Yes |
| 31...40 | High | Yes | Fair | Yes |
| >40 | Medium | No | Excellent | No |

โดยมี Class labels (Product Y) กำหนดไว้ 2 ค่า คือ Yes , No และ สิ่งที่นำมาพิจารณา คือ อายุ รายได้ การทำงาน และ เครดิต โดย

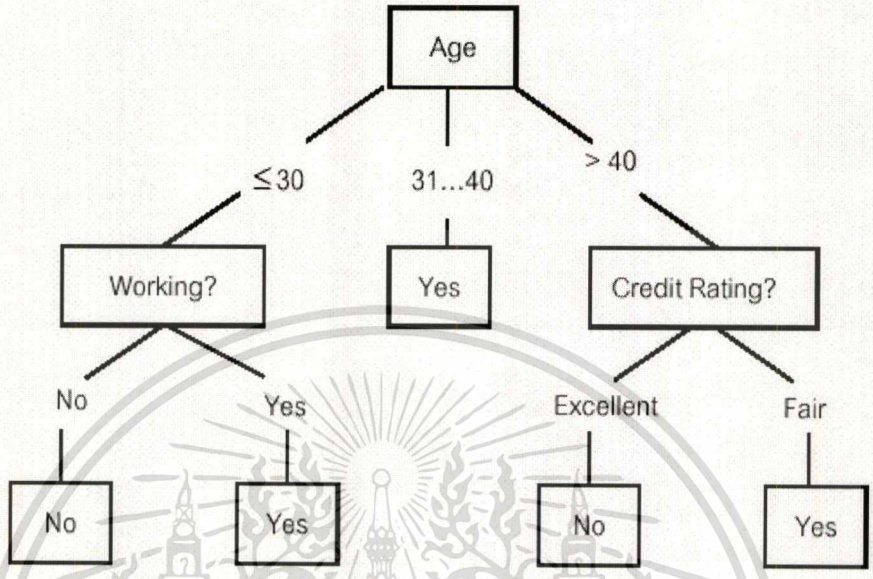
อายุ แบ่งออกไว้ 3 ค่า คือ อายุน้อยกว่า 30, อายุระหว่าง 31 - 40 และ อายุ 40 ปีขึ้นไป

รายได้ แบ่งออกได้ 3 ค่า คือ High, Medium และ Low

การทำงาน แบ่งออกได้ 2 ค่า คือ Yes, No

เครดิต แบ่งออกได้ 2 ค่า คือ Fair, Excellent

**Decision Tree For " Product_Y "**




**รูปที่ 3.4** แสดง Decision Tree ของ Product Y

จากรูป เราจะดูจากอายุก่อนว่าเข้าอยู่ในกลุ่มใด โดยมี 3 กลุ่ม คือ อายุ ≤ 30, 31 - 40 และ อายุ > 40 จากนั้นมาดูต่อว่า

- กลุ่ม ≤ 30 ปี ต้องดูว่าทำงานหรือไม่
  - ถ้า ไม่ทำงาน ผล คือ จะไม่ซื้อ
  - ถ้า ทำงาน ผล คือ จะซื้อ
- กลุ่ม 31 - 40 ทุกคน จะซื้อสินค้า
- กลุ่ม >40 ต้องดูว่าเครดิตเป็นอย่างไร
  - ถ้า เครดิต = Excellent จะไม่ซื้อ
  - ถ้า เครดิต = Fair จะซื้อ

จากตัวอย่างข้างต้น เราจะทราบได้อย่างไรว่าควรใช้อายุเป็นตัวแบ่งเริ่มต้น แล้วตามด้วยการทำงาน หรือ เครดิต การใช้ตัวแบ่งแตกต่างกันไป ก็จะได้โครงสร้างที่แตกต่างกันด้วย ทำให้เกิด Algorithm ต่าง ๆ กันไป

### 3.3.1 ID3 Algorithm

เป็น Algorithm ที่ใช้หลักการของการใช้ทฤษฎีข่าวสาร ค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่ง โดยมีสูตรการคำนวณดังนี้

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad \textbf{(3.1)}$$

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i) \quad \textbf{(3.2)}$$

$$Gain\,(A) = I(p,n) - E(A) \quad \textbf{(3.3)}$$

เช่น ถ้าข้อมูลมีด้วยกัน 6 records

<u>กรณีที่ 1</u> : ตอบ Yes = 6 No = 0

คือ $I(6,0) = -\left(\frac{6}{6}\log 2\frac{6}{6}\right) - \left(\frac{6}{6}\log 2\frac{0}{6}\right)$

$= 0$

แบบนี้ ได้ค่า 0 คือไม่มี ความสับสนเลย

<u>กรณีที่ 2</u> : ตอบ Yes = 3 No = 3

คือ $I(3,3) = -\left(\frac{3}{6}\log 2\frac{3}{6}\right) - \left(\frac{3}{6}\log 2\frac{3}{6}\right)$

$= 1$

แบบนี้ได้ค่า 1 คือ มีความสับสนมากที่สุด เพราะความเป็นไปได้ 50 : 50

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

<u>ตัวอย่าง</u> การหาตัวแบ่งในการทำ Decision Tree จากตารางเดิมจะได้ คือ มี Attribute 4 รายการ นำมาพิจารณาคำตอบสำหรับการซื้อหรือไม่ซื้อผลิตภัณฑ์ Y คือ

1. Age
2. Income
3. Working
4. Credit Rating

ข้อมูลเดิม (Original Data) ทั้งหมด 14 Record ได้ค่าความสับสนเบื้องต้นคือ

P = Yes มี 9 Record, N = No มี 5 Record

$I(p,n) = I(9,5)$

$= 0.945$

เราจะพิจารณา ทีละ Attribute โดยมี

1. พิจารณา Age ซึ่ง Age มีอยู่ด้วยกันอยู่ 3 กลุ่ม แล้วมีผลต่อ Product Y ดังตาราง คือ

| Age | รวม | Product –Y = Yes | Product – Y = NO |
|---|---|---|---|
| ≤30 | 5 | 2 | 3 |
| 31- 40 | 4 | 4 | 0 |
| >40 | 5 | 3 | 2 |

จากสมการที่ 3.2 $$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= I([2,3], [4,0], [3,2])$$

$$= \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$$

$$= 0.694$$

$$\text{Gain (Age)} = I(p,n) - I([2,3], [4,0], [3,2])$$

$$= 0.94 - 0.694$$

$$= 0.246$$

2. พิจารณาที่ Attribute Income มีอยู่ด้วยกัน 3 กลุ่ม ที่มีผลต่อ Product Y ดังตาราง คือ

| Income | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| High | 4 | 2 | 2 |
| Medium | 6 | 4 | 2 |
| Low | 4 | 2 | 1 |

$$I([2,2], [4,2], [3,1]) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1)$$

$$= 0.911$$

$$\text{Gain (Income)} = I(p,n) - I([2,2], [4,2], [3,1])$$

$$= 0.94 - 0.911$$

$$= 0.29$$

3. พิจารณา Attribute Working มีอยู่ ด้วยกัน 2 กลุ่ม ที่มีผลต่อ Product Y ดังตาราง คือ

| Working | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Yes | 7 | 6 | 1 |
| No | 7 | 3 | 4 |

$$I([6,1],[3,4]) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4)$$

$$= 0.789$$

$$\text{Gain (Working)} = I(p,n) - I([6,1],[3,4])$$

$$= 0.94 - 0.789$$

$$= 0.151$$

4. พิจารณาที่ Attribute Credit Rating มีอยู่ด้วยกัน 2 กลุ่ม ที่มีผลต่อ Product Y ดังตาราง คือ

| Credit Rating | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Fair | 8 | 6 | 2 |
| Excellent | 6 | 3 | 3 |

$$I([6,2],[3,3]) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$$

$$\text{Gain (Credit Rating)} = I(p,n) - I([6,2],[3,3])$$

$$= 0.94 - 0.892$$

$$= 0.048$$

เมื่อพิจารณาครบทุก Attribute จะได้ผลดังนี้

Gain (Age) = 0.246

Gain (Income) = 0.029

Gain (Working) = 0.151

Gain (Credit Rating ) = 0.048

จะเห็นว่า Gain (Age) มีค่ามากที่สุด ซึ่งเลือกเป็น Attribute ในการนำมาเป็นทางเลือกตัวแรก

หาลำดับที่ 2 จากการแบ่งด้วย Age แล้ว ซึ่งมี 3 ทางเลือก

2.1 อายุ ≤ 30 ปี มี Yes = 2 , No = 3 รวม 5 ตัว

| Age | รวม | Product –Y = Yes | Product – Y = NO |
|---|---|---|---|
| ≤30 | 5 | 2 | 3 |

$$I(\text{Age ไปยัง Age}) = \frac{5}{5} I(2,3) + \frac{0}{5} (0,0) + \frac{0}{5} (0,0)$$

$$= I(2,3)$$

$$\text{Gain (Age ไปยัง Age)} = I(2,3) - I(2,3)$$

$$= 0$$ แสดงว่าไม่เกิดประโยชน์ในการใช้ซ้ำ

Age ≤ 30 ไปดูต่อที่ Income

| Income | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| High | 2 | 0 | 2 |
| Medium | 2 | 1 | 1 |
| Low | 1 | 1 | 0 |

$$I(\text{Age ไปยัง Income}) = \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0)$$

$$= 0.400$$

$$\text{Gain (Age ไปยัง Income)} = I(2,3) - I([0,2],[1,1],[1,0])$$

$$= 0.971 - 0.400$$

$$= 0.571$$

Age ≤ 30 ไปดูต่อที่ Working

| Working | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Yes | 2 | 2 | 0 |
| No | 3 | 3 | 0 |

$$I(\text{Age ไปยัง Working}) = \frac{2}{5} I(2,0) + \frac{3}{5} I(3,0)$$

$$= 0$$

$$\text{Gain (Age ไปยัง Working)} = I(2,3) - I([2,0],[3,0])$$

$$= 0.971 - 0$$

$$= 0.971$$

Age ≤ 30 ไปดูต่อที่ Credit Rating

| Credit Rating | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Fair | 3 | 5 | 1 |
| Excellent | 2 | 1 | 1 |

$$I(\text{Age ไปยัง Credit Rating}) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1)$$

$$= 0.951$$

$$\text{Gain ( Age ไปยัง Credit Rating)} = I(2,3) - I([2,1],[1,1])$$

$$= 0.020$$

จาก Age ≤ 30 จะได้ผล ดังนี้

Gain (Age ≤ 30) ไปยัง Age = 0

Gain (Age ≤ 30) ไปยัง Income = 0.571

Gain (Age ≤ 30) ไปยัง Working= 0.971

Gain (Age ≤ 30) ไปยัง Credit Working = 0.020

ทำให้สรุปได้ว่า Age ≤ 30 จะต้องใช้ Working เป็นตัวแบ่งตัวต่อไป

2.2 Age 31 – 40 มี Yes = 4 , No = 0

| Age | รวม | Product –Y = Yes | Product – Y = NO |
|---|---|---|---|
| 31- 40 | 4 | 4 | 0 |

จะเห็นว่ามีแต่ Product Y = Yes ทั้งหมดแล้ว จึงไม่ต้องคำนวณต่อ สามารถบอกคำตอบได้ทันทีว่าถ้า อายุ 31- 40 แล้ว จะมี Product Y = Yes

2.3 Age > 40 มีทั้งหมด 5 Record เป็น ค่า Yes = 3 , No = 2

| Age | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| > 40 | 5 | 3 | 2 |

$$I(\text{Age ไปยัง Age}) = \frac{0}{5} I(0,0) + \frac{0}{5} I(0,0) + \frac{5}{5} I(3,2)$$

$$= I(3,2)$$

$$\text{Gain(Age ไปยัง Age)} = I(3,2) - I(2,3)$$

$$= 0$$ แสดงว่าไม่เกิดประโยชน์ในการใช้ซ้ำ

Age > 40 ไปดูต่อที่ Income

| Income | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| High | 0 | 0 | 0 |
| Medium | 3 | 2 | 1 |
| Low | 2 | 1 | 1 |

$$I(\text{Age ไปยัง Income}) = \frac{0}{5} I(0,0) + \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$$

$$= 0.951$$

$$\text{Gain(Age ไปยัง Income)} = I(3,2) - I([0,0], [2,1], [1,1])$$

$$= 0.971 - 0.951$$

$$= 0.020$$

Age > 40 ไปดูต่อที่ Working

| Working | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Yes | 3 | 2 | 1 |
| No | 2 | 1 | 1 |

$$I(\text{Age ไปยัง Working}) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$$

$$= 0.951$$

$$\text{Gain (Age ไปยัง Working)} = I(3,2) - I([2,1], [1,1])$$

$$= 0.971 - 0.951$$

$$= 0.020$$

Age > 40 ไปดูต่อที่ Credit Rating

| Credit Rating | รวม | Product_Y = Yes | Product_Y = NO |
|---|---|---|---|
| Fair | 3 | 3 | 0 |
| Excellent | 2 | 2 | 0 |

$$I(\text{Age ไปยัง Credit Rating}) = \frac{3}{5} I(3,0) + \frac{2}{5} I(2,0)$$

$$= 0$$

$$\text{Gain (Age ไปยัง Credit Rating)} = I(3,2) - I([3,0],[2,0])$$

$$= 0.971 - 0$$

$$= 0.971$$

จากอายุ Age > 40 จะได้ผล ดังนี้

Gain(Age > 40 ไปยัง Age) = 0

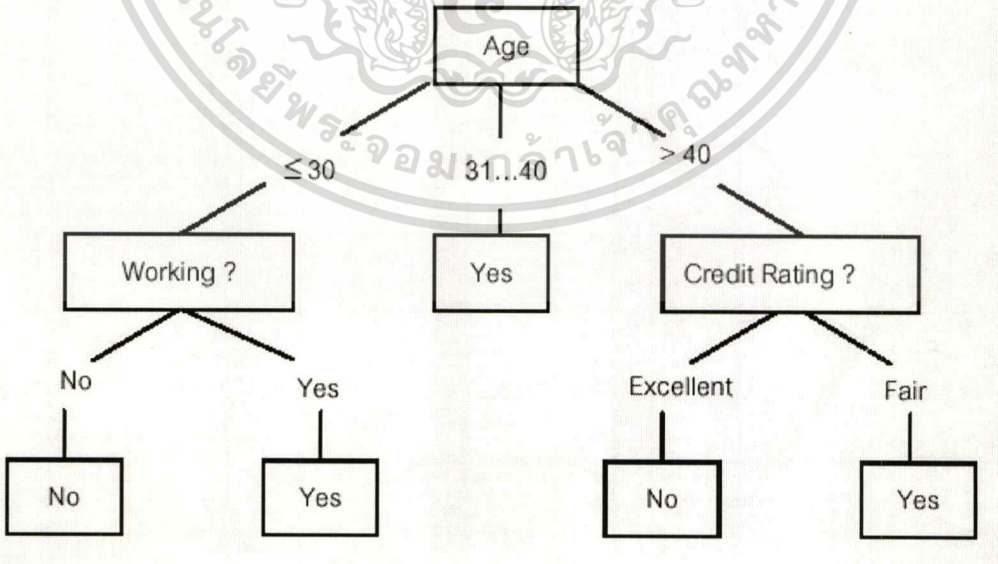
Gain(Age > 40 ไปยัง Income) = 0.020

Gain(Age > 40 ไปยัง Working) = 0.020

Gain(Age > 40 ไปยัง Credit Rating) = 0.971

จึงทำให้สรุปได้ว่า จาก Age > 40 จะต้องใช้ Credit Rating เป็นตัวแบ่งตัวต่อไป

จากการคำนวณจะสรุปได้ดังรูป 3.5




**รูปที่ 3.5** แสดง Decision Tree ของ Product Y จากการคำนวณ

และหากต้องการเขียนเป็น Algorithm หรือเขียน Coding เราสามารถเขียนได้ดังนี้

Example

IF age = "≤ 30 " AND working = " no" THEN product Y = " no"

IF age = "≤ 30 " AND working = " yes" THEN product Y = "yes"

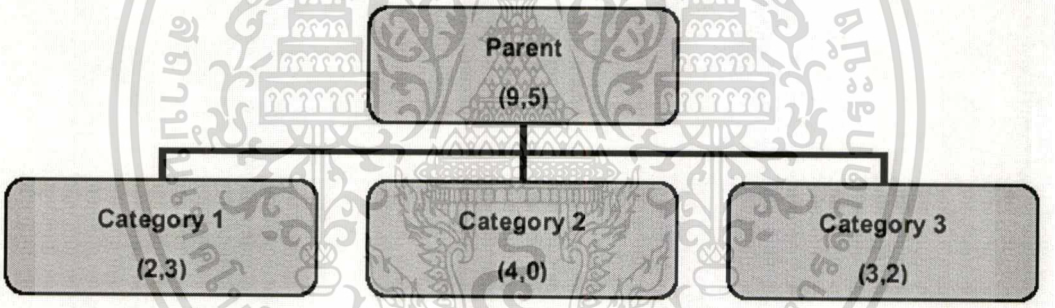
IF age = " 31…40 " THEN product Y = "yes"

IF age = " > 40 " AND Credit Rating = " excellent " THEN product Y = "yes"

IF age = "≤ 30 " AND Credit Rating = " fair " THEN product Y = " no "

### 3.3.2 C5.0 Algorithm

เป็นรูปแบบหนึ่งของการใช้ Decision Tree ที่พัฒนาเพิ่มเติม โดยไม่ได้ใช้ Gian เป็นตัวแบ่ง แต่ใช้ Gain ratio เป็นตัวแบ่งของ Tree ในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือ ต้องหา Info และ gain ออกมาก่อน




Parent : $\text{Info}([9,5]) = -(\frac{9}{14}\log_2\frac{9}{14}) - (\frac{5}{14}\log_2\frac{5}{14})$

$= 0.940$

Category 1 : $\text{Info}([2,3]) = -(\frac{2}{5}\log_2\frac{2}{5}) - (\frac{3}{5}\log_2\frac{3}{5})$

$= 0.971$

Category 2 : $\text{Info}([4,0]) = -(\frac{4}{4}\log_2\frac{4}{4}) - (\frac{0}{4}\log_2\frac{0}{4})$

$= 0.0$

Category 3 : $\text{Info}([3,2]) = -(\frac{3}{5}\log_2\frac{3}{5}) - (\frac{2}{5}\log_2\frac{2}{5})$

$= 0.971$

Category 3 : $\text{Info}([3,2],[4,0],[2,3]) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971$

= 0.693

Gain = Info([9,5]) - Info ([3,2],[4,0],[2,3])

= 0.940-0.693

= 0.247

สิ่งที่ C5.0 ต้องทำเพิ่มเติมคือ การหาค่า gain ratio ซึ่งจะใช้ตัวนี้เป็นตัวแบ่งซึ่งหาได้จากสูตร

$$\text{gain ratio} = \frac{\text{gain}}{\text{split info}} \quad \textbf{(3.4)}$$

ซึ่ง split info ใช้หลักการเดิม ในการหาค่า information ออกมาเป็น3 กลุ่มเป็นการนับจำนวนตัวของข้อมูล ไม่ได้นับจำนวน ค่า Yes หรือ No จากข้อมูลดังกล่าวจะทราบว่า

กลุ่ม อายุน้อยกว่าหรือเท่ากับ 30 มีจำนวน 5 ตัว

กลุ่ม อายุระหว่าง 31-40 มีจำนวน 4 ตัว

กลุ่ม อายุมากกว่า 40 มีจำนวน 5 ตัว

จะได้ค่า Split Info คือ

$$\text{Split info ([5,4,5])} = -\frac{5}{14}\log_2\frac{5}{14} - \frac{4}{14}\log_2\frac{4}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 1.577$$

$$\text{gain ratio} = \frac{\text{gain}}{\text{split info}}$$

$$= \frac{0.247}{1.577}$$

$$= 0.156$$

เราจะได้ค่า gain ratio และทำการคำนวณในรายการอื่น ๆ จนครบ เราก็จะสามารถแสดงได้ว่าเราจะแบ่งการทำงานอยู่ที่ค่าของรายการใด

<u>ข้อเสียของการใช้ ID3</u>

ID3 จะมีค่าของการ gain มีความโน้มเอียงมาก จากการที่เราใช้ค่าที่แตกต่างกันมากที่จะไป split เช่นการใช้ ค่า ID โดยนำ ID ซึ่งทราบอยู่แล้วว่า แต่ละ record จะไม่มีค่าซ้ำกันและ มีจำนวนมาก เช่น หากมีจำนวน 14 record จะต้องทำการสร้าง 14 กิ่ง โดยที่แต่ละกิ่งจะมีค่าเพียงอย่างเดียว ระหว่าง Yes กับ No โดยแต่ละกิ่งจะมีค่า Info([1,0]) หรือ Info([0,1])

Info([1,0],[0,1] ... [1,0],[0,1])

$$= \frac{1}{14}(\text{I}[1,0]) + \frac{1}{14}(\text{I}[0,1]) \ldots + \frac{1}{14}(\text{I}[1,0]) + \frac{1}{14}(\text{I}[0,1])$$

$$= -(\frac{1}{14}\log_2\frac{1}{14}) - (\frac{1}{14}\log_2\frac{1}{14}) \ldots - (\frac{1}{14}\log_2\frac{1}{14}) - (\frac{1}{14}\log_2\frac{1}{14})$$

$$= 0$$

โดยค่า gain = 0.94 – 0 จะได้ค่าเดิม คือไม่เกิดประโยชน์ต่อการทำงาน

โดย C5.0 จะสามารถทำงานแบบนี้ได้ดีกว่า เพราะเป็นการมองจำนวนของข้อมูลในแต่ละกิ่งมากกว่า เมื่อเอามา weight ทำให้ gain ratio ต่ำกว่า ซึ่งหากคำนวณออกมาค่าที่ได้จะมากกว่า 1.577 คือนำตัวมากไปหารค่าเดิมจะได้ค่าที่น้อยกว่าเดิม ซึ่งเป็นเหตุผลที่ต้องทำตัว Split info ขึ้นมา การที่ไม่ต้องการตัวแปรที่มีความเป็นไปได้มาก เพราะทำแล้วทำให้ prediction ได้ยาก

### 3.3.3 CART (Classification and Regression Tree)

CART มีหลักการทำงานเหมือนกับ gain ใน ID3 หรือ gain ratio ใน C5.0 ใน CART จะเรียกว่า Goodness

$$\varphi(s/t) = 2P_L P_R \sum_{J=1}^{\#class} |P(j/t_L) - P(j/t_R)| \qquad (3.5)$$

ซึ่งตัววัดตัวนี้จะมีสาระสำคัญออกเป็น 2 ส่วนคือ

กลุ่มที่ 1 คือ $2P_L P_R$

กลุ่มที่ 2 คือ $\sum_{J=1}^{\#class} |P(j/t_L) - P(j/t_R)|$

โดยตัวแบ่งที่ใช้ต้องทำให้ 2 กลุ่มนี้มีค่ามาก ๆ CART สำคัญคือมีการแบ่งออกเป็น 2 กิ่งเท่านั้น หรือ ซ้ายกับขวา หากมีตัวแปรที่มีความเป็นไปได้มากกว่า 2 จะพยายามจัดกลุ่มให้เหลือเพียง 2 กลุ่มเท่านั้น โดยไม่จำเป็นต้องมีจำนวนเท่ากัน เช่น ถ้ามีข้อมูล 14 ตัวอาจมีด้านซ้าย 8 ตัว ด้านขวา 6 ตัวก็ สามารถทำงานได้ ซึ่งจะได้

$P_L$ = สัดส่วนของ recode ด้านซ้าย จะได้ = 8/14

$P_R$ = สัดส่วนของ recode ด้านขวา จะได้ = 6/14

และค่ามีค่าสูงสุดได้ของสมการนี้แบ่งออกเป็น 2 กลุ่ม

<u>กลุ่มที่ 1</u> คือ $2P_LP_R$ จะต้องให้ $P_L$ และ $P_R$ มีค่าเท่ากับ 0.5 โดยมีความหมายว่าได้แบ่งจำนวนตัวตั้ง 2 ด้านได้เท่าเท่ากัน หรือให้มีค่าใกล้เคียงกันที่สุดเท่าที่จะสามารถทำได้ สาระสำคัญของกลุ่มที่ 1 คือ จะมีค่าสูงเมื่อตัวแบ่ง record ทำให้จำนวนตัวของทั้ง 2 ข้างเท่า ๆ กัน

<u>กลุ่มที่ 2</u> คือ $\sum_{J=1}^{\#class} |P(j/t_L) - P(j/t_R)|$

จะมีค่าสูงสุดได้ เมื่อค่าของ 2 class เช่น ที่เป็น Yes หรือ No โดยจะเห็นว่าค่าที่ได้อยู่ใน Absolute จะได้ค่าสูงเมื่อคำนวณแล้วมีค่าแตกต่างกันมากๆ คือ มีค่า Yes หรือ ค่า No อย่างเดียว ก็จะได้ค่าแตกต่างกันมากสรุปจากสูตร $\varphi(s/t) = 2P_L P_R \sum_{J=1}^{\#class} | P(j/t_L) - P(j/t_R) |$

จะทำให้มีค่ามากจะต้องประกอบด้วย 2 ส่วนคือ

1. ทำให้จำนวน record ด้านซ้ายและขวามีจำนวนเท่ากัน คืออย่างละ 50:50
2. ทำให้จำนวนค่า class ที่กำหนดแบ่งค่าอย่างชัดเจน เช่นมี yes อยู่ด้านซ้ายทั้งหมด และมีค่า No อยู่ด้านขวาทั้งหมด

หลักการในการ Split จะเหมือน ๆ กัน ทุก Algorithm จะมีความแตกต่างกันบ้างก็คือการให้ตัวแบ่งโดย

- ID3 ใช้ gain
- C5 ใช้ gain ratio
- CART ใช้ goodness

# บทที่ 4
# การวิเคราะห์และประมวลผลข้อมูล

ในบทนี้จะกล่าวถึงกระบวนการและขั้นตอนต่างๆในการนำเอาข้อมูลของลูกค้าเกี่ยวกับพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ มาวิเคราะห์และประมวลผลผ่านทางโปรแกรมสำเร็จรูปที่ได้เตรียมไว้ โดยเริ่มจากการเตรียมข้อมูล ซึ่งมีขั้นตอนตั้งแต่การคัดเลือกข้อมูล, การประมวลผลข้อมูลก่อน และการแปลงรูปแบบข้อมูล จากนั้นจะนำข้อมูลที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วไปทำดาต้าไมนิ่ง จากนั้นจะทำการวิเคราะห์และแปลความหมายของผลลัพธ์ที่ได้ ตลอดจนการนำความรู้ที่ได้จากการทำดาต้าไมนิ่งไปใช้งาน

## 4.1 การกำหนดวัตถุประสงค์และขอบเขตของการศึกษา

ขั้นตอนแรกของการทำดาต้าไมนิ่งก็คือ การกำหนดวัตถุประสงค์และขอบเขตของการดำเนินงานว่าเป็นอย่างไร ซึ่งวัตถุประสงค์ของการศึกษาโครงการนี้ คือ ต้องการทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ โดยใช้หลักการของ Predictive Model มาใช้ในการวิเคราะห์ข้อมูล โดยผ่านโปรแกรมสำเร็จรูปในการประมวลผล แล้วนำผลที่ได้จากการวิเคราะห์ไปใช้เป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด เพื่อป้องกันไม่ให้ลูกค้ายกเลิกการใช้บริการ

## 4.2 การเตรียมข้อมูล

เป็นขั้นตอนที่ใช้เวลานานที่สุดในการทำดาต้าไมนิ่ง โดยข้อมูลที่นำมาใช้ในการประมวลผลในโครงการนี้ เป็นข้อมูลที่ได้จากฐานข้อมูลลูกค้าของ บมจ.โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น ซึ่งเป็นผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ระบบ DTAC โดยข้อมูลที่นำมานั้นเป็นข้อมูลของลูกค้า 2 ประเภท คือ ข้อมูลของลูกค้าที่ยกเลิกบริการโทรศัพท์เคลื่อนที่แบบจ่ายรายเดือนในเดือน กรกฎาคม 2549 และลูกค้าที่ยังคงใช้บริการโทรศัพท์เคลื่อนที่แบบจ่ายรายเดือน ณ เดือน กรกฎาคม 2549 โดยข้อมูลที่เกี่ยวกับลูกค้ายกเลิกนั้นจะเรียกว่าข้อมูล Churn และข้อมูลลูกค้าที่ยังคงใช้บริการโทรศัพท์เคลื่อนที่อยู่นั้นจะเรียกว่า ลูกค้า Active ซึ่งวัตถุประสงค์ของการนำเอาข้อมูลทั้ง 2 ประเภทนี้มาใช้ เนื่องจากต้องการศึกษาดูถึงปัจจัยที่มีผลกระทบที่ทำให้ลูกค้ายกเลิกบริการ เพื่อจะได้นำมาใช้ประกอบการตัดสินใจในการกำหนดกลยุทธ์ทางการตลาดขององค์กร และใช้ในการทดสอบโมเดลที่สร้างขึ้นมาเพื่อใช้ในการพยากรณ์ผลการทำนายว่ามีความแม่นยำมากน้อยเพียงใด ซึ่งถ้าหากโมเดลที่นำมาใช้สามารถทำนายผลได้อย่างถูกต้องในระดับความเชื่อมั่นที่สูงมาก ก็จะ

สามารถนำมาใช้เป็นแนวทางในการประมาณการว่าโอกาสที่ลูกค้ามีแนวโน้มที่จะยกเลิกบริการมีมากน้อยเพียงใด เพื่อที่จะได้หาแนวทางที่จะป้องกันไม่ให้ลูกค้ายกเลิกบริการได้สำเร็จ

จากข้อมูลที่นำมาประมวลผลนั้น แบ่งเป็นข้อมูลลูกค้า Active จำนวน 224,573 ราย และลูกค้า Churn จำนวน 18,288 ราย โดยเป็นการสุ่มตัวอย่างประมาณ 30% จากฐานข้อมูลลูกค้า Active และ 70% จากฐานข้อมูลลูกค้า Churn ซึ่งข้อมูลลูกค้าประกอบไปด้วยตัวแปรต่างๆ เกี่ยวกับ ข้อมูลประวัติของลูกค้าและข้อมูลพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้าในช่วง 3 เดือนก่อนที่จะยกเลิกการใช้บริการ ซึ่งมีรายละเอียดแสดงดังตารางที่ 4.1และ ตารางที่ 4.2

**ตารางที่ 4.1** แสดงตัวแปรที่เกี่ยวข้องกับข้อมูลประวัติของลูกค้า

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| Customer No | Text | รหัสลูกค้า |
| Subscriber No | Text | หมายเลขโทรศัพท์เคลื่อนที่ |
| Subscriber Name | Text | ชื่อ-นามสกุลของลูกค้า |
| Subscriber Status | Text | สถานะของลูกค้า |
| Switch on Analysis | Text | ประเภทของการจดทะเบียน |
| Switch on Type | Text | ชนิดของการจดทะเบียน |
| Switch on Reason | Text | สาเหตุของการจดทะเบียน |
| Product Name | Text | ระบบที่ใช้ |
| PackageGroup(Previous) | Text | กลุ่มของโปรโมชั่นที่ใช้มาก่อนหน้า |
| PackageType(Previous) | Text | ประเภทของโปรโมชั่นที่ใช้มาก่อนหน้า |
| Package(Previous) | Text | รายละเอียดของโปรโมชั่นที่ใช้มาก่อนหน้า |
| PackageGroup(Current) | Text | กลุ่มของโปรโมชั่นที่ใช้ในปัจจุบัน |
| PackageType(Current) | Text | ประเภทของโปรโมชั่นที่ใช้ในปัจจุบัน |
| Package(Current) | Text | รายละเอียดของโปรโมชั่นที่ใช้ในปัจจุบัน |
| PackageGroup(Next) | Text | กลุ่มของโปรโมชั่นหลังจากโปรโมชั่นปัจจุบันหมดอายุ |
| PackageType(Next) | Text | ประเภทของโปรโมชั่นหลังจากโปรโมชั่นปัจจุบันหมดอายุ |
| Package(Next) | Text | รายละเอียดของโปรโมชั่นหลังจากโปรโมชั่นปัจจุบันหมดอายุ |

**ตารางที่ 4.1 (ต่อ)**

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| Suspend Type | Text | ประเภทของการถูกระงับบริการ |
| Suspend Reason | Text | สาเหตุของการถูกระงับบริการ |
| No of Suspend AR | Numeric | จำนวนครั้งที่ถูกระงับเนื่องจากมียอดค้างค่าใช้บริการ (Account Receivable) |
| No of Suspend CR | Numeric | จำนวนครั้งที่ถูกระงับเนื่องจากมีค่าใช้จ่ายเกินวงเงินค่าใช้บริการ (Credit Limit) |
| AOUMonth | Numeric | อายุการใช้งานของลูกค้ามีหน่วยเป็นเดือน |
| AgeMonth | Numeric | อายุของลูกค้ามีหน่วยเป็นเดือน |
| Area Description | Text | กลุ่มเบอร์ที่จดทะเบียน |
| Dealer Name | Text | ชื่อร้านค้าที่จำหน่ายเบอร์แก่ลูกค้า |
| Credit Limit Amount | Numeric | จำนวนวงเงินค่าใช้บริการ |

**ตารางที่ 4.2** แสดงตัวแปรที่เกี่ยวข้องกับการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้า

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| Total ARPU 04 | Numeric | จำนวนเงินค่าใช้บริการของใบแจ้งยอดค่าใช้บริการเดือนเมษายน 2549 |
| Total ARPU 05 | Numeric | จำนวนเงินค่าใช้บริการของใบแจ้งยอดค่าใช้บริการเดือนพฤษภาคม 2549 |
| Total ARPU 06 | Numeric | จำนวนเงินค่าใช้บริการของใบแจ้งยอดค่าใช้บริการเดือนมิถุนายน 2549 |
| No of Call 04 (Voice) | Numeric | จำนวนครั้งของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนเมษายน 2549 |
| No of Call 05 (Voice) | Numeric | จำนวนครั้งของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนพฤษภาคม 2549 |
| No of Call 06 (Voice) | Numeric | จำนวนครั้งของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนมิถุนายน 2549 |
| MOU 04 (Voice) | Numeric | จำนวนนาทีของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนเมษายน 2549 |

**ตารางที่ 4.2 (ต่อ)**

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| MOU 05 (Voice) | Numeric | จำนวนนาทีของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนพฤษภาคม 2549 |
| MOU 06 (Voice) | Numeric | จำนวนนาทีของการโทรออกของใบแจ้งยอดค่าใช้บริการเดือนมิถุนายน 2549 |
| MOU 04 (Vas) | Numeric | จำนวนนาทีที่ใช้บริการเสริมของใบแจ้งยอดค่าใช้บริการเดือนเมษายน 2549 |
| MOU 05 (Vas) | Numeric | จำนวนนาทีที่ใช้บริการเสริมของใบแจ้งยอดค่าใช้บริการเดือนพฤษภาคม 2549 |
| MOU 06 (Vas) | Numeric | จำนวนนาทีที่ใช้บริการเสริมของใบแจ้งยอดค่าใช้บริการเดือนมิถุนายน 2549 |
| MOU 04 (IR) | Numeric | จำนวนนาทีการใช้ Inter Roaming ของใบแจ้งยอดค่าใช้บริการเดือนเมษายน 2549 |
| MOU 05 (IR) | Numeric | จำนวนนาทีการใช้ Inter Roamingของใบแจ้งยอดค่าใช้บริการเดือนพฤษภาคม2549 |
| MOU 06 (IR) | Numeric | จำนวนนาทีการใช้ Inter Roaming ของใบแจ้งยอดค่าใช้บริการเดือนมิถุนายน 2549 |

นอกจากนี้ได้มีการสร้างตัวแปรที่ชื่อว่า type ขึ้นมา เพื่อใช้เป็นตัวแปรในการพยากรณ์ว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่หรือไม่ โดยที่ค่าของตัวแปรจะแสดงเป็น Churn และ Non Churn ซึ่ง Churn หมายถึง ลูกค้าที่ยกเลิกบริการไปแล้ว และ Non Churn หมายถึง ลูกค้าที่มีสถานะ Active หรือ ลูกค้าที่ยังคงใช้บริการอยู่

โดยข้อมูลที่ได้มาจากฐานข้อมูลนั้น จะมีความหลากหลายของข้อมูลมาก จึงจำเป็นที่จะต้องมีการกำหนดและแปลงข้อมูลให้อยู่ในลักษณะเดียวกัน เพื่อให้ผลการพยากรณ์มีความถูกต้องแม่นยำมากที่สุด และสามารถตีความได้อย่างมีประสิทธิภาพ ซึ่งจากชุดข้อมูลที่ได้นำมาศึกษานั้น มีการนำข้อมูลมาผ่านกระบวนการคัดเลือกข้อมูล, การประมวลผลข้อมูลก่อน และการแปลงรูปแบบข้อมูล ซึ่งจะกล่าวถึงรายละเอียดของขั้นตอนต่างๆในหัวข้อถัดไป

### 4.2.1 การคัดเลือกข้อมูล

จากชุดข้อมูลที่ได้จากฐานข้อมูลลูกค้า Active และ Churn นั้น เนื่องจากข้อมูลทั้ง 2 ชุดนั้นมีรายละเอียดและจำนวนของตัวแปรที่แตกต่างกัน ดังนั้นจึงต้องนำข้อมูลทั้ง 2 ชุดนั้นมาทำการ

คัดเลือก และปรับแต่งข้อมูลให้มีจำนวนตัวแปรที่เท่ากัน เพื่อที่จะสามารถนำมาทำการวิเคราะห์ข้อมูลร่วมกันได้ โดยรายละเอียดเกี่ยวกับตัวแปรได้แสดงไว้แล้วในตารางที่ 4.1 และตารางที่ 4.2 ทั้งนี้จากรายงานยอดลูกค้ายกเลิกประจำเดือน ทำให้ทราบว่ากลุ่มลูกค้าที่มีจำนวนการยกเลิกบริการค่อนข้างสูง ได้แก่ กลุ่มลูกค้าที่ใช้โปรโมชั่น ZAD และโปรโมชั่น WORK ดังนั้นจึงสนใจที่จะนำลูกค้าทั้ง 2 โปรโมชั่นมาทำการวิเคราะห์ก่อน โดยทำการคัดเลือกข้อมูลเฉพาะลูกค้าที่ใช้โปรโมชั่น ZAD และ WORK โดยใช้ตัวแปรเกี่ยวกับกลุ่มของโปรโมชั่นที่ใช้ในปัจจุบัน หรือ PackageGroup(Current) เป็นตัวแปรในการกรองข้อมูล

### 4.2.2 การประมวลข้อมูลก่อน

เป็นขั้นตอนของการทำข้อมูลให้มีคุณภาพดี โดยในที่นี้ เนื่องจากข้อมูลที่ได้มาเป็นข้อมูลที่ได้จากการสุ่มตัวอย่าง โดยมีรายละเอียดของข้อมูลที่แตกต่างกันไป มีบางชุดข้อมูลที่ข้อมูลบางส่วนขาดหายไป (Missing value) อันเนื่องมาจากการกรอกข้อมูลไม่ครบ หรือเกิดจากความผิดพลาดของระบบที่ทำให้ข้อมูลขาดหายไป ซึ่งจากการประมวลข้อมูลก่อน พบว่าลูกค้าที่มีอายุการใช้งานไม่ถึง 1 เดือนจะไม่มีข้อมูลเกี่ยวกับพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้า โดยข้อมูในระบบจะบันทึกเป็นค่าว่าง ทำให้ไม่สามารถนำข้อมูลมาวิเคราะห์ได้ จึงได้ทำการตัดกลุ่มลูกค้าที่ไม่มีการบันทึกข้อมูลเกี่ยวกับการใช้งานโทรศัพท์เคลื่อนที่ออกไป เนื่องจากไม่เกิดประโยชน์ที่จะนำมาวิเคราะห์ และอาจทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งเกิดความผิดพลาดได้

### 4.2.3 การแปลงรูปแบบข้อมูล

โดยในที่นี้ได้มีการแปลงข้อมูล ของตัวแปรบางตัว เพื่อให้สอดคล้องกับดาต้าไมนิ่งโมเดลที่ใช้ และให้ผลลัพธ์ของการทำไมนิ่งมีประสิทธภาพมากขึ้น โดยในที่นี้ได้มีการสร้างตัวแปรขึ้นมาใหม่ ดังต่อไปนี้

**ตารางที่ 4.3** ตัวแปรที่เกิดจาการคำนวณหาค่าเฉลี่ย

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| ARPU3Month | Numeric | จำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือน |
| NoOfCallVoice3Month | Numeric | จำนวนครั้งของการโทรออกเฉลี่ยต่อเดือน |
| MOUVoice3Month | Numeric | จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือน |
| MOUVas3Month | Numeric | จำนวนนาทีที่ใช้บริการเสริมเฉลี่ยต่อเดือน |
| MOUIR3Month | Numeric | จำนวนนาทีการใช้ Inter Roaming เฉลี่ยต่อเดือน |
| Duration3Month | Numeric | ระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้ง |

ตัวแปรในตารางที่ 4.3 ได้มาจากการคำนวณหาค่าเฉลี่ยของการใช้โทรศัพท์มือถือในช่วงเดือนเมษายน 2549 ถึง เดือนมิถุนายน 2549 กล่าวคือ

- จำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือน (ARPU3Month) คือ ค่าเฉลี่ยของตัวแปร Total ARPU 04, Total ARPU 05 และ Total ARPU 06
- จำนวนครั้งของการโทรออกเฉลี่ยต่อเดือน (NoOfCallVoice3Month) คือค่าเฉลี่ยของตัวแปร No of Call 04 (Voice), No of Call 05 (Voice) และ No of Call 06 (Voice)
- จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือน (MOUVoice3Month) คือค่าเฉลี่ยของตัวแปร MOU 04 (Voice), MOU 05 (Voice) และ MOU 06 (Voice)
- จำนวนนาทีที่ใช้บริการเสริมเฉลี่ยต่อเดือน (MOUVas3Month) คือค่าเฉลี่ยของตัวแปร MOU 04 (Vas), MOU 05 (Vas) และ MOU 06 (Vas)
- จำนวนนาทีการใช้ Inter Roaming เฉลี่ยต่อเดือน (MOUIR3Month) คือค่าเฉลี่ยของตัวแปร MOU 04 (IR), MOU 05 (IR) และ MOU 06 (IR)
- ระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้ง (Duration3Month) คือ จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือน (MOUVoice3Month) หารด้วย จำนวนครั้งของการโทรออกเฉลี่ยต่อเดือน (NoOfCallVoice3Month)

นอกจากนี้ยังมีการแปลงข้อมูล โดยได้ทำการจัดช่วงของข้อมูลที่เกี่ยวกับข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ ดังแสดงในตารางที่ 4.4

**ตารางที่ 4.4** แสดงตัวแปรที่เกิดจากการจัดช่วงของข้อมูล

| ชื่อตัวแปร | ชนิดของตัวแปร | ความหมายของตัวแปร |
|---|---|---|
| ARPUGroup | Text | จำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนที่แบ่งเป็นช่วง |
| MOUVoiceGroup | Text | จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนที่แบ่งเป็นช่วง |
| DurationGroup | Text | ระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งที่แบ่งเป็นช่วง |

ตัวแปรในตารางที่ 4.4 มีรายละเอียดเกี่ยวกับการจัดช่วงของตัวแปรดังนี้

- ARPUGroup เป็นตัวแปรที่เกิดจากการจัดช่วงข้อมูลของตัวแปร จำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือน (ARPU3Month) โดยที่

ถ้า ARPU3Month มีค่าเท่ากับ 0 แทนค่าด้วย “F ARPU = 0”

ถ้า ARPU3Month มีค่ามากกว่า 0 ถึง 500 แทนค่าด้วย “F >0-500”

ถ้า ARPU3Month มีค่ามากกว่า 500 ถึง 750 แทนค่าด้วย “E >500-750”

ถ้า ARPU3Month มีค่ามากกว่า 750 ถึง 1,200 แทนค่าด้วย “D >750-1,200”

ถ้า ARPU3Month มีค่ามากกว่า 1,200 ถึง 2,000 แทนค่าด้วย “C >1,200-2,000”

ถ้า ARPU3Month มีค่ามากกว่า 2,000 ถึง 3,000 แทนค่าด้วย “B >2,000-3,000”

ถ้า ARPU3Month มีค่ามากกว่า 3,000 ขึ้นไป แทนค่าด้วย “A >3,000”

- MOUVoiceGroup เป็นตัวแปรที่เกิดจากการจัดช่วงข้อมูลของตัวแปร จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือน (MOUVoice3Month) โดยที่

ถ้า MOUVoice3Month มีค่าเท่ากับ 0 แทนค่าด้วย “0 min”

ถ้า MOUVoice3Month มีค่ามากกว่า 0 ถึง 50 แทนค่าด้วย “>0-50”

ถ้า MOUVoice3Month มีค่ามากกว่า 50 ถึง 100 แทนค่าด้วย “>50-100”

ถ้า MOUVoice3Month มีค่ามากกว่า 100 ถึง 200 แทนค่าด้วย “>100-200”

ถ้า MOUVoice3Month มีค่ามากกว่า 200 ถึง 300 แทนค่าด้วย “>200-300”

ถ้า MOUVoice3Month มีค่ามากกว่า 300 ถึง 500 แทนค่าด้วย “>300-500”

ถ้า MOUVoice3Month มีค่ามากกว่า 500 ถึง 1,000 แทนค่าด้วย “>500-1,000”

ถ้า ARPU3Month มีค่ามากกว่า 1,000 ขึ้นไป แทนค่าด้วย “>1,000”

- DurationGroup เป็นตัวแปรที่เกิดจากการจัดช่วงข้อมูลของตัวแปร ระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้ง (Duration3Month) โดยที่

ถ้า Duration3Month มีค่าเท่ากับ 0 แทนค่าด้วย “0 min”

ถ้า Duration3Month มีค่ามากกว่า 0 ถึง 1 แทนค่าด้วย “>0-1”

ถ้า Duration3Month มีค่ามากกว่า 1 ถึง 2 แทนค่าด้วย “>1-2”

ถ้า Duration3Month มีค่ามากกว่า 2 ถึง 3 แทนค่าด้วย “>2-3”

ถ้า Duration3Month มีค่ามากกว่า 3 ถึง 5 แทนค่าด้วย “>3-5”

ถ้า Duration3Month มีค่ามากกว่า 5 ขึ้นไป แทนค่าด้วย “>5”

หลังจากที่ได้ทำการเตรียมข้อมูลโดยการคัดเลือกเอาเฉพาะข้อมูลที่เราต้องการนำมาใช้ในการวิเคราะห์ข้อมูลจริงๆ โดยผ่านกระบวนการกรองข้อมูล และการแปลงรูปแบบของข้อมูล ทำให้ได้ข้อมูลที่มีคุณภาพ พร้อมที่จะนำไปใช้ในขั้นตอนของการสร้างโมเดลนั้น ได้ทำการแบ่งชุดข้อมูลดังกล่าวออกเป็น 2 ชุด โดยข้อมูลชุดแรก คือ ข้อมูลที่ใช้สำหรับการสร้างโมเดล (Training data) จำนวน 194,233 ราย หรือคิดเป็นร้อยละ 80 ของข้อมูลทั้งหมด ส่วนข้อมูลชุดที่สอง คือ ข้อมูลที่ใช้สำหรับการทดสอบโมเดล (Testing data) ใช้เพื่อทดสอบว่าโมเดลที่สร้าง

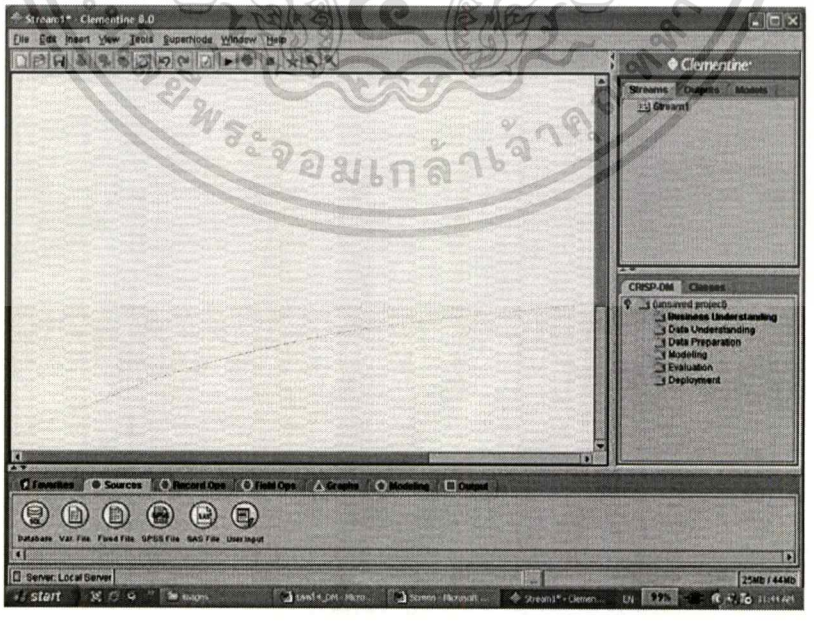
ขึ้นมานั้นมีความถูกต้องและน่าเชื่อถือหรือมากน้อยเพียงใด ซึ่งข้อมูลชุดนี้มีจำนวน 48,628 ราย หรือคิดเป็นร้อยละ 20 ของข้อมูลทั้งหมด

## 4.3 การทำดาต้าไมนิ่ง

ในขั้นตอนการทำดาต้าไมนิ่งของโครงการนี้ จะใช้โปรแกรมสำเร็จรูปที่ชื่อว่า Clementine ซึ่งเป็นโปรแกรมสำเร็จรูปทางด้านดาต้าไมนิ่งของบริษัทเอสพีเอสเอส (SPSS) โดยจะใช้เทคนิคของ Decision Tree และใช้ C5.0 Algorithm ในการสร้างโมเดลเพื่อทำนายลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ โดยขั้นตอนการทำไมนิ่งข้อมูล มีรายละเอียดดังต่อไปนี้

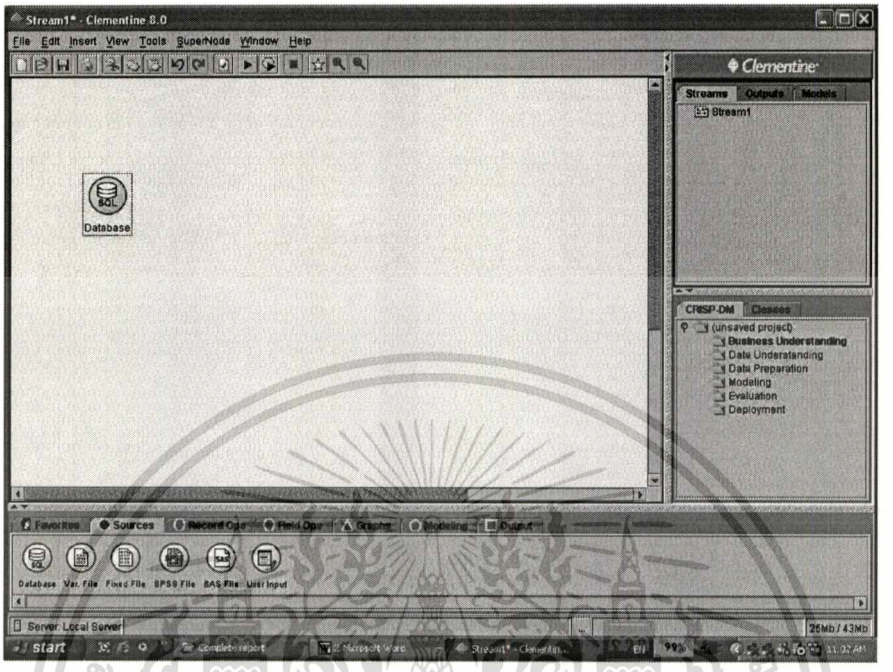
การนำเข้าข้อมูลที่ได้จากการเตรียมข้อมูล โดยโปรแกรม Clementine สามารถนำเข้าข้อมูลได้หลายชนิด เช่น ข้อมูลที่เป็น Worksheet จาก Excel, ข้อมูลที่เป็น Text File, ข้อมูลที่เป็น SPSS File และ SAS File ตลอดจนข้อมูลจากฐานข้อมูล เช่น Oracle, SQL Server และ Microsoft Access เป็นต้น โดยที่ Source Node ที่ใช้จะขึ้นกับชนิดของข้อมูลที่จะนำเข้า โดยข้อมูลที่ใช้สำหรับโครงการนี้อยู่ในรูปของตารางบนฐานข้อมูลของ Microsoft Access ดังนั้นการนำเข้าข้อมูลจึงใช้ Database Node โดยการติดต่อผ่าน ODBC (Open Database Connectivity) ซึ่งจะแสดงรายละเอียดและภาพประกอบของการนำเข้าข้อมูลดังนี้

เมื่อเปิดโปรแกรม Clementine ขึ้นมา และไปคลิ๊กที่ Sources จะแสดง Sources Node ที่ใช้สำหรับการนำข้อมูลเข้าของโปรแกรม Clementine ดังแสดงในรูปที่ 4.1



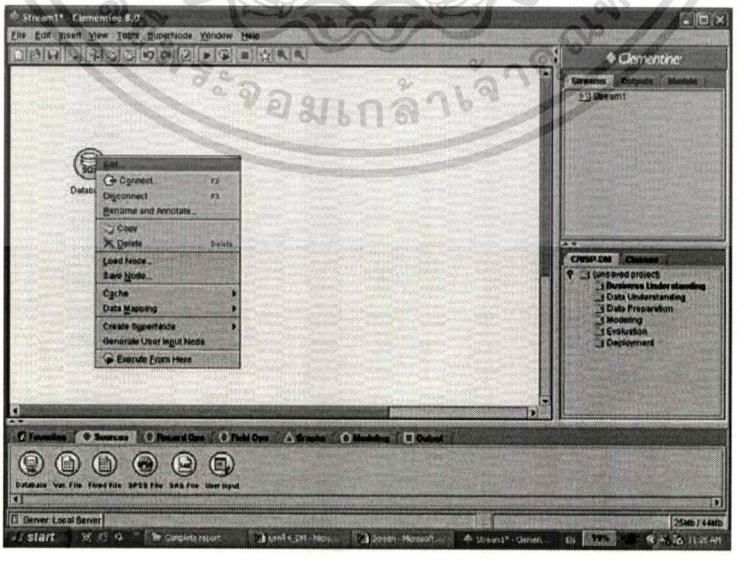

**รูปที่ 4.1** แสดง Sources Node ของโปรแกรม Clementine

จากนั้นคลิ๊กเลือกที่ Database Node และนำไปวางไว้ที่ Stream Canvas ดังรูปที่ 4.2



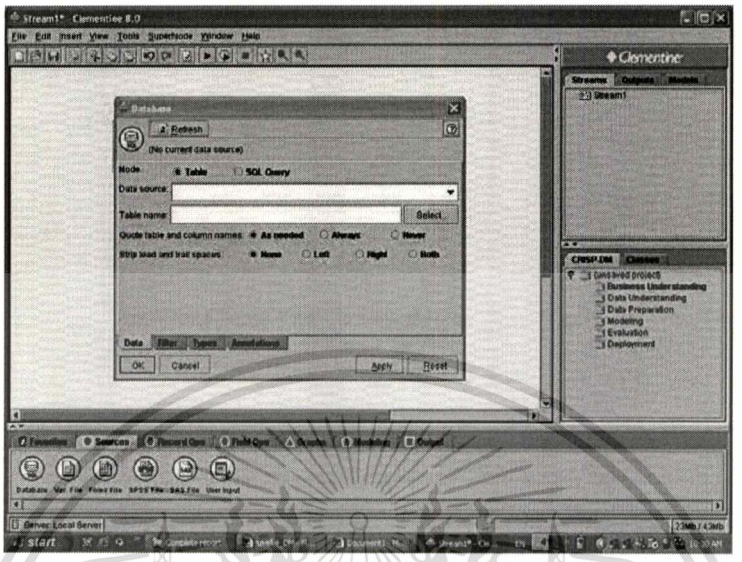

**รูปที่ 4.2** แสดง Database Node บน Stream Canvas

จากนั้นคลิ๊กขวาที่ Database Node บน Stream Canvas และเลือกคำสั่ง Edit เพื่อทำการเลือกข้อมูลจาก Access Database ดังรูปที่ 4.3



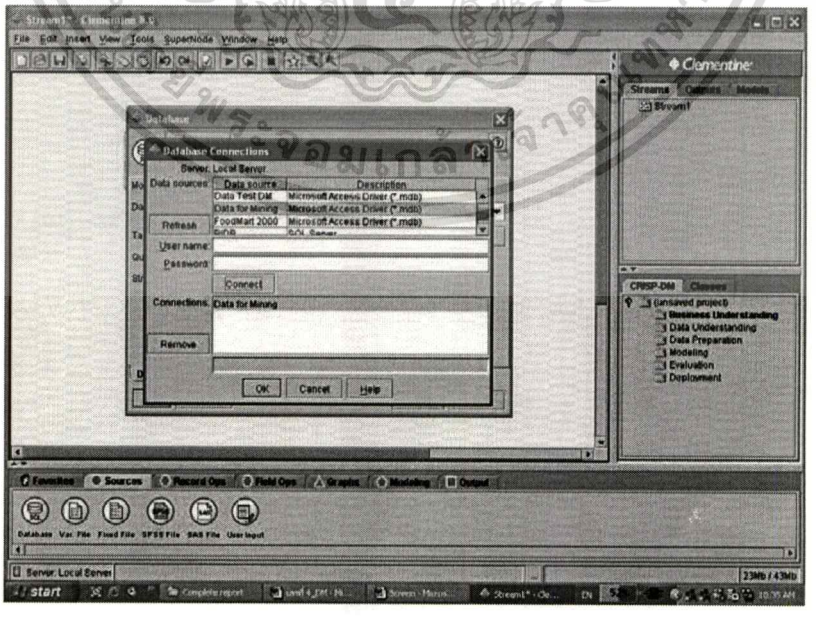

**รูปที่ 4.3** แสดงคำสั่ง Edit ของ Database Node

เมื่อเลือกคำสั่งEdit จะปรากฏหน้าจอ สำหรับการกำหนดค่าต่างๆของ Database Node ดังรูปที่ 4.4



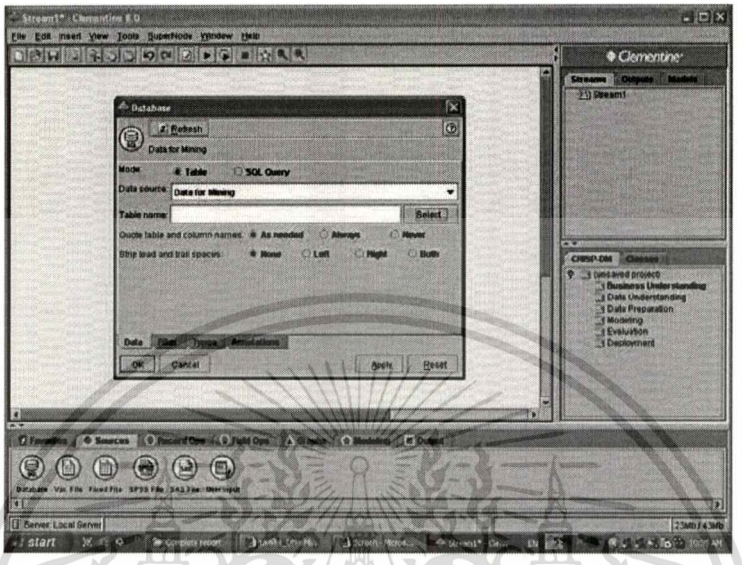
**รูปที่ 4.4** แสดง Database Node Dialog Box

จากนั้นทำการระบุ Data Source โดยการคลิ๊กที่เครื่องหมาย drop down แล้วเลือก Add New Database Connection จะปรากฏหน้าจอของ Database Connections เพื่อให้ทำการเลือก database ที่เราต้องการ ซึ่งในโครงการนี้มีชื่อว่า Data for Mining โดยการ connect ผ่าน ODBC ดังรูปที่ 4.5



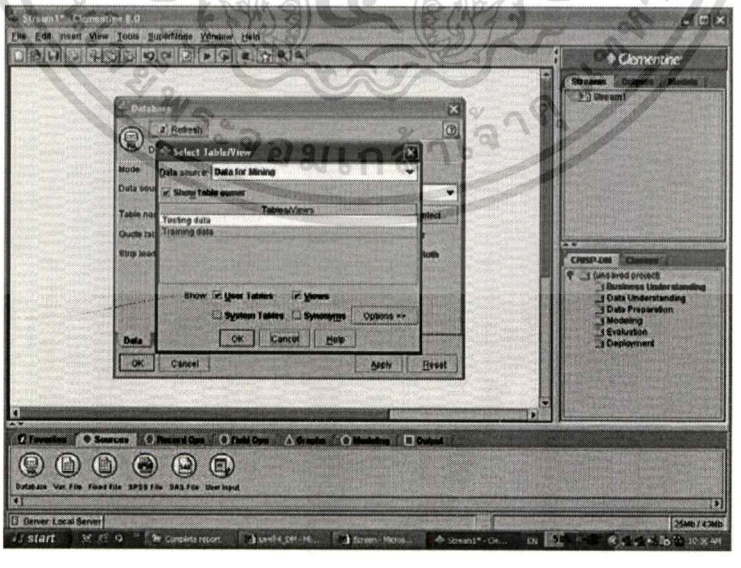
**รูปที่ 4.5** แสดงหน้าจอ Database Connections

จากรูปที่ 4.5 เมื่อคลิ๊ก OK จะกลับไปที่ Database Node dialog และจะปรากฏชื่อ Data for Mining ที่ตำแหน่งของ Data source ดังแสดงในรูปที่ 4.6



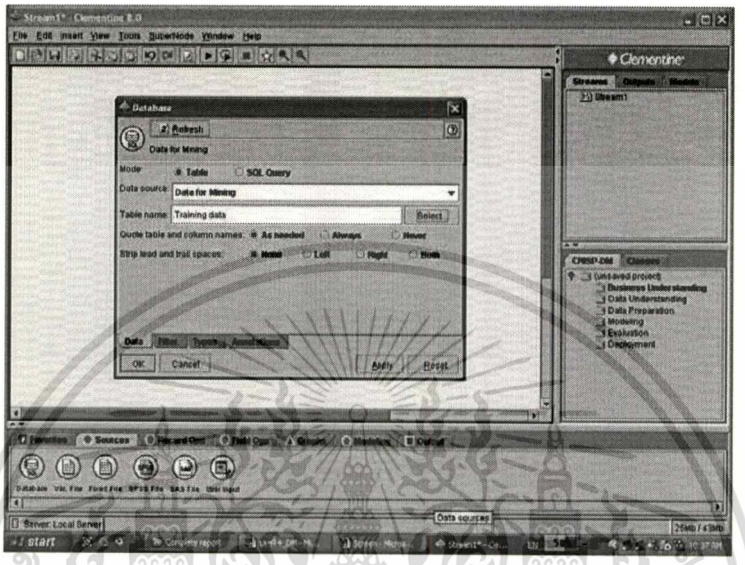
**รูปที่ 4.6** แสดง Database Node dialog ที่มีการระบุ Data source แล้ว

จากนั้น จะทำการเลือกชื่อ Table ของ database โดยการกดปุ่ม Select จะปรากฏ Select Table/View Dialog ดังรูปที่ 4.7



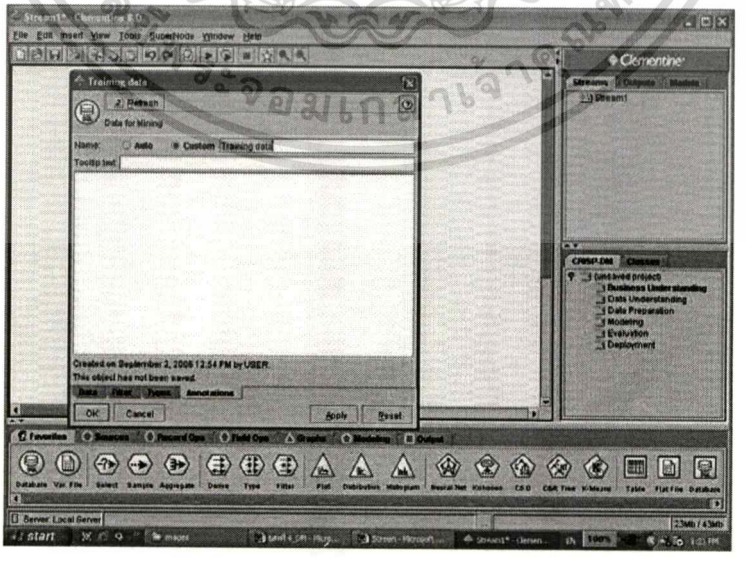
**รูปที่ 4.7** แสดง Select Table/View Dialog

จากนั้นเลือก Table ที่ต้องการ ในที่นี้เลือก table Training data จากนั้นคลิ๊ก OK จะกลับไปที่ Database Node dialog และจะปรากฏชื่อ Training data ที่ตำแหน่งของ Table name ดังแสดงในรูปที่ 4.8



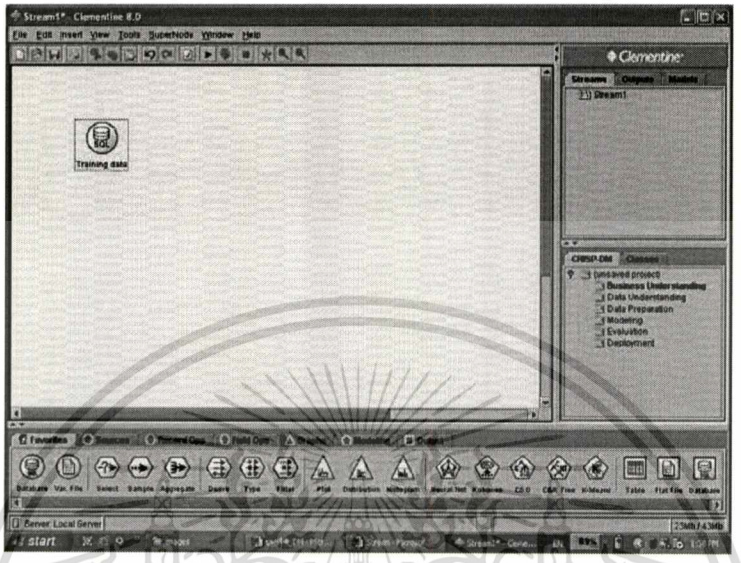
**รูปที่ 4.8** แสดง Database Node dialog ที่มีการระบุ Table name แล้ว

จากนั้นคลิ๊กที่ Tab ของ Annotations และไปตั้งชื่อ Database เป็น Training data ดังแสดงในรูปที่ 4.9



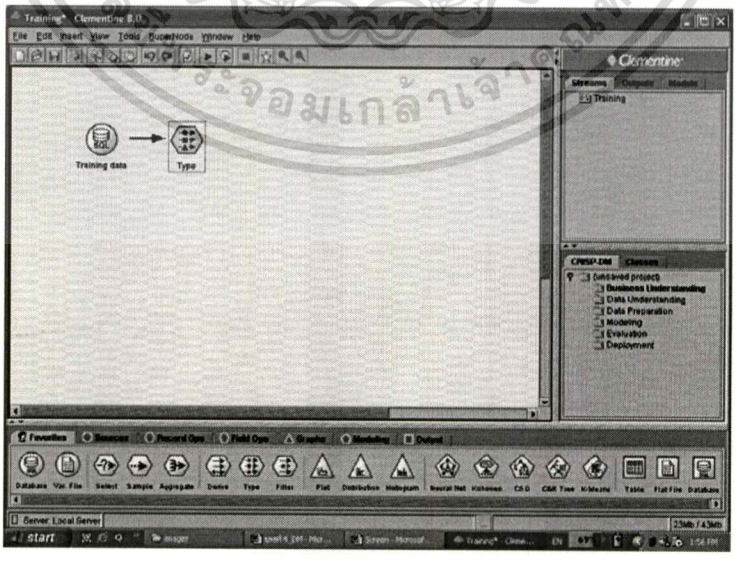
**รูปที่ 4.9** แสดงการตั้งชื่อของ Database Node

จากนั้นกดปุ่ม OK จะพบว่ามี Database Node ที่ชื่อ Training data ปรากฏอยู่บน Stream Canvas โดยมีการนำเข้าข้อมูลเรียบร้อยแล้วดังรูปที่ 4.10



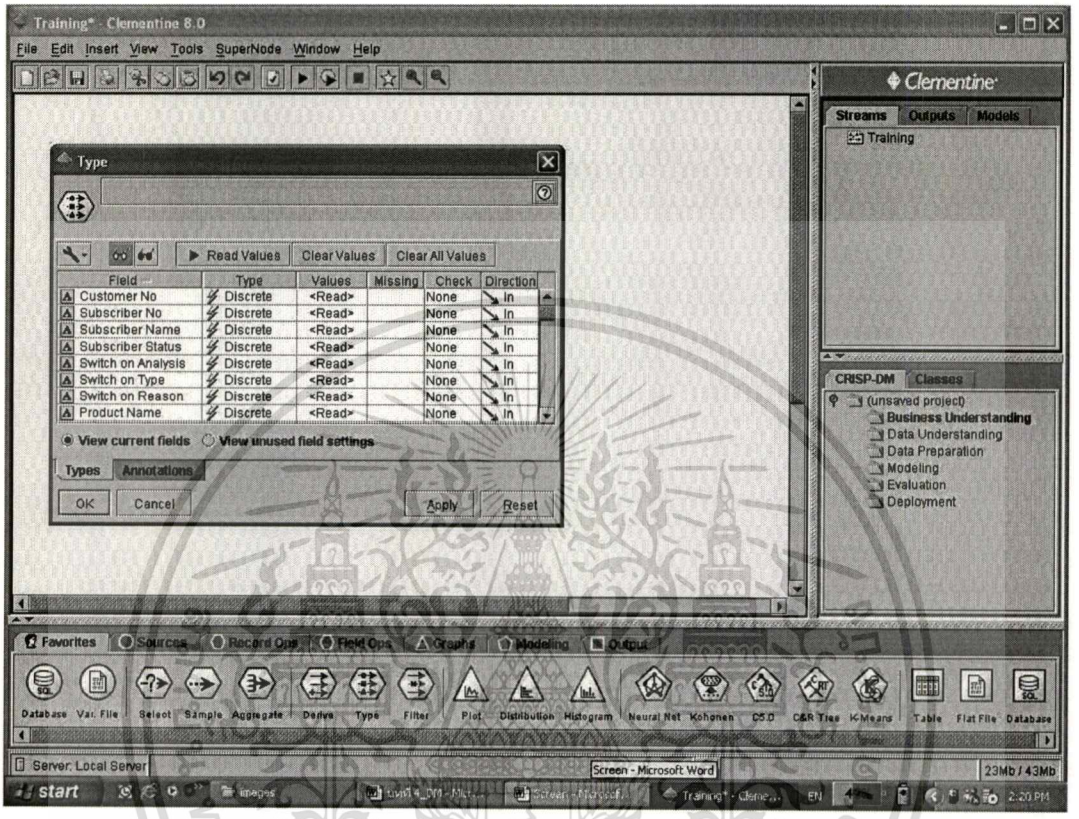

**รูปที่ 4.10** แสดง Database Node ที่มีการนำเข้าข้อมูลเรียบร้อยแล้ว บน Stream Canvas

ขั้นตอนต่อไปจะเป็นการกำหนดเกี่ยวกับประเภทของข้อมูลที่จะนำเข้าไปเพื่อทำการสร้างโมเดล ซึ่งจะใช้ Type Node เป็นตัวกำหนด โดยนำไปวางต่อจาก Database Node และทำการเชื่อมต่อเข้าด้วยกัน ดังรูปที่ 4.11




**รูปที่ 4.11** แสดงการเชื่อมต่อกันของ Database Node และ Type Node

การกำหนดคุณสมบัติของตัวแปร ทำได้ด้วยการคลิ๊กขวา แล้วเลือกคำสั่ง Edit จะปรากฏ Type Dialog ดังรูปที่ 4.12




**รูปที่ 4.12** แสดง Type Dialog

เราสามารถกำหนดคุณสมบัติต่างๆของตัวแปรได้จาก Type Dialog นี้โดยในข้อมูลชุดนี้ได้มีการกำนดเกี่ยวกับ Input variables และ Output variable ที่จะใช้ในการสร้างโมเดลดังตารางที่ 4.5

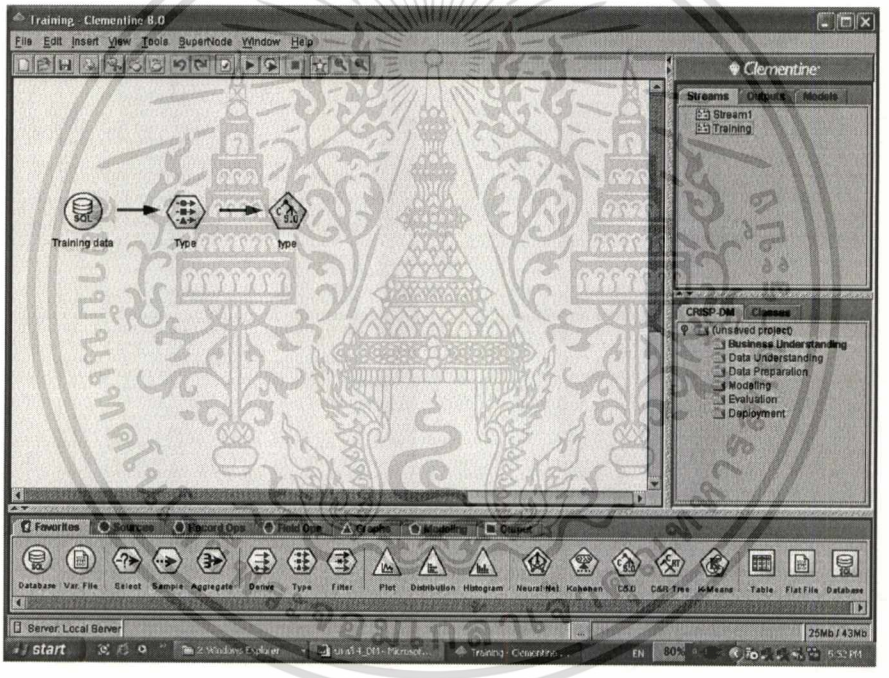
**ตารางที่ 4.5** แสดง Input variables และ Output variable ที่ใช้ในการสร้างโมเดล

| ชื่อตัวแปร | ความหมายของตัวแปร | Direction |
|---|---|---|
| PackageGroup(Current) | กลุ่มของโปรโมชั่นที่ใช้ในปัจจุบัน | Input |
| AOUMonth | อายุการใช้งานของลูกค้ามีหน่วยเป็นเดือน | Input |
| ARPU3Month | จำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือน | Input |
| NoOfCallVoice3Month | จำนวนครั้งของการโทรออกเฉลี่ยต่อเดือน | Input |
| MOUVoice3Month | จำนวนนาทีของการโทรออกเฉลี่ยต่อเดือน | Input |
| MOUVas3Month | จำนวนนาทีที่ใช้บริการเสริมเฉลี่ยต่อเดือน | Input |

**ตารางที่ 4.5 (ต่อ)**

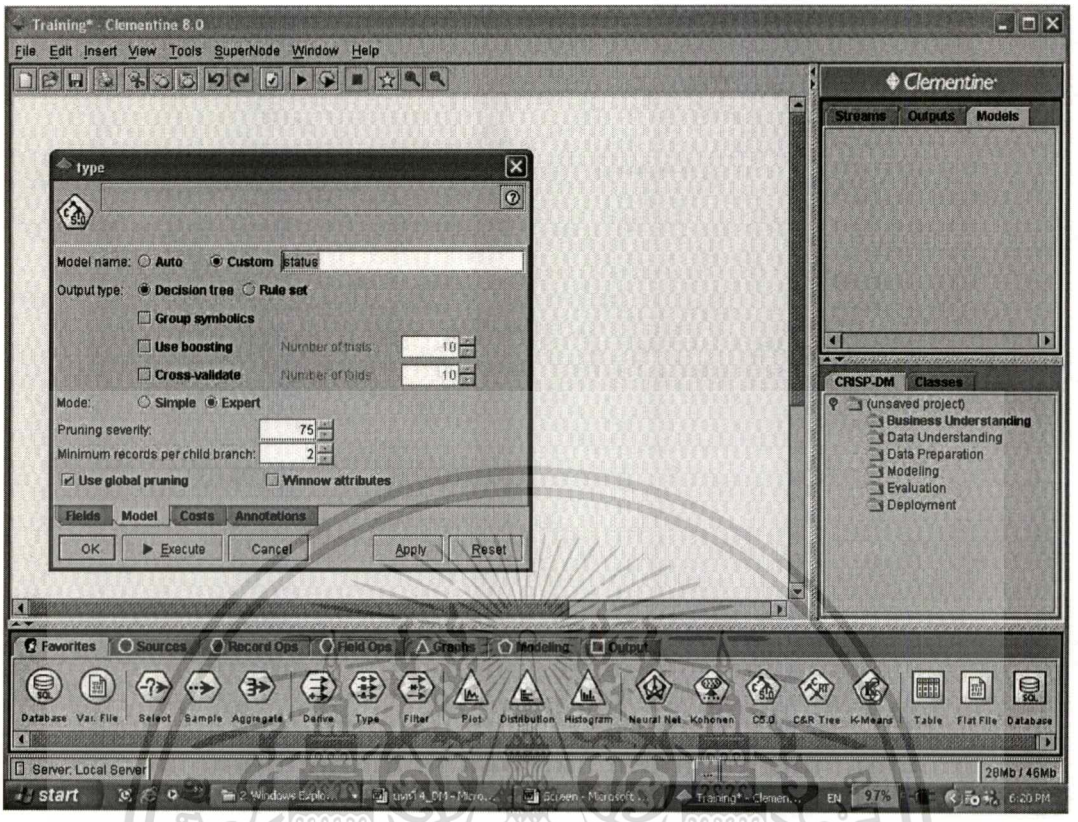
| ชื่อตัวแปร | ความหมายของตัวแปร | Direction |
|---|---|---|
| MOUIR3Month | จำนวนนาทีการใช้ Inter Roaming เฉลี่ยต่อเดือน | Input |
| Duration3Month | ระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้ง | Input |
| Type | เป็นตัวแปรที่ระบุสถานะของลูกค้า | Output |

เมื่อได้กำหนดตัวแปรต่างๆ ตามตารางที่ 4.5 เรียบร้อยแล้ว จะทำการสร้างโมเดลโดยใช้ C5.0 Algorithm ทำได้โดยการคลิ๊กเลือก C5.0 Node ไปเชื่อมต่อกับ Type Node ดังรูปที่ 4.13



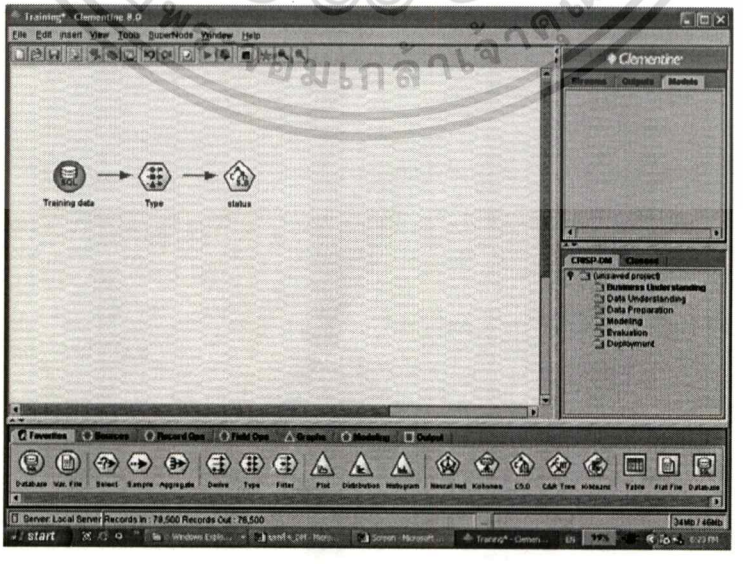

**รูปที่ 4.13** แสดงการเชื่อมต่อของ Type Node และ C5.0 Node

จากนั้นคลิ๊กขวาที่โหนดของ C5.0 แล้วเลือกคำสั่ง Edit จะปรากฏหน้าจอให้ทำการแก้ไขรายละเอียดต่างๆของโหนด C5.0 โดยในที่นี้มีการแก้ไข Model name เป็น status ซึ่งถ้าหากไม่มีการแก้ไขเกี่ยวกับ Model name ค่าเริ่มต้นที่ถูกกำหนดให้จะเป็นชื่อเดียวกับตัวแปรที่เราใช้เป็นตัวแปรในการทำนายซึ่งในที่นี้คือตัวแปร Type จากนั้นกำหนด Output type เป็น Decision tree และกำหนด Mode เป็น Expert ดังรูปที่ 4.14



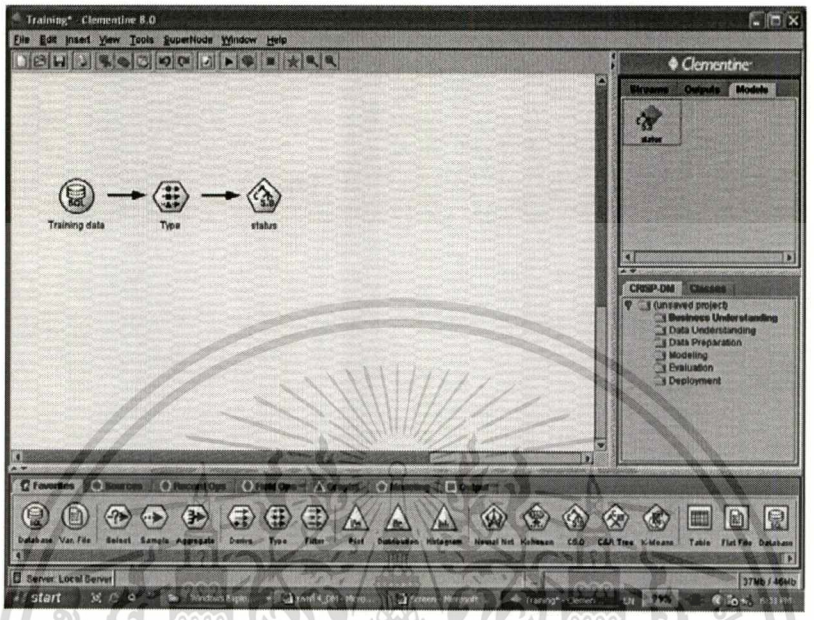

**รูปที่ 4.14** แสดงการกำนดค่าต่างๆ ของ C5.0 Node

จากนั้นกดปุ่ม Execute โปรแกรมจะทำการรันโมเดล ซึ่งขณะที่โมเดลกำลังรันอยู่นั้นจะมีสีเขียวที่เครื่องหมายลูกศรที่ทำหน้าที่เชื่อมต่อระหว่างโหนด ดังรูปที่ 4.15



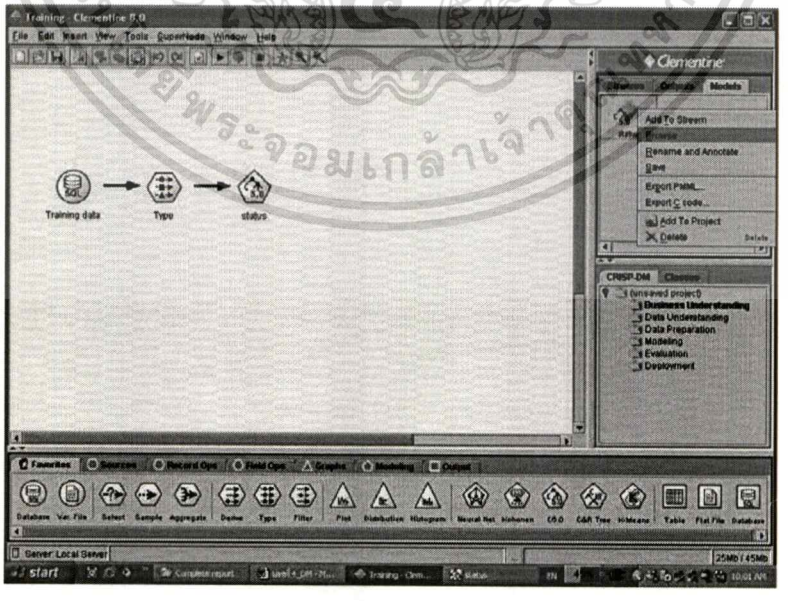

**รูปที่ 4.15** แสดงสถานะของโปรแกรมขณะกำลังรันข้อมูล

เมื่อโปรแกรมทำการรันโมเดลเสร็จแล้ว จะทำการสร้างโมเดลโหนดที่เกิดจากการรันโมเดลไปเก็บเอาไว้ที่ Tab ของ Models ทางด้านขวามือ ดังรูปที่ 4.16



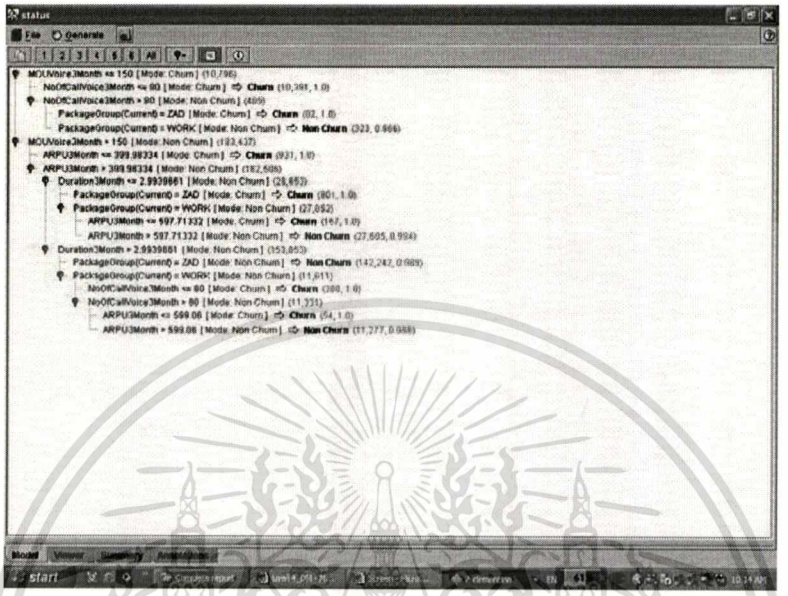

**รูปที่ 4.16** แสดงโมเดลโหนดที่เกิดจากการรันโมเดลของโปรแกรม Clementine

การดูรายละเอียดของโมเดลต้องไปคลิ๊กขวาที่โมเดลโหนด แล้วเลือกคำสั่ง Browse ดังรูปที่ 4.17



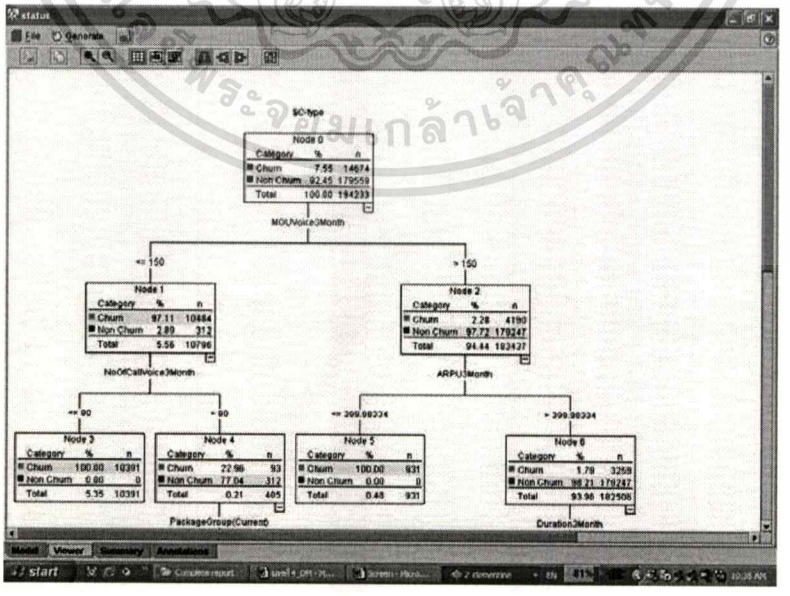

**รูปที่ 4.17** แสดงคำสั่ง Browse ของโมเดลโหนด

เมื่อเลือกคำสั่ง Browse แล้วจะปรากฏหน้าจอที่แสดงรายละเอียดเกี่ยวกับโมเดลที่รันได้เป็นลักษณะโครงสร้างแบบ Branch ดังรูปที่ 4.18



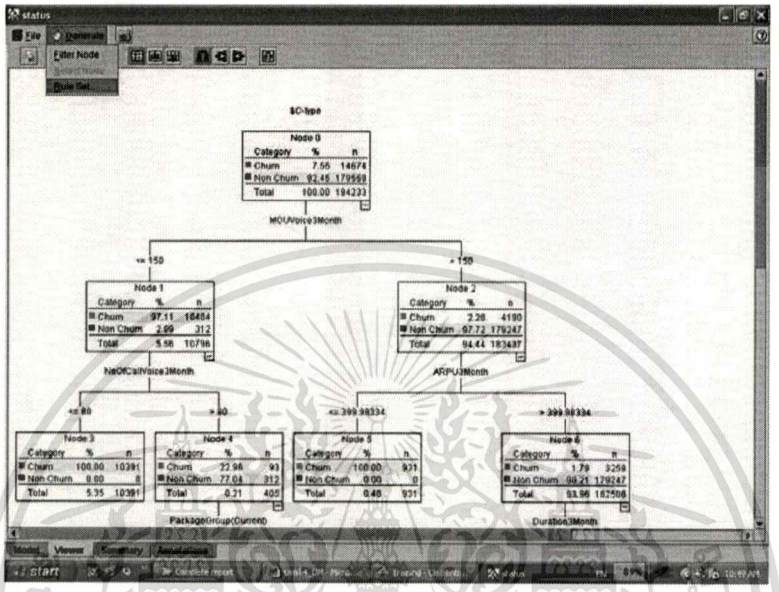
**รูปที่ 4.18** แสดงรายละเอียดของโมเดลในลักษณะโครงสร้างแบบ Branch

จากรูปที่ 4.18 เมื่อคลิ๊กเลือกที่ Tab ของ Viewer จะเห็นรายละเอียดของโมเดลในลักษณะของแผนภาพต้นไม้ ดังรูปที่ 4.19



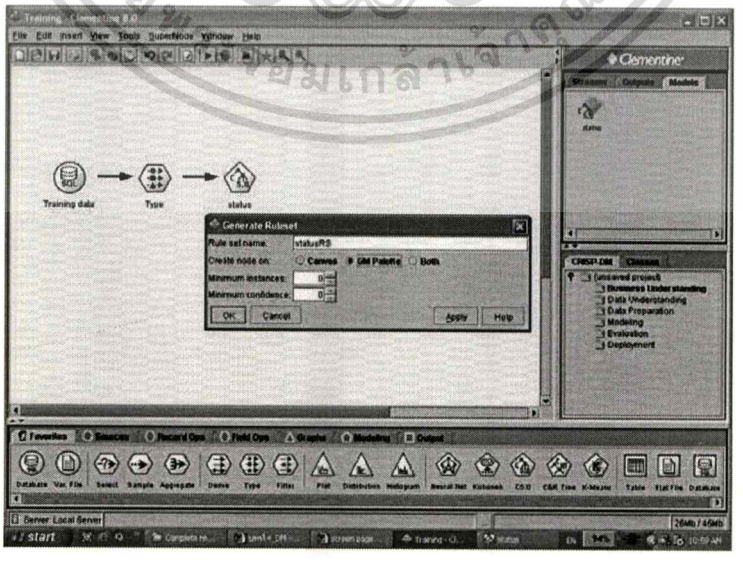

**รูปที่ 4.19** แสดงรายละเอียดของโมเดลในลักษณะแผนภาพต้นไม้

จากรูปที่ 4.18 และรูปที่ 4.19 ผลลัพธ์ที่ได้มีรายละเอียดค่อนข้างมาก ตัวอย่างเช่น แผนภาพต้นไม้ที่ได้นั้นค่อนข้างใหญ่ ไม่สามารถมองเห็นได้ภายในหน้าเดียวกัน ดังนั้นจึงสร้างเป็น Rule Set เพื่อนำไปวิเคราะห์และประมวลผลต่อไป โดยใช้คำสั่ง Generate …Rule Set ดังรูปที่ 4.20



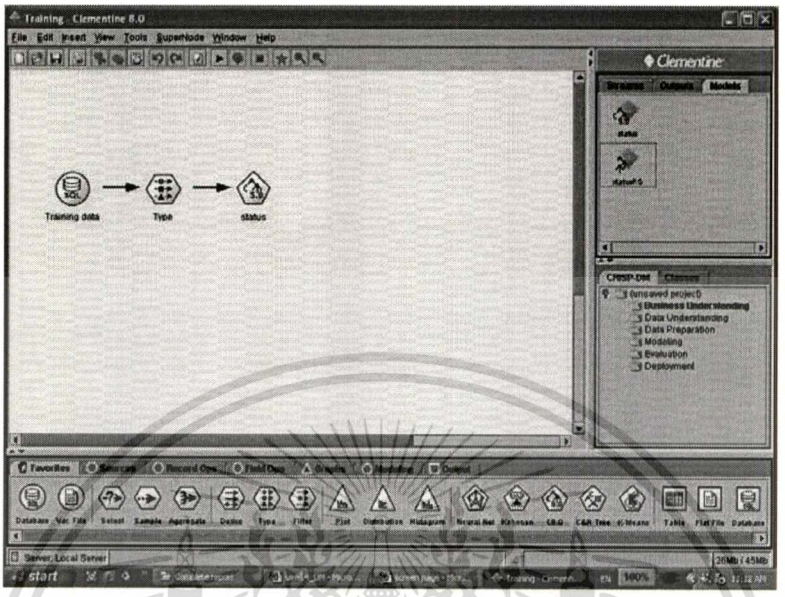

**รูปที่ 4.20** แสดงคำสั่งการสร้าง Rule Set

เมื่อใช้คำสั่ง Generate…Rule Set จะกลับไปที่หน้าของ Stream และมี Generate Ruleset Dialog Box ขึ้นมา ให้คลิ๊กเลือกที่ GM Palette ดังรูปที่ 4.21



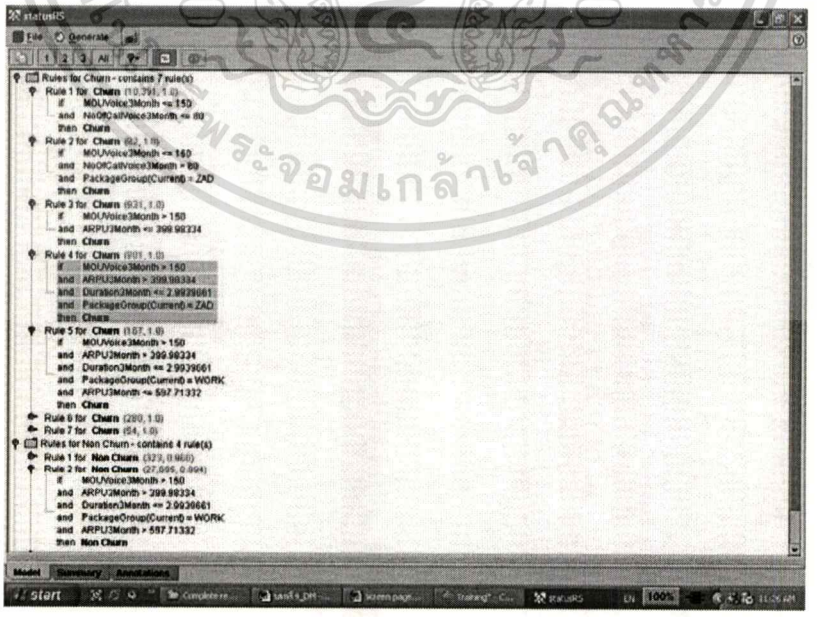

**รูปที่ 4.21** แสดง Generate Ruleset Dialog Box

จากนั้นกด OK จะปรากฏ Rule Set โหนดที่ Tab ของ Models ทางด้านขวามือ ดังรูปที่ 4.22



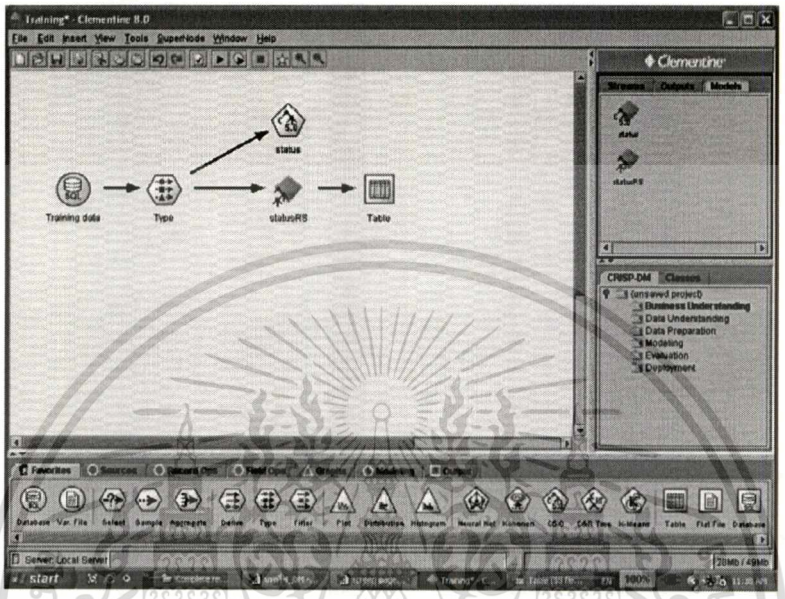

**รูปที่ 4.22** แสดง Rule Set โหนดบนTab ของ Models

เมื่อคลิกขวาที่ Rule Set โหนดแล้วเลือกคำสั่ง Browse จะได้หน้าจอที่แสดงรายละเอียดของ Rule Set ในลักษณะของ If...then ดังรูปที่ 4.23



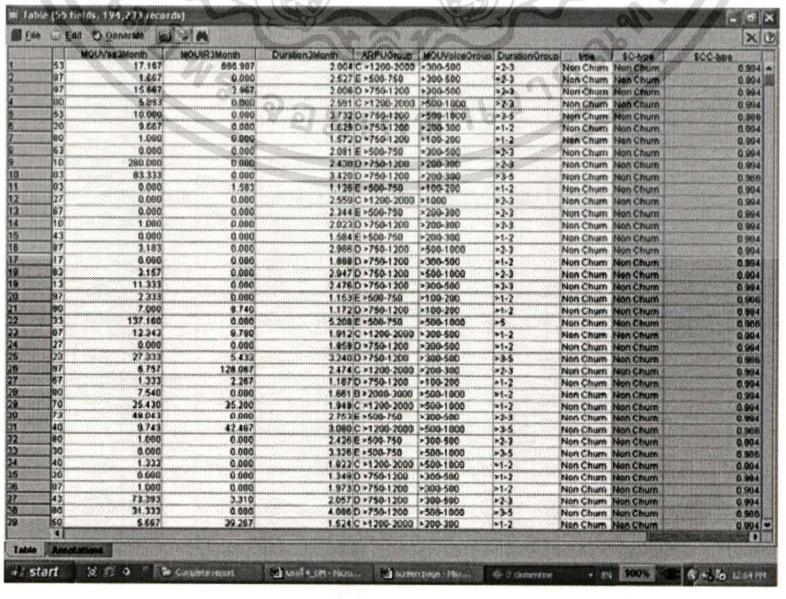

**รูปที่ 4.23** แสดงรายละเอียด Rule Setในรูปของ If...then

กลับไปที่ Stream Canvas นำ Rule Set Node ไปเชื่อมต่อกับ Database Node และ Type Node ก่อนจากนั้นจึงนำ Table Node ไปเชื่อมต่อกับ Rule Set Node เพื่อใช้เป็น Output สำหรับการแสดงตัวแปรใหม่ที่เกิดขึ้นจากการรันโมเดล ดังรูปที่ 4.24



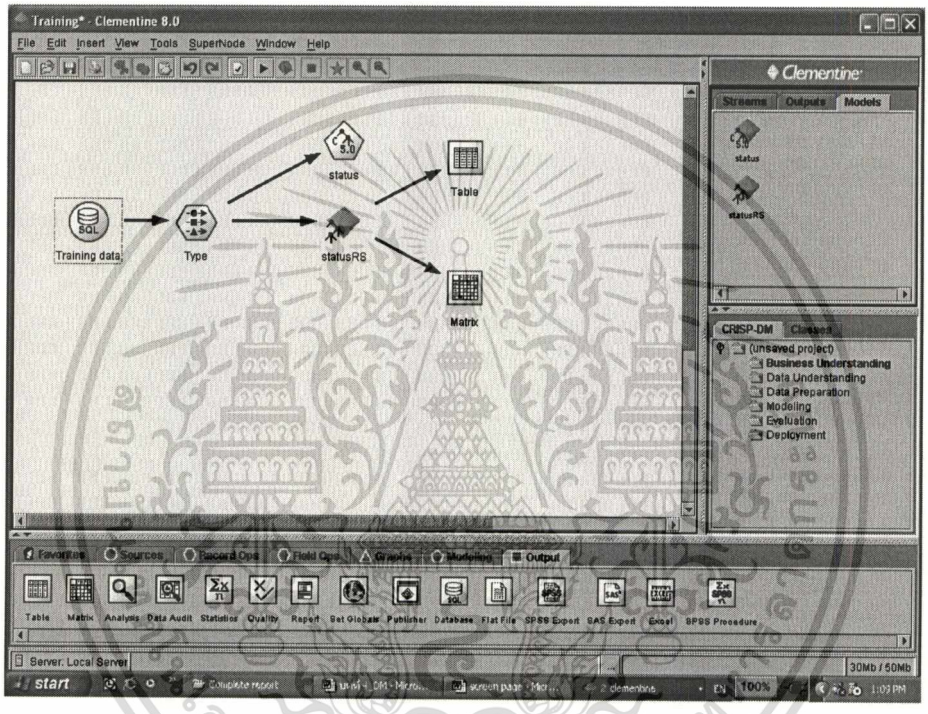

**รูปที่ 4.24** แสดงการเชื่อมต่อ Rule Set Node กับ Node ต่างๆ บน Stream

จากนั้นคลิ๊กขวาที่ Table Node แล้วเลือกคำสั่ง Execute จะได้ผลลัพธ์ ดังรูปที่ 4.25



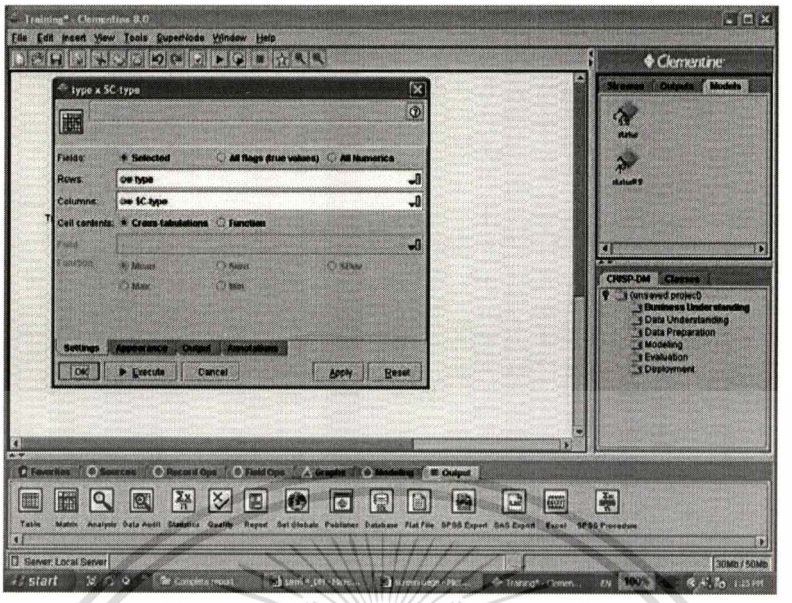
**รูปที่ 4.25** แสดงตัวแปรใหม่ที่เกิดขึ้นจากการรันโมเดล

ตัวแปรใหม่ที่เกิดขึ้น ได้แก่ตัวแปร $C-type และ $CC-type โดยที่ $C-type คือ ตัวแปรที่เกิดจากการทำนายว่าลูกค้าจะเป็น Churn หรือ Non Churn และตัวแปร $CC-type คือระดับความน่าเชื่อถือของการทำนาย ซึ่งเราสามารถทดสอบความน่าเชื่อถือของโมเดลได้โดยการนำตัวแปรที่ได้จากการทำนาย ($C-type) มาเปรียบเทียบกับค่าของข้อมูลที่เกิดขึ้นจริง (type) ว่าผลการทำนายที่ได้มีความถูกต้องและน่าเชื่อถือมากน้อยเพียงใด โดยการใช้ Matrix Node ไปเชื่อมต่อกับ Rule Set Node ดังรูปที่ 4.26



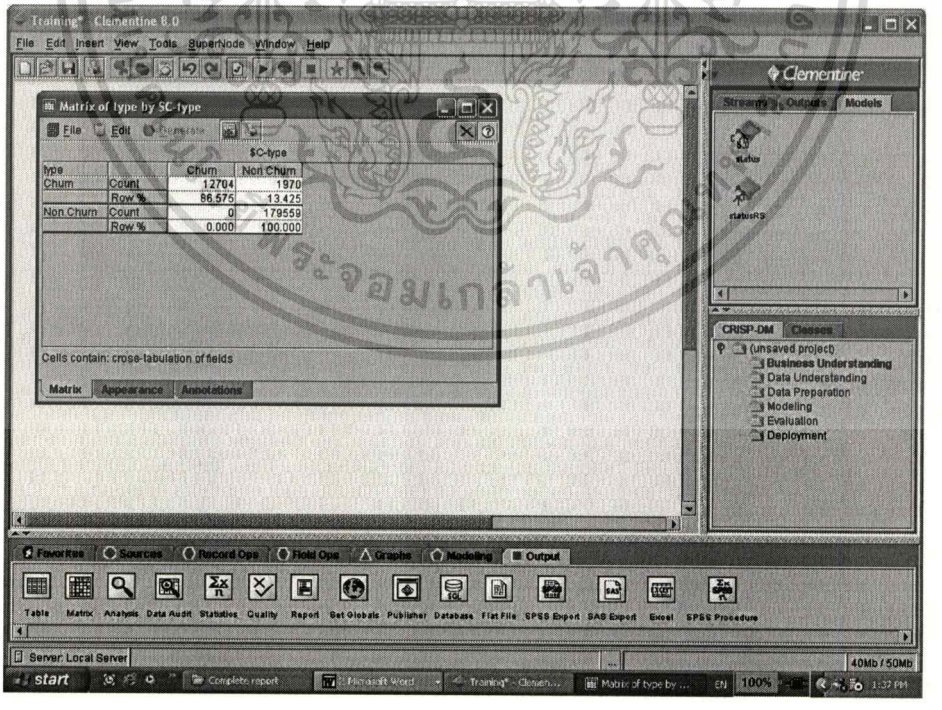

**รูปที่ 4.26** แสดงการเชื่อมต่อของ Matrix Node

จากนั้นคลิ๊กขวาที่ Matrix Node แล้วเลือกคำสั่ง Edit จะปรากฏ Matrix Dialog Box ขึ้นมาให้กำหนดตัวแปร type ไว้ที่ตำแหน่งของ Rows และกำหนดตัวแปร $C-type ไว้ที่ตำแหน่งของ Columns ดังรูปที่ 4.27



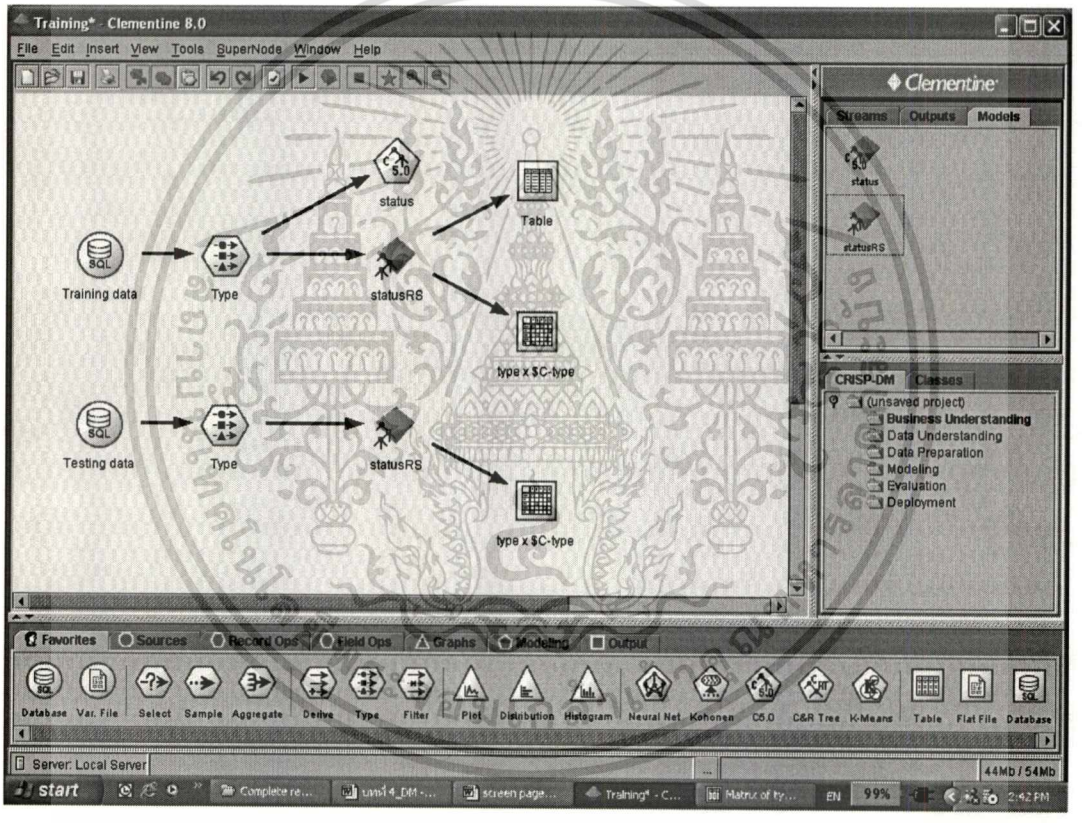
**รูปที่ 4.27** แสดงการกำหนดค่าตัวแปรของ Matrix Node

จากนั้นไปที่ Tab ของ Appearance คลิ๊กเลือก Percentage of row เสร็จแล้วกดปุ่ม Execute จะได้ผลลัพธ์ดังรูปที่ 4.28



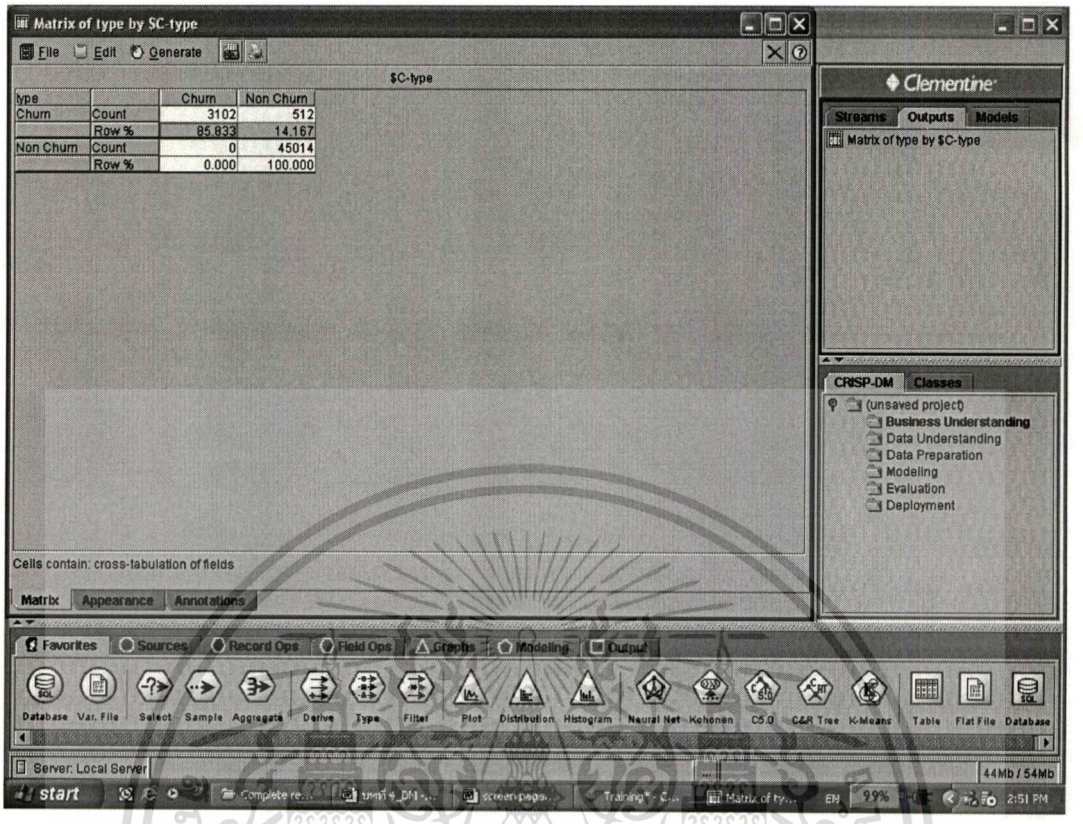

**รูปที่ 4.28** ผลลัพธ์แสดงความถูกต้องของการทำนายจาก Training data

จากรูปที่ 4.28 อธิบายได้ว่า โมเดลที่ได้นั้นสามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ได้ถูกต้อง 12,704 ราย หรือคิดเป็น 86.6% ขณะที่ทำนายผิดพลาดไปเป็นจำนวน 1,970 ราย หรือคิดเป็น 13.4% ซึ่งผลที่ได้นั้นเป็นการทดสอบโดยใช้ข้อมูลที่เป็น Training data ต่อจากนี้ จะทดสอบความถูกต้องของโมเดลจากชุดข้อมูลที่เป็น Testing data แล้วนำผลลัพธ์ที่ได้มาเปรียบเทียบกันว่าโมเดลที่ได้นั้นมีความถูกต้องและน่าเชื่อถือมากน้อยเพียงใดก่อนที่จะนำโมเดลที่ได้ไปใช้งาน ซึ่งขั้นตอนที่ทำการทดสอบกับชุดข้อมูลของ Testing data แสดงดังรูปที่ 4.29 โดยรายละเอียดเกี่ยวกับโหนดต่างๆจะเหมือนกับการสร้างโมเดลที่ได้อธิบายไปแล้ว เพียงแต่ข้อมูลที่นำมาใช้เปลี่ยนจาก Training data เป็น Testing data




**รูปที่ 4.29** แสดงการทดสอบโมเดลด้วยชุดข้อมูล Testing data

เมื่อคลิ๊กขวาที่ Matrix Node ของชุดข้อมูล Testing data แล้วเลือกคำสั่ง Execute จะได้ผลลัพธ์ดังรูปที่ 4.30




**รูปที่ 4.30** ผลลัพธ์แสดงความถูกต้องของการทำนายจาก Testing data

จากรูปที่ 4.30 อธิบายได้ว่า โมเดลที่ได้นั้นสามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ได้ถูกต้อง 3,102 ราย หรือคิดเป็น 85.8% ขณะที่ทำนายผิดพลาดไปเป็นจำนวน 512 ราย หรือคิดเป็น 14.2%

เมื่อเปรียบเทียบผลลัพธ์ที่ได้จากการทำนายของทั้ง 2 ชุดข้อมูลสามารถสรุปได้ว่าโมเดลที่ได้นั้นสามารถนำไปใช้ได้ เนื่องจากมีความถูกต้องมากกว่า 85%

## 4.4 การวิเคราะห์ผลลัพธ์

จากการทำดาต้าไมนิ่งในหัวข้อที่แล้ว ทำให้ได้โมเดลที่จะนำไปทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ (Churn Prediction Model) ซึ่งแบ่งเป็น โมเดลของ Churn 7 โมเดล และ Non Churn 4 โมเดล แต่ก่อนที่จะนำโมเดลไปใช้ได้นั้นต้องทำการวิเคราะห์และแปลความหมายของโมเดลก่อน โดยในโครงการนี้มีรายละเอียดของโมเดลทั้งหมด ที่จะแสดงในรูปแบบของ If-Then Rule ได้ดังต่อไปนี้

**Rules for Churn - contains 7 rule(s)**

***Rule 1 for Churn (10,391, 1.0)***

***if MOUVoice3Month <= 150 and NoOfCallVoice3Month <= 80 then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 80 ครั้ง

***Rule 2 for Churn (82, 1.0)***

***if MOUVoice3Month <= 150 and NoOfCallVoice3Month > 80***

***and PackageGroup(Current) = ZAD then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้ง และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น ZAD

***Rule 3 for Churn (931, 1.0)***

***if MOUVoice3Month > 150 and ARPU3Month <= 399.98334 then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนต่ำกว่า 400 บาท

***Rule 4 for Churn (801, 1.0)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***

***and Duration3Month <= 2.9939661 and PackageGroup(Current) = ZAD***

***then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนมากกว่า 400 บาท และระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งต่ำกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น ZAD

***Rule 5 for Churn (167, 1.0)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***

***and Duration3Month <= 2.9939661 and PackageGroup(Current) = WORK***

***and ARPU3Month <= 597.71332 then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนอยู่ระหว่าง 400 บาทถึง 598 บาท และมีระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งต่ำกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK

***Rule 6 for Churn (280, 1.0)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***
***and Duration3Month > 2.9939661 and PackageGroup(Current) = WORK***
***and NoOfCallVoice3Month <= 80 then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนมากกว่า 400 บาท และมีระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งมากกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 80 ครั้ง

***Rule 7 for Churn (54, 1.0)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***
***and Duration3Month > 2.9939661 and PackageGroup(Current) = WORK***
***and NoOfCallVoice3Month > 80 and ARPU3Month <= 599.06***
***then Churn***

- หมายความว่าลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนอยู่ระหว่าง 400 บาทถึง 600 บาท และมีระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งมากกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้ง

**Rules for Non Churn - contains 4 rule(s)**

***Rule 1 for Non Churn (323, 0.966)***

***if MOUVoice3Month <= 150 and NoOfCallVoice3Month > 80***
***and PackageGroup(Current) = WORK then Non Churn***

- หมายความว่าลูกค้าจะไม่ยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการ

โทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้งและโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK

***Rule 2 for Non Churn (27,685, 0.994)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***

***and Duration3Month <= 2.9939661 and PackageGroup(Current) = WORK***

***and ARPU3Month > 597.71332 then Non Churn***

- หมายความว่าลูกค้าจะไม่ยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนอยู่ระหว่าง 400 บาทถึง 598 บาท และมีระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งมากกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK

***Rule 3 for Non Churn (142,242, 0.989)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***

***and Duration3Month > 2.9939661 and PackageGroup(Current) = ZAD***

***then Non Churn***

- หมายความว่าลูกค้าจะไม่ยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนมากกว่า 400 บาท และระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งมากกว่า 3 นาที และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น ZAD

***Rule 4 for Non Churn (11,277, 0.986)***

***if MOUVoice3Month > 150 and ARPU3Month > 399.98334***

***and Duration3Month > 2.9939661 and PackageGroup(Current) = WORK***

***and NoOfCallVoice3Month > 80 and ARPU3Month > 599.06***

***then Non Churn***

- หมายความว่าลูกค้าจะไม่ยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนมากกว่า 150 นาที และมีจำนวนเงินค่าใช้บริการเฉลี่ยต่อเดือนอยู่มากกว่า 400 บาทและ 600 บาท โดยมีระยะเวลาเฉลี่ยของการใช้โทรศัพท์ต่อครั้งมากกว่า 3 นาที และมีโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้ง

สิ่งที่ได้จากการแปลความหมายของโมเดล ก็คือความรู้เกี่ยวกับพฤติกรรมการใช้งานโทรศัพท์ของกลุ่มลูกค้า Churn และ Non Churn โดยเราจะนำความรู้ที่ได้นี้ไปใช้เป็นแนวทางใน

การกำหนดกลยุทธ์ทางการตลาด เพื่อเป็นการป้องกันไม่ให้ลูกค้ายกเลิกการใช้บริการโทรศัพท์เคลื่อนที่

## 4.5 การนำความรู้ที่ได้ไปใช้งาน

จากการวิเคราะห์และแปลความหมายโมเดลของ Churn และ Non Churn ทำให้เราทราบถึงพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้าทั้ง 2 กลุ่ม และสามารถนำความรู้ที่ได้ไปประยุกต์ใช้กับกิจกรรมทางการตลาดขององค์กรเพื่อป้องกันไม่ให้ลูกค้ายกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ กล่าวคือ นำความรู้ที่ได้ไปประยุกต์ใช้กับโครงการ Right Package Advisory ซึ่งโครงการนี้เป็นการโทรไปหาลูกค้าเพื่อแนะนำโปรโมชั่นที่เหมาะสมให้กับลูกค้า เราสามารถนำความรู้ที่ได้จากโมเดลที่ 2 ของ Churn ที่ทำให้เราทราบว่า ***ลูกค้ามีแนวโน้มที่จะยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้ง และโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น ZAD*** ในขณะที่ โมเดลที่ 1 ของ Non Churn บอกให้เราทราบว่า ***ลูกค้าจะไม่ยกเลิกการใช้บริการ ถ้าลูกค้ามีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้งและโปรโมชั่นที่ใช้อยู่ในปัจจุบันเป็น WORK***

ดังนั้นเราจึงนำความรู้ที่ได้จากโมเดลทั้งสองไปใช้ประโยชน์ โดยการโทรไปหาลูกค้าที่ใช้โปรโมชั่น ZAD อยู่ในปัจจุบัน และมีพฤติกรรมการใช้โทรศัพท์เคลื่อนที่ตามที่โมเดลระบุ คือ มีจำนวนนาทีของการโทรออกเฉลี่ยต่อเดือนน้อยกว่าหรือเท่ากับ 150 นาที และมีจำนวนครั้งของการโทรออกเฉลี่ยต่อเดือนมากกว่า 80 ครั้ง โดยการแนะนำให้ลูกค้าเปลี่ยนมาใช้โปรโมชั่น WORK ซึ่งเป็นโปรโมชั่นที่เหมาะสมกับลูกค้ามากกว่าโปรโมชั่น ZAD

จากการที่เรานำความรู้ที่ได้ไปใช้ในโครงการ Right Package Advisory ทำให้ลูกค้าเกิดความพึงพอใจ เนื่องจากการที่ลูกค้าเปลี่ยนจากโปรโมชั่น ZAD มาเป็นโปรโมชั่น WORK นั้นทำให้ลูกค้าสามารถประหยัดค่าใช้จ่ายได้มากขึ้นถึง 20% ดังนั้นลูกค้าจึงเกิดความพึงพอใจและไม่ยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ ทำให้องค์กรสามารถรักษาฐานลูกค้าเอาไว้ได้

# บทที่ 5
# บทสรุปและข้อเสนอแนะ

ในบทนี้จะกล่าวสรุปถึงผลการวิเคราะห์ที่ได้จากการนำข้อมูลเกี่ยวกับการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้าทั้งที่ยกเลิกการใช้บริการไปแล้ว และยังคงใช้บริการอยู่ มาวิเคราะห์โดยผ่านกระบวนการของการทำดาต้าไมนิ่ง ว่าผลลัพธ์ที่ได้นั้น มีลักษณะเป็นอย่างไร มีความถูกต้องและน่าเชื่อถือเพียงใด และสามารถนำไปใช้ประโยชน์ได้อย่างไร ตลอดจนข้อเสนอแนะที่ได้จากการศึกษาในโครงการนี้

## 5.1 บทสรุป

จากการศึกษาโครงการฉบับนี้ พบว่าผลลัพธ์ที่ได้จากการสร้างโมเดลโดยใช้เทคนิคของ Decision Tree ด้วย C5.0 Algorithm ของโปรแกรม Clementine นั้นสามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ได้ถูกต้องมากกว่า 85% โดยได้ทำการทดสอบกับข้อมูลทั้ง 2 ชุด คือ Training data และ Testing data

ข้อมูลชุดที่เป็น Training data สามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ได้ถูกต้อง 12,704 ราย หรือคิดเป็น 86.6% ขณะที่ทำนายผิดพลาดไปเป็นจำนวน 1,970 ราย หรือคิดเป็น 13.4% ส่วนชุดข้อมูลที่เป็น Testing data สามารถทำนายลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ได้ถูกต้อง 3,102 ราย หรือคิดเป็น 85.8% ขณะที่ทำนายผิดพลาดไปเป็นจำนวน 512 ราย หรือคิดเป็น 14.2%

จากการใช้โปรแกรม Clementine โดยใช้เทคนิคของ Decision Tree และใช้ C5.0 Algorithm นั้น ทำให้เราได้โมเดลที่อธิบายให้เราทราบถึงพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของกลุ่มลูกค้าที่ยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ไปแล้ว (Churn) และพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของกลุ่มลูกค้าที่ยังคงใช้บริการโทรศัพท์เคลื่อนที่อยู่ (Non Churn) ว่ามีพฤติกรรมที่เหมือนหรือแตกต่างกันอย่างไร โดยแสดงออกมาในรูปแบบของโมเดลหรือ Rule ซึ่งสิ่งที่ได้จากโมเดล หรือ Rule ก็คือความรู้เกี่ยวกับพฤติกรรมการใช้งานของลูกค้าที่เราสามารถนำไปประยุกต์ใช้ในการทำนายว่าลูกค้าที่มีแนวโน้มที่จะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ในอนาคตนั้นจะเป็นใครบ้าง และจะทำอย่างไรเพื่อไม่ให้ลูกค้ายกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ โดยการนำความรู้ที่ได้จากโมเดลนี้ ไปใช้เป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด เพื่อป้องกันไม่ให้ลูกค้ายกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ ดังตัวอย่างเกี่ยวกับการแนะนำโปรโมชั่นที่เหมาะสมให้กับลูกค้า (Right Package Advisory) ที่กล่าวถึงไปแล้วในบทที่4 หรืออาจจะนำความรู้ที่ได้ไปเป็น

แนวทางในการออกโปรโมชั่นหรือแคมเปญใหม่ๆที่เหมาะสมกับพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้า ซึ่งจะทำให้ลูกค้าเกิดความพึงพอใจ เนื่องจากเป็นสิ่งที่ตรงตามความต้องการของลูกค้า ทั้งนี้เมื่อลูกค้าเกิดความพึงพอใจแล้ว โอกาสที่ลูกค้าจะยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ก็จะลดน้อยลง ทำให้องค์กรสามารถรักษาฐานลูกค้าให้ยังคงอยู่กับองค์กรต่อไป

## 5.2 ข้อเสนอแนะ

การศึกษาโครงการในครั้งนี้ ใช้ข้อมูลเกี่ยวกับพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้ามาวิเคราะห์เป็นส่วนใหญ่ ซึ่งข้อมูลที่นำมาใช้ในการวิเคราะห์นั้นอาจจะยังมีรายละเอียดไม่มากพอ เนื่องด้วยผู้เขียนไม่สามารถที่จะนำเอาข้อมูลในรายละเอียดเชิงลึกเกี่ยวกับการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้ามาวิเคราะห์ได้ ทั้งนี้เนื่องด้วยข้อจำกัดเกี่ยวกับการเข้าถึงข้อมูล ซึ่งผู้เขียนคิดว่าถ้าหากมีข้อมูลเชิงลึกเกี่ยวกับ การใช้งานโทรศัพท์เคลื่อนที่ เช่น ช่วงเวลาที่ลูกค้าใช้งานโทรศัพท์มากที่สุด มาประกอบการวิคราะห์ อาจทำให้ผลลัพธ์ที่ได้มีจุดที่น่าสนใจ และสามารถนำไปใช้เป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด ตลอดจนการออกแคมเปญหรือโปรโมชั่นใหม่ ที่เกี่ยวกับช่วงเวลาของการใช้โทรศัพท์ ยกตัวอย่าง เช่น ถ้าเราทราบว่าลูกค้าของเราใช้โทรศัพท์ในช่วงเวลา18.00 น. ถึง 20.00 น. มากที่สุด เราก็อาจจะออกโปรโมชั่นเป็นค่าโทรอัตราพิเศษในช่วงเวลา 18.00 น. ถึง 20.00 น. เป็นต้น ซึ่งการที่เราสามารถตอบสนองต่อความต้องการของลูกค้าได้มากเท่าไร ก็จะทำให้เราสามารถเข้าถึงกลุ่มลูกค้าได้มากขึ้นด้วย ส่งผลให้การดำเนินธุรกิจขององค์กรเป็นไปในแนวทางที่ดีขึ้น

# บรรณานุกรม

สุวิสาข์ โกพล และสุวิมล คงศักดิ์ตระกูล .2544.**Data Mining.** [Online]. เข้าถึงได้จาก : http://project.cs.kku.ac.th/2544/seminar/day2/413356-4and413357-5/datamining.doc

นิดา หุตะจูฑะ, มาลินี วินิจสร, ชัยวุฒิ พุฒพิสุทธิ์ และประเสริฐ กิ่งพุทธพงศ์. **Data Mining Techniques For Classification Problems.** การทำเหมืองข้อมูลประยุกต์ คณะสถิติประยุกต์. กรุงเทพฯ: สถาบันบัณฑิตพัฒนบริหารศาสตร์.

บุญเสริม กิจศิรกุล. 2546. **อัลกอริทึมการทำเหมืองข้อมูล.** ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย.

Alex Berson, Stephen Smith, Kurt Thearling, **Building Data Mining Application for CRM,** McGraw-Hill, 1999.

Berry, Michael J. and G. Linoff. (1997) **Data Mining Technique : For Marketing, Sales and Customer Support.** New York: Wiley

Berson, Alex and S. J. Smith. (1997) **Data Warehousing, Data Mining, & OLAP.** New York: McGraw Hill.

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

SPSS Inc. (2002) **Data Mining with Confidence. 2nd Edition.** Chicago: SPSS Inc.

SPSS Inc. (2003) **Introduction to Clementine.** Chicago: SPSS Inc.

# ประวัติผู้เขียน

| | |
|---|---|
| ชื่อ-นามสกุล | นางสาวสุภาภรณ์ ศิริติกุล |
| วัน เดือน ปี เกิด | 2 กันยายน พ.ศ. 2519 |
| ประวัติการศึกษา | พ.ศ. 2542 สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาสถิติ จากมหาวิทยาลัยเกษตรศาสตร์ |
| | พ.ศ. 2547 เข้าศึกษาต่อในระดับปริญญาโท หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิชาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง |
| ประวัติการทำงาน | ปัจจุบันทำงานในตำแหน่งเจ้าหน้าที่วิเคราะห์ข้อมูลสังกัดหน่วยงาน Customer Retention and Touch Point บริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) |