การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อศึกษาพฤติกรรมของลูกค้าสมาชิกกลุ่มนกแฟนคลับ สายการบินนกแอร์

ศริชล ชัยชาญทิพยุทธ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตวิทยาลัย สาขาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2557

THE APPLICATION OF DATA MINING TECHNIQUES TO STUDY THE BEHAVIOR OF NOK AIR FANCLUB MEMBERSHIP

Sarishon Chaichantipayoot

A Term Paper Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate School
Thai-Nichi Institute of Technology
Academic Year 2014

หัวข้อสารนิพนธ์	การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อศึกษาพฤติกรรมของลูกค้าสมาชิกกลุ่มนกแฟนคลับ สายการบินนกแอร์
โดย	ศริชล ชัยชาญทิพยุทธ
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษาสารนิพนธ์	ดร. กรกฎ เหมสถาปัตย์

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับสารนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร. พิชิต สุขเจริญพงษ์)

วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบสารนิพนธ์

.....ประธานกรรมการ

(ดร. ภาสกร อภิรักษ์วรพินิต)

.....กรรมการ

(อาจารย์อดิศักดิ์ เสือสมิง)

.....อาจารย์ที่ปรึกษาสารนิพนธ์

(ดร. กรกฎ เหมสถาปัตย์)

ศริชล ชัยชาญทิพยุทธ : การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อศึกษาพฤติกรรมของลูกค้าสมาชิกกลุ่มนกแฟนคลับ สายการบินนกแอร์. อาจารย์ที่ปรึกษา : ดร. กรกฎ เหมสถาปัตย์, 83 หน้า.

ปัจจุบันองค์กรธุรกิจส่วนใหญ่เผชิญกับปัญหาของ ข้อมูลดิบจำนวนมากแต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย Data Mining จึงเป็นเทคนิคสารสนเทศที่สามารถกลั่นกรองข้อมูล เพื่อให้ได้ข้อมูลที่มีประโยชน์ โดยใช้เทคนิคในการทำ Data Mining ที่นิยมใช้ในการทำเหมืองข้อมูลกับฐานข้อมูล และได้ใช้เทคนิคของการค้นหากฎความสัมพันธ์เพื่อใช้ในการค้นหาความสัมพันธ์ของข้อมูลการสืบค้นความรู้บนฐานข้อมูลที่มีโครงสร้างซับซ้อน การกรองความรู้ การจำแนกข้อมูล การสรุปผลและนำเสนอข้อมูลโดยจะแสดงในรูปแบบที่สามารถเข้าใจได้ง่าย

การศึกษาครั้งนี้จึงได้นำกลุ่มข้อมูลดิบของสมาชิกนกแฟนคลับมาวิเคราะห์โดยใช้เทคนิค Data Mining และ RFM มาค้นหากฎความสัมพันธ์ ซึ่งมีวัตถุประสงค์เพื่อศึกษาหาพฤติกรรมของลูกค้ากลุ่มที่เป็นสมาชิกนกแฟนคลับ เพื่อนำมาใช้ในการหาความสัมพันธ์ของกลุ่มลูกค้า

จากการศึกษาพบว่ามีข้อจำกัดบางอย่างของข้อมูล เช่น การขาดหายของข้อมูลทำให้ไม่สามารถอธิบายผลได้ชัดเจนแต่อย่างไรดีการศึกษายังสามารถแบ่งกลุ่มความสัมพันธ์ได้ออกมาเป็น 4 กลุ่ม ซึ่งพบกลุ่มที่น่าสนใจ 2 กลุ่มซึ่งมีผลตอบแทนรายได้ใน 2 ปีที่ผ่านมาเป็นพี่น่าสนใจในการทำการตลาดด้วยคือ กลุ่มที่ 2 และกลุ่มที่ 3 ซึ่งสามารถมองได้เป็นลูกค้าประเภท True Friends และ Butterflies โดยจากผลการศึกษายังสามารถมองเห็นเส้นทางเล็กๆ ที่ไม่ใช่เส้นทางหลักของ 2 กลุ่มนี้เดินทางและน่าสนใจในการทำการตลาดเพิ่มขึ้น

บัณฑิตวิทยาลัย ลายมือชื่อนักศึกษา .....

สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2557

SARISHON CHAICHANTIPAYOOT : THE APPLICATION OF DATA MINING TECHNIQUES TO STUDY THE BEHAVIOR OF NOK AIR FANCLUB MEMBERSHIP. ADVISOR : DR. KORAKOT HEMSATHAPAT, 83 PP.

Nowadays, most of the business organizations are encountering issue of large scale raw data that are little to none inapplicable to be applied into real world utilizations. Therefore, Data Mining is a branch of information technology that can filter and extract complicated information from data sets and transform it into understandable useful and usable data or information. Data mining technique is highly popular in database technology because the data can be searched by utilizing the relationship between its data set to retrieve another data or information.

Therefore, this research and study used raw data of Nok Fan Club for an analysis by utilizing the technique of Data Mining and RFM to explore relationship basis. An objective of this research and study was to use Nok Fan Club customer behaviors data and information then utilize those data or information to search for potential related group of customers in database system.

From this research and study discovery, there were some insufficient or missing pieces of data or information that cannot be exact explained or leaded to conclusion. However, the research and study of this project were divided into four relationship subgroups. There were two very interesting subgroups out of four that showed profitability marketing information or data during the past two years which were Group Two and Group Three customers. These customers were classified as True Friends and Butterflies. The result of this research and study has concluded that these two subgroup customer are regularly flying in less traffic routes. There routes are more interested for future market development.

Graduate School Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2014

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลงได้ด้วยความกรุณาของอาจารย์ ดร. กรกฎ เหมสถาปัตย์ ผู้ให้คำปรึกษาและตรวจสอบงานสารนิพนธ์จนเสร็จสมบูรณ์ผู้วิจัยขอกราบขอบพระคุณด้วยความรู้สึกซาบซึ้งและสำนึกในพระคุณอย่างสูง

ขอขอบพระคุณท่านอาจารย์ดร.สะพรั่งสิทธิ์มฤทุสาธรที่เปิดหลักสูตรและทุ่มเททั้งแรงกายแรงใจคอยให้คำแนะนำต่างๆเป็นอย่างดี

ขอขอบคุณทีมงาน IT Nok Airlines และทีมงานนกแฟนคลับที่ช่วยให้การสนับสนุนคำปรึกษาในการศึกษาปริญญาโทและ ให้ใช้ข้อมูลเพื่อนำมาใช้ในการศึกษาทำสารนิพนธ์นี้สำเร็จลงได้ด้วยดี

ขอบคุณ ดร. เอกสิทธิ์ พัชรวงศ์ศักดา ที่เปิดอบรมหลักสูตร Practical Data Mining with RapidMiner Studio 6 และให้คำปรึกษาสำหรับเทคนิคการใช้โปรแกรม RapidMiner Studio 6 ซึ่งสามารถนำเทคนิคมาใช้กับงานสารนิพนธ์ฉบับนี้เป็นอย่างมาก

ขอกราบขอบพระคุณบิดามารดาและครอบครัวที่ให้โอกาสผู้วิจัยในการศึกษาต่อรวมทั้งการทำสารนิพนธ์ในครั้งนี้ขอบขอบคุณน้องๆ เพื่อนๆ พี่ๆ ที่คณะเทคโนโลยีสารสนเทศสถาบันเทคโนโลยีไทย-ญี่ปุ่น ที่เป็นกำลังใจและให้ความช่วยเหลือด้วยดีเสมอมา

สุดท้ายนี้ผู้จัดทำหวังว่าสารนิพนธ์ฉบับนี้จะเป็นประโยชน์ในการศึกษาและนำไปประยุกต์ใช้ได้ต่อไปในอนาคต

ศริชล ชัยชาญทิพยุทธ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันรัฐบาลมีโครงการพัฒนาระบบการคมนาคมขนส่งให้มีมาตรฐาน สามารถเข้าถึงประชาชนทุกภูมิภาคทั่วประเทศ รวมถึงเพิ่มขีดความสามารถในการแข่งขันของประเทศ ในการก้าวเข้าสู่ประชาคมเศรษฐกิจอาเซียน (ASEAN Economic Community: AEC) ที่ให้ความสำคัญต่อการขับเคลื่อนเศรษฐกิจให้เป็นฐานการผลิตเดียวกัน มีการเคลื่อนย้ายการผลิตอย่างเสรี สามารถสร้างอำนาจทางการต่อรองทางการค้าและเศรษฐกิจในเวทีโลกได้อย่างเข้มแข็งมากยิ่งขึ้น รัฐบาลจึงเดินหน้าพัฒนาระบบรถไฟความเร็วสูงคาดว่าแล้วจะเสร็จภายในปีพ.ศ. 2563 ซึ่งนับเป็นการพัฒนาโครงสร้างพื้นฐานด้านการคมนาคมที่สำคัญยิ่ง เพื่อส่งเสริมศักยภาพให้กับประเทศ ยกระดับคุณภาพชีวิตของประชาชน ตลอดจนขับเคลื่อนประเทศสู่ศูนย์กลางคมนาคมในภูมิภาค และการคมนาคมภายในประเทศทางอากาศนับว่ากำลังได้รับความนิยมเพิ่มขึ้นเป็นอย่างมากในหมู่ผู้ใช้บริการคนไทย ทั้งนี้สาเหตุสำคัญประการหนึ่งเป็นผลมาจากการที่ประเทศไทยได้เปิดให้มีบริการธุรกิจสายการบินต้นทุนต่ำขึ้นภายในประเทศตั้งแต่เมื่อปลายปีพ.ศ. 2546 เป็นต้นมา ทำให้ประชาชนทั่วไปสามารถเลือกใช้บริการชนิดนี้แทนระบบการคมนาคมแบบเดิมๆ อันได้แก่ทางรถยนต์และรถไฟกันได้มากยิ่งขึ้น และนับวันแนวโน้มความต้องการใช้บริการชนิดนี้คาดว่าจะมีเพิ่มมากขึ้นต่อไปเรื่อยๆ ซึ่งธุรกิจการบินต้นทุนต่ำนับเป็นธุรกิจที่มีการแข่งขันกันอย่างมาก ทั้งในด้านราคาตั๋วเครื่องบิน ประสิทธิภาพการให้บริการ ยิ่งตลาดการบินในอาเซียนหลังเปิดประชาคมเศรษฐกิจอาเซียน ต้องยอมรับว่ามีศักยภาพในการเติบโตสูง ทั้งการเดินทางท่องเที่ยว ติดต่อการค้าทำธุรกิจร่วมกันในภูมิภาค ทำให้ธุรกิจการบินต้นทุนต่ำมีการแข่งขันกันรุนแรงมากขึ้น

ธุรกิจสายการบินต้นทุนต่ำเกิดขึ้นครั้งแรกในช่วงปลายปีพ.ศ. 2546 โดยการเปิดตัวของสายการบิน 2 รายในเวลาไล่เลี่ยกัน เริ่มจากไทยแอร์เอเชีย (Thai AirAsia) เปิดให้บริการเป็นรายแรกในเดือนพฤศจิกายน พ.ศ. 2546 หลังจากนั้นประมาณ 1 เดือน สายการบินวันทูโก (One Two Go) ในเครือของโอเรียนท์ ไทย แอร์ไลน์ส (Orient Thai Airlines) ก็เปิดให้บริการตามมา และในเดือนมิถุนายน พ.ศ. 2547 นกแอร์ (NokAir) สายการบินลูกของการบินไทยก็ได้เปิดให้บริการขึ้นเป็นรายล่าสุดและตามมาด้วยสายการบินนกสกู๊ต แอร์ไลน์

การที่มีธุรกิจสายการบินต้นทุนต่ำมีการแข่งขันกันเพิ่มมากขึ้น ส่งผลให้ผู้บริโภคมีทางเลือกในการตัดสินใจที่จะใช้บริการสายการบิน ทำให้เกิดการแข่งขันทางด้านการตลาด ซึ่งมีปัจจัยหลายๆ อย่างที่ผู้บริโภคเลือกที่จะใช้บริการ ไม่ว่าจะเป็นในเรื่องของรูปแบบการให้บริการ ราคาตั๋วเครื่องบิน

สายการบินต่างๆ จึงจำเป็นที่จะต้องใช้กลยุทธ์ทางการตลาดรูปแบบต่างๆ เพื่อที่จะดึงดูดใจผู้บริโภคเข้ามาเป็นสมาชิกและใช้บริการของสายการบินของตน

การเปิดเสรีน่านฟ้าในปี 2015 ทำให้สายการบินของไทยต้องเผชิญกับสภาวะการแข่งขันที่รุนแรงยิ่งขึ้นจากสายการบินในอาเซียนที่ต้องการเข้ามาแข่งขันในประเทศไทย ในอดีต ประเทศไทยอนุญาตให้สายการบินเพียงไม่กี่สายเข้ามาให้บริการภายในประเทศ และจำกัดเส้นทางในการให้บริการของสายการบินเพื่อไม่ให้บินทับเส้นทางหลักของสายการบินแห่งชาติ (การบินไทย) และในภายหลังปี 2002 ได้มีการผ่อนคลายกฎเกณฑ์ในการให้บริการของสายการบิน ทำให้ธุรกิจการบินของไทยมีการแข่งขันมากขึ้น แต่เนื่องจากในปี 2015 ประเทศไทยต้องก้าวเข้าสู่ประชาคมเศรษฐกิจอาเซียน (AEC) ซึ่งได้มีการตกลงในเรื่องนโยบายการเปิดเสรีน่านฟ้า (Open Skies Policy) เพิ่มมากขึ้น เพื่ออำนวยความสะดวกในการประกอบธุรกิจการขนส่งทางอากาศ ภายใต้ความร่วมมือด้านการขนส่งทางอากาศที่มีการให้สิทธิรับขนทั้งในด้านการขนส่งผู้โดยสารและด้านการขนส่งสินค้าทางอากาศตามการจราจรเสรีภาพที่สาม สี่ และห้า อย่างไม่จำกัดระหว่างเมืองในอาเซียน กล่าวคือ สายการบินสามารถให้บริการขนส่งทางอากาศทั้งขาไปและขากลับระหว่างเมืองได้ อีกทั้งสามารถลงจอดเพื่อรับผู้โดยสารและสินค้าจากเมืองตามเส้นทางการบินที่บินผ่านได้ ทำให้สายการบินในอาเซียนต่างก็เตรียมพร้อมเข้ามาให้บริการขนส่งทางอากาศในประเทศไทย ดังเช่น กลุ่มไลอ้อนแอร์ซึ่งเป็นสายการบินต้นทุนต่ำของอินโดนีเซียได้ประกาศว่า จะเข้ามาเปิดบริการสายการบินไทยไลอ้อนแอร์ในไทยและได้มีการเปิดให้บริการแล้วในช่วงปลายปี 2013 และในปี 2014 นี้ กานต์แอร์ได้ร่วมทุนกับเวียดเจ็ทแอร์ประกาศจัดตั้งสายการบินไทยเวียดเจ็ทแอร์ซึ่งเป็นสายการบินต้นทุนต่ำเพื่อเปิดให้บริการในประเทศไทย จะเห็นได้ว่าการเปิดเสรีน่านฟ้าจะทำให้สายการบินในอาเซียนสามารถเข้ามาให้บริการในประเทศไทยได้ทันที ย่อมหมายความว่า ธุรกิจการบินของไทยจะมีการแข่งขันที่รุนแรงมากยิ่งขึ้น

สภาวะการแข่งขันที่รุนแรงมากขึ้นในธุรกิจการบินไทยผู้ให้บริการสายการบินของไทยทั้งสายการบินต้นทุนต่ำและสายการบินให้บริการเต็มรูปแบบควรมีมาตรการเพื่อรองรับการแข่งขันจากสายการบินในอาเซียน จากการเปิด AEC ในปี 2015 ธุรกิจการบินของไทยจะมีการเข้ามาแข่งขันจากสายการบินในอาเซียนโดยเฉพาะสายการบินต้นทุนต่ำ หากผู้ให้บริการสายการบินของไทยไม่มีมาตรการรองรับการแข่งขัน จะส่งผลต่อความอยู่รอดของธุรกิจในระยะยาว แม้ปัจจุบันสายการบินของไทยต่างก็มีการขยายธุรกิจการบินเพิ่มขึ้นเพื่อให้บริการเส้นทางการบินที่ไกลขึ้น ดังเช่น ไทยแอร์เอเชียเอ็กซ์ นกสกู๊ต หรือเพื่อให้บริการสายการบินราคาประหยัด ดังเช่น การบินไทยสมายล์ แต่อย่างไรก็ตาม สายการบินในอาเซียนต่างก็มีความพร้อมที่จะเปิดให้บริการในเส้นทางเหล่านั้นเช่นกัน ดังนั้น สายการบินของไทยควรมีมาตรการเพิ่มเติม เช่น การลดต้นทุนการดำเนินการ เพื่อรักษากำไรไม่ให้ลดลงมากนัก หรือเพิ่มคุณภาพของการให้บริการและสร้างภาพลักษณ์ที่ดี เพื่อรักษาความสามารถในการแข่งขันกับคู่แข่งที่จะมีมากขึ้นในอนาคต

ด้วยปัญหาดังกล่าวผู้ศึกษาจึงได้จัดทำการศึกษานี้ขึ้นเพื่อเพิ่มอัตราการพบโดยใช้เทคนิค Data Mining นำมาใช้ในการศึกษาพฤติกรรม และช่วยในการเพิ่มยอดกลุ่มลูกค้าสมาชิกใหม่ของสมาชิกกลุ่มนกแฟนคลับอนึ่ง การสร้างความสัมพันธ์กับลูกค้าโดยอาศัยเทคโนโลยีและการใช้บุคลากรอย่างมีหลักการ จะช่วยให้เกิดการบริการที่ดีขึ้น สามารถเก็บข้อมูลและวิเคราะห์พฤติกรรมของลูกค้าในการใช้บริการแต่ละประเภท เพื่อที่จะค้นหาความต้องการของลูกค้า อีกทั้งยังสามารถนำข้อมูลมาเพื่อมาปรับปรุงพัฒนาการบริการให้ดีขึ้น ฉะนั้นแล้วการนำเทคโนโลยีเหมืองข้อมูล ซึ่งเป็นเทคโนโลยีที่ได้รับการยอมรับว่าเป็นเครื่องมือที่นำไปสู่การเข้าใจลูกค้า จึงถูกนำมาใช้เพื่อวิเคราะห์พฤติกรรมความต้องการ ตลอดจนแนวโน้มในการใช้จ่ายของลูกค้า อีกทั้งหลายองค์กรยังหยิบมาใช้เพื่อช่วยในการวิเคราะห์สิ่งต่างๆ เช่น ช่วยในการพยากรณ์การยอดการให้บริการในอนาคต ช่วยหาความสัมพันธ์ของสินค้าหรือการบริการ โดยผลที่ได้จากการวิเคราะห์จะนำข้อมูลไปปรับปรุงการทำงานรวมถึงสร้างเป็นโปรโมชั่นให้ตรงกับความต้องการ หรือนำไปสู่การตัดสินใจในการลงทุนภายในองค์กร

ซึ่งการรักษาลูกค้านั้นจำเป็นต้องลดหรือจำกัดอิทธิพลเชิงลบเกี่ยวกับการบริการที่มีอยู่ โดยการสร้างความมั่นใจให้ลูกค้า นอกจากนี้ทางสายการบินจะต้องตรวจสอบอย่างใกล้ชิดในการกำหนดคุณลักษณะของลูกค้าให้ถูกต้องกับกลุ่มเป้าหมาย ซึ่งลูกค้าที่จะนำมาพิจารณานั้นไม่เป็นแค่เพียงกลุ่มสมาชิกนกแฟนคลับ แต่ยังรวมถึงพฤติกรรมการบริโภคและการรับรู้ถึงคุณภาพในการให้บริการ เพื่อที่จะรักษากลุ่มลูกค้าจึงถือได้ว่าความภักดีของลูกค้าเป็นรากฐานที่สำคัญในการพัฒนาความได้เปรียบทางการแข่งขันแบบยั่งยืนของธุรกิจประเภทต่างๆ ทั้งธุรกิจผู้ผลิตสินค้า บริการ และการค้าปลีก [1]

วัฏจักรขั้นตอนการทำงานของ Data Mining ประกอบไปด้วย 4 ขั้นตอนหลักๆ ดังนี้

1. การระบุโอกาสทางธุรกิจหรือการระบุปัญหาที่เกิดขึ้นกับธุรกิจ เป็นการระบุขอบเขตของข้อมูลที่จะนำมาทาการวิเคราะห์เพื่อหาความได้เปรียบทางการตลาดหรือเพื่อนำมาทาการแก้ไขปัญหา

2. ส่วนของ Data Mining เป็นการนำเทคนิคของ Data Mining ไปใช้ถ่ายทอดหรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลที่จะนำไปใช้ได้จริงในทางธุรกิจ

3. การทดลองปฏิบัติตามข้อมูล คือการนำเอาข้อมูลที่เป็นผลลัพธ์ของส่วน Data Mining มาทดลองปฏิบัติจริงกับธุรกิจ

4. การวัดประสิทธิภาพจากผลลัพธ์ การวัดประสิทธิภาพของเทคนิคของ Data Mining ที่จะนำมาใช้จากผลลัพธ์ ซึ่งสามารถตรวจสอบได้หลายทาง เช่น วัดจากส่วนแบ่งของตลาด วัดจากปริมาณลูกค้า หรือ วัดจากกำไรสุทธิ เป็นต้น จากทั้ง 4 ขั้นตอนที่กล่าวมาข้างต้นคือการนำเอา Data Mining ไปใช้กับระบบทางธุรกิจ โดยแต่ละขั้นตอนจะพึ่งพาอาศัยกันผลลัพธ์จากขั้นตอนหนึ่งจะกลายมาเป็นอินพุทจากอีกขั้นตอนต่อไป ซึ่ง Data Mining จะเปลี่ยนข้อมูลดิบให้เป็นข้อมูลประยุกต์ ดังนั้นการระบุแหล่งข้อมูลที่ถูกต้องจึงเป็นสิ่งที่สำคัญอย่างยิ่งต่อผลลัพธ์ที่ได้จากการวิเคราะห์

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษาพฤติกรรมของกลุ่มลูกค้าสมาชิกของกลุ่มนกแฟนคลับ โดยใช้เทคนิค Data mining มาวิเคราะห์กับข้อมูลดิบที่มีอยู่

1.2.2 เพื่อศึกษาปัจจัยต่างๆ ที่จะสามารถนำมาใช้เป็นตัวแบ่งข้อมูลในการทำเหมืองเพื่อค้นหาความสัมพันธ์ของกลุ่มลูกค้า เพื่อให้ได้ความสัมพันธ์ที่มีความถูกต้องแม่นยำ

1.2.3 เพื่อให้ได้ความสัมพันธ์ของกลุ่มลูกค้า สำหรับใช้เป็นข้อมูลช่วยสนับสนุนการตัดสินใจในการรักษากลุ่มลูกค้าและเพิ่มประสิทธิภาพการแข่งขัน หรือการในการเจาะหากลุ่มลูกค้าใหม่

1.3 ขอบเขตการศึกษา

ในการศึกษาครั้งนี้ จะทำการศึกษาถึงพฤติกรรมการใช้บริการที่มีผลต่อการตัดสินใจสมัครสมาชิกกลุ่มนกแฟนคลับโดยจะเน้นศึกษาจากข้อมูลสมาชิกที่มีอยู่แล้วจำนวน 290,000 ราย ที่มีการใช้บริการในปี 2012-2013 จำนวน 149,831 ราย

1.4 ขั้นตอนการดำเนินงาน

1.4.1 ศึกษา ค้นคว้า และสรุปหลักการของเทคนิคการทำเหมืองข้อมูล (Data Mining) คุณสมบัติ ลักษณะ และการนำเทคนิคเหมืองข้อมูลไปใช้ในการวิเคราะห์ข้อมูลรูปแบบต่างๆ

1.4.2 วิเคราะห์หารูปแบบความสัมพันธ์ของข้อมูลโดยเทคนิคที่เลือก เพื่อให้ได้ผลลัพธ์จากการทำเหมืองข้อมูลที่แสดงส่วนที่สนใจ

1.4.3 ประเมินความถูกต้องของผลลัพธ์ที่ได้ รวมถึงความน่าสนใจของผลลัพธ์นั้น

1.4.4 นำผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลมาประยุกต์ใช้

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ทราบถึงปัจจัยที่สามารถนำมาใช้เป็นข้อมูลในการทำ Data Mining เพื่อให้ได้ความสัมพันธ์ของลูกค้าที่ถูกต้องแม่นยำ

1.5.2 ได้ความสัมพันธ์ของกลุ่มลูกค้าสมาชิกกลุ่มนกแฟนคลับ สำหรับใช้เป็นข้อมูลช่วยสนับสนุนการตัดสินใจ ในการวางแผนกลยุทธ์ทางด้านการตลาด และเพิ่มประสิทธิภาพการแข่งขัน หรือการในการเจาะหากลุ่มลูกค้าใหม่ และรักษากลุ่มลูกค้าเดิม

1.5.3 ช่วยให้เข้าใจพฤติกรรมของลูกค้าสมาชิกลุ่มนกแฟนคลับแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะเสนอโปรโมชั่นได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม

1.6 นิยามศัพท์เฉพาะ

ผู้โดยสารที่จงรักภักดี (Passenger Loyalty) หรือ ลูกค้าที่จงรักภักดี (Customer Loyalty) โดยทั่วไปความภักดีของผู้โดยสาร (Passenger Loyalty) ถูกกำหนดว่าเป็นตัวชี้วัดพฤติกรรมการใช้บริการของลูกค้าที่มีอยู่จึงให้รางวัลกับผู้โดยสารโดยพิจารณาจากปริมาณการใช้บริการ ยิ่งใช้บริการมากเท่าไร ก็จะได้รับรางวัลตอบแทนจากการใช้บริการมากเท่านั้น เช่น การสะสมไมล์ การสะสมยอดซื้อ การสะสมแต้ม เป็นต้น ซึ่ง Werner J. Reinartz และ V.Kumar [2] ได้นำเสนอวิธีการแบ่งกลุ่มลูกค้าโดยใช้ระยะเวลาการเป็นลูกค้าและความสามารถในการทำกำไรเป็นเกณฑ์แบ่งลูกค้าออกเป็น 4 กลุ่ม คือ 1) Strangers (คนแปลกหน้า) 2) Butterflies (พร้อมเปลี่ยนได้ทุกเมื่อเปรียบเสมือนผีเสื้อที่พร้อมจะโบยบิน) 3) True Friends (เพื่อนแท้) 4) Barnacles (เกาะติดกับบริษัท เปรียบเสมือนเพรียงที่เกาะใต้ท้องเรือ)

กลุ่มที่ 1 Strangers ลูกค้าคนแปลกหน้า (ลูกค้าที่มีกำไรต่ำและซื้อสินค้าหรือบริการจากเราในระยะสั้น) เป็นกลุ่มลูกค้าที่ไม่มีความภักดีต่อบริษัทและแทบจะไม่ได้สร้างกำไรให้กับบริษัท สินค้าหรือบริการที่เราเสนอสามารถตอบสนองความต้องการของลูกค้ากลุ่มนี้ได้น้อยมาก ดังนั้น จึงมีโอกาสที่จะทำกำไรจากลูกค้ากลุ่มนี้ต่ำ เราจึงต้องระบุตัวลูกค้ากลุ่มนี้ให้ได้ตั้งแต่เนิ่นๆ และควรหลีกเลี่ยงการลงทุนใดก็ตามเพื่อสร้างความสัมพันธ์กับลูกค้ากลุ่มนี้ เราควรแสวงหากำไรในการซื้อขายแต่ละครั้งกับลูกค้ากลุ่มนี้ เพราะการซื้อในแต่ละครั้งอาจเป็นครั้งสุดท้ายก็ได้

กลุ่มที่ 2 Butterflies ลูกค้าที่พร้อมจะเปลี่ยนไปได้ทุกเมื่อ (ลูกค้าที่มีกำไรสูง และซื้อสินค้าหรือบริการจากเราในระยะสั้น) ลูกค้ากลุ่มนี้เอาแน่เอานอนไม่ได้ แต่ก็สามารถสร้างกำไรให้กับเราได้อย่างมาก และมักจะไม่แสดงความภักดีทางพฤติกรรม ลูกค้ากลุ่มนี้พร้อมเปลี่ยนได้ทุกเมื่อ มีอยู่เป็นจำนวนมากในหลายอุตสาหกรรม ลูกค้ากลุ่มนี้มักจะซื้อเป็นจำนวนมากในระยะเวลาสั้นๆ แล้วก็เปลี่ยนไปซื้อจากคู่แข่งรายอื่น พวกเขาจะหลีกเลี่ยงการสร้างความสัมพันธ์ระยะยาวกับบริษัทใดบริษัทหนึ่ง พร้อมที่จะเปลี่ยนเมื่อได้ข้อเสนอที่ดีกว่า

กลุ่มที่ 3 True Friends เพื่อนแท้ (ลูกค้าที่มีกำไรสูงและซื้อสินค้าหรือบริการจากเราเป็นระยะเวลานาน) ลูกค้าที่มีทั้งความภักดีและสร้างกำไรให้กับบริษัทนั้นเรียกว่า เพื่อนแท้ ลูกค้ากลุ่มนี้จะซื้อจากเราเป็นประจำและสม่ำเสมอ (แต่ไม่ได้ซื้อในปริมาณที่มากเกินไปในแต่ละครั้ง) เป็นระยะเวลานาน โดยทั่วไปลูกค้ากลุ่มนี้จะรู้สึกพอใจกับสิ่งที่เราเสนอให้ในปัจจุบัน ซึ่งสามารถเห็นได้จากความภักดีและความสามารถในการทำกำไรจากพวกเขา ลูกค้ากลุ่มนี้เต็มใจที่จะมีส่วนร่วมในกระบวนการหรือกิจกรรมต่างๆ ของบริษัท เราจึงควรให้ความระมัดระวังและใส่ใจในการสร้างความสัมพันธ์กับลูกค้ากลุ่มนี้ เพราะลูกค้าเหล่านี้มีโอกาสที่จะสร้างกำไรให้กับเราได้ในระยะยาวมากที่สุด

กลุ่มที่ 4 Barnacles ลูกค้าที่เกาะติดบริษัท (ลูกค้าที่มีกำไรต่ำ และซื้อสินค้าหรือบริการจากเราเป็นเวลานาน) ลูกค้ากลุ่มนี้เปรียบเสมือนเพรียง (Barnacles) ที่เกาะติดอยู่ใต้ท้องเรือซึ่งเป็นตัวถ่วงน้ำหนักของเรือ หากเราจัดการลูกค้ากลุ่มนี้ไม่ดีพอ อาจทำให้เราต้องสูญเสียทรัพยากรจำนวนมากไปโดยเปล่าประโยชน์จำนวนและปริมาณการซื้อของลูกค้ากลุ่มนี้อาจไม่คุ้มกับค่าใช้จ่ายที่เกิดจากการส่งเสริมการตลาด และรักษาความสัมพันธ์กับพวกเขา แต่ถ้าเราสามารถบริหารลูกค้ากลุ่มนี้ได้อย่างเหมาะสมและสามารถเปลี่ยนเป็นลูกค้ากลุ่ม True Friends ได้ จะกลายเป็นลูกค้าที่สร้างกำไรในอนาคต

1.7 โครงการสมาชิกนกแฟนคลับ

โครงการสมาชิกนกแฟนคลับเกิดขึ้นเพื่อดูแลผู้โดยสารที่รักการท่องเที่ยว และมีการเดินทางอย่างต่อเนื่องกับสายการบินนกแอร์ โดยมีเป้าหมายที่จะเพิ่มรอยยิ้มให้แก่สมาชิกทุกท่านผ่านสิทธิพิเศษและสิทธิประโยชน์ต่างๆ ของโครงการฯ โดยปัจจุบัน โครงการมีสมาชิกประมาณ 240,000 คน แบ่ง 2 ระดับสมาชิก คือ

1. Nok Smile เปิดรับสมัครผู้สนใจทุกท่าน
2. Nok Smile Plus (สมาชิกที่ได้รับการเลื่อนระดับขึ้นมาจากสมาชิก Nok Smile เท่านั้น ใช้บริการเกิน 35 เที่ยวบิน หรือ 15000 คะแนนต่อปี)

โดยทุก 5 บาทของราคาตั๋วโดยสารสมาชิกนกแฟนคลับจะได้รับ Nok Point 1 คะแนน เพื่อสะสมไว้แลกสิทธิประโยชน์ และของรางวัลต่างๆ จากโครงการฯโดยการสะสม Nok Point หากค่าที่ได้เป็นทศนิยมจากการคำนวณ จะถูกปัดลงและแสดงเป็นจำนวนเต็มเสมอและ Nok Point มีอายุ 12 เดือนนับจากเดือนคะแนนเข้า

<u>สิทธิประโยชน์สมาชิกนกแฟนคลับ</u>

สมาชิกระดับ Nok Smile

- แลกตั๋วเครื่องบินฟรี 1 ใบ เมื่อสะสม Nok Point ครบตั้งแต่ 7,500 คะแนนภายในอายุสมาชิก
- รับน้ำหนักกระเป๋าเพิ่มอีก 5 กิโลกรัมจากน้ำหนักกระเป๋าปกติ
- รู้ตั๋วโปรโมชั่นก่อนใคร
- ข้อเสนอพิเศษจาก Partners นกแฟนคลับ

สมาชิกระดับ Nok Smile Plus

- แลกตั๋วเครื่องบินฟรี 1 ใบ เมื่อสะสม Nok Point ครบตั้งแต่ 5,500 คะแนนภายในอายุสมาชิก
- รับน้ำหนักกระเป๋าเพิ่มอีก 10 กิโลกรัมจากน้ำหนักกระเป๋าปกติ
- รู้ตั๋วโปรโมชั่นก่อนใคร
- อัพเกรดที่นั่งเป็นนกพลัส Premium Seat
- ข้อเสนอพิเศษจาก Partners นกแฟนคลับ
- Nok Calls Home บริการติดต่อสมาชิกเพื่อให้บริการเช็คอิน

บทที่ 2
แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

2.1 การประเมินความจงรักภักดีของผู้โดยสาร

แนวคิดของความภักดี (Loyalty) เกิดขึ้นครั้งแรกในปี ค.ศ. 1950 คือพฤติกรรมการซื้อซ้ำของลูกค้ามีความตั้งใจและยินดีที่จะแนะนำให้ผู้อื่นซื้อสินค้าหรือใช้การหรือการบอกต่อของลูกค้า [1] มีผู้ที่นิยามความหมายของความจงรักภักดีและวิธีการวัดความจงรักภักดีไว้ค่อนข้างกว้างและหลากหลาย ขึ้นอยู่กับข้อมูลและพฤติกรรมของลูกค้า รวมทั้งบริบทของธุรกิจต่างๆอีกด้วย ความภักดีของลูกค้าเป็นรากฐานที่สำคัญของผลกำไรทางธุรกิจ ความจงรักภักดีเป็นการตอบสนองของผู้บริโภคในการบ่งบอกถึงระดับของสิ่งที่ลูกค้ารู้สึกและแสดงออกว่าเป็นที่พอใจหรือไม่พอใจอย่างไร [2] ความจงรักภักดีนั้นเป็นความมุ่งมั่น อย่างต่อเนื่องที่จะดำเนินความสัมพันธ์ของลูกค้าโดยมีปัจจัยที่เกี่ยวข้อง เช่น ความพึงพอใจของลูกค้า ความจงรักภักดีเป็นพฤติกรรมที่เกิดขึ้นหลังจากการซื้อหรือได้รับการบริการโดยมีความตั้งใจที่จะซื้อซ้ำและความตั้งใจที่จะบอกต่อในทางบวก [3] และความจงรักภักดีนั้นหมายรวมถึงพฤติกรรมที่แสดงออกและทัศนคติที่มีต่อบริษัทผู้ขาย/ผู้ให้บริการ [4]

จากความหมายของคำนิยามเหล่านี้ได้กำหนดความจงรักภักดีของลูกค้าในรูปแบบที่แตกต่างกัน 2 รูปแบบ คือ

1) กำหนดความลงรักภักดีของลูกค้าเป็นทัศนคติ (Attitudinal loyalty) เช่น ความพึงพอใจ การรับรู้ถึงความสัมพันธ์อันดี ความตั้งใจที่จะซื้อซ้ำ เป็นต้น และ

2) กำหนดเป็นพฤติกรรม (Behavior Loyalty) สามารถประเมินในรูปแบบของการซื้อซ้ำบอกต่อและเพิ่มขนาดขอบเขตความสัมพันธ์เช่น ความพยายามในการซื้อเพิ่มมากขึ้น เป็นต้น นอกจากนี้ความลงรักภักดีของลูกค้ายังสามารถแบ่งออกเป็นสามกลุ่มโดยไม่คำนึงถึงความหมายและการวัดคือ 1) พฤติกรรมในการซื้อซ้ำและการบอกต่อ 2) ความจงรักภักดีเป็นส่วนประกอบรวมของการอุดหนุนและองค์ประกอบเจตคติ และ 3) ความจงรักภักดีเป็นความสัมพันธ์ทางด้านจิตใจ ซึ่งเป็นความสำคัญเกี่ยวกับการบอกต่อว่า เป็นหนึ่งในปัจจัย ที่สำคัญในการแสวงหาลูกค้าใหม่ [5]

2.2 การทำเหมืองข้อมูล (Data Mining)

2.2.1 ความหมายของการทำเหมืองข้อมูล

Data Mining เป็นกระบวนการของการกลั่นกรองข้อมูลสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูล เพื่อทำนายพฤติกรรม โดยอาศัยข้อมูลในอดีต เพื่อใช้ข้อมูลสารสนเทศเหล่านี้มาช่วยในการสนับสนุนการตัดสินใจในในทางธุรกิจ

วิวัฒนาการของเหมืองข้อมูล (Data Mining)

ปี 1960 - Data Collection คือ การนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือและป้องกันการสูญหายได้เป็นอย่างดี

ปี 1980 - Data Access คือ การนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ต่อกันในข้อมูลเพื่อประโยชน์ในการนำไปวิเคราะห์ และการตัดสินใจอย่างมีคุณภาพ

ปี 1990 - Data Warehouse & Decision Support คือ การรวบรวมข้อมูลมาจัดเก็บลงไปในฐานข้อมูลขนาดใหญ่โดยครอบคลุมทุกแง่ ทุกมุมขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ

ปี 2000 - Data Mining คือ การนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผลโดยการสร้างแบบจำลอง และความสัมพันธ์ทางสถิติ

โดยสรุปการทำเหมืองข้อมูล หรือ Data Mining หมายถึงการที่ผู้ใช้ดึงข้อมูลโดยการสังเคราะห์และตรวจสอบข้อมูลอย่างละเอียด โดยการสังเคราะห์ดังกล่าวอาจเป็นการเรียนรู้ข้อมูลในอดีตหรือข้อมูลในปัจจุบัน ผลลัพธ์ที่ได้มาต้องมีลักษณะของข้อมูลที่เป็นข้อมูลแบบ Unknown, ข้อมูลแบบ Valid และข้อมูลแบบ Actionable มาจากฐานข้อมูลที่เป็นขนาดใหญ่ซึ่งมาจาก Transection ต่างๆ ในฐานข้อมูล

ข้อมูลแบบ Unknown คือข้อมูลที่ถูกใช้จะต้องเป็นข้อมูลผู้ใช้งานไม่รู้มาก่อนและไม่ชัดเจนจนไม่สามารถตั้งสมมติฐานได้ล่วงหน้าว่าควรเป็นแบบใด ตัวอย่างเช่น เจ้าของห้างสรรพสินค้าแห่งหนึ่งเพิ่งจะค้นพบว่าพฤติกรรมของผู้บริโภคใหม่ที่เป็นพ่อบ้าน มักจะซื้อสินค้าประเภทเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น ดังนั้นเพื่อเป็นสัญญาณให้เจ้าของกิจการควรจะต้องเตรียมสินค้าไว้เพื่อจำหน่าย

ข้อมูลแบบ Valid เมื่อผู้ใช้ได้เริ่มใช้เทคนิค Data Mining จะค้นพบสิ่งที่น่าสนใจโดยต้องพิจารณาด้วยว่าสิ่งนั้น Valid หรือไม่เช่นผู้ใช้มักจะพบว่ามีความสัมพันธ์ของการซื้อของ 2 สิ่งเสมอเมื่อจำนวนความหลากหลายของสินค้ามีมากขึ้นแต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้าเก็บสินค้ามากขึ้นเพราะข้อมูลที่ได้อาจเกิดความคลาดเคลื่อนเพราะฉะนั้นจะต้องทำการValidation และ Checking ความถูกต้องของข้อมูลและวิเคราะห์ความถูกต้องอีกครั้ง

ข้อมูลแบบ Actionable ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจให้เป็นความได้เปรียบเชิงธุรกิจบางครั้งข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้วจึงต้องมีวิจารณญาณในการใช้ด้วยบางทีข้อมูลดังกล่าวอาจจะไม่มีประโยชน์ก็ได้

2.3 หลักการทั่วไปของ Knowledge Discovery in Database (KDD) and Data Mining

KDD หมายถึงกระบวนการในการค้นหาลักษณะแฝงของข้อมูลที่อยู่ในกลุ่มข้อมูลจำนวนมากซึ่งมีขั้นตอนการทำ Data Mining เป็นกระบวนการที่สำคัญในการค้นหาลักษณะที่น่าสนใจของข้อมูลเหล่านี้เช่นรูปแบบความสัมพันธ์การเปลี่ยนแปลงโครงสร้างที่เด่นชัดหรือลักษณะที่ผิดปกติของข้อมูลจากข้อมูลจำนวนมากๆที่เก็บอยู่ในฐานข้อมูลหรือแหล่งที่เก็บข้อมูลอื่นๆ ซึ่งวิธีการต่างๆ ที่นำมาใช้ในการทำ Data Mining นี้ก็มีวัตถุประสงค์ต่างๆ กันขึ้นอยู่กับผลลัพธ์ของกระบวนการโดยรวมที่ต้องการดังนั้นจึงควรมีการนำเสนอวิธีการที่หลากหลายสำหรับเป้าหมายที่แตกต่างกันเพื่อให้ได้ผลลัพธ์ที่เหมาะสมตามที่ต้องการหลังจากนำไปใช้งานแล้วและเนื่องจากความแพร่หลายของการจัดเก็บข้อมูลในลักษณะที่เป็นรูปแบบทางอิเล็กทรอนิกส์และความต้องการในการเปลี่ยนข้อมูลเหล่านั้นให้เป็นข้อมูลที่มีประโยชน์ต่อการนำไประยุกต์ใช้ในงานด้านต่างๆเช่นการวิเคราะห์ด้านการตลาดการบริหารธุรกิจรวมถึงระบบที่ช่วยสนับสนุนการตัดสินใจเป็นต้นดังนั้นจึงทำให้การนำ Data Mining มาใช้ได้รับความสนใจมากในช่วง 2-3 ปีที่ผ่านมา

จากที่ได้กล่าวแล้วว่า Data Mining เป็นขั้นตอนหนึ่งที่สำคัญในกระบวนการค้นหาลักษณะของข้อมูลแฝงที่มีประโยชน์ในฐานข้อมูล (Knowledge Discovery in Database : KDD) ซึ่งโดยทั่วไปกระบวนการของ KDD นั้นประกอบด้วยขั้นตอนต่างๆดังนี้

1. การคัดเลือกข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำ Data Mining รวมถึงการนำข้อมูลที่ต้องการออกมาจากฐานข้อมูลเพื่อทำการพิจารณาในเบื้องต้นต่อไป

2. การกรองข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่าถูกต้องโดยการนำข้อมูลที่ไม่ถูกต้องออก

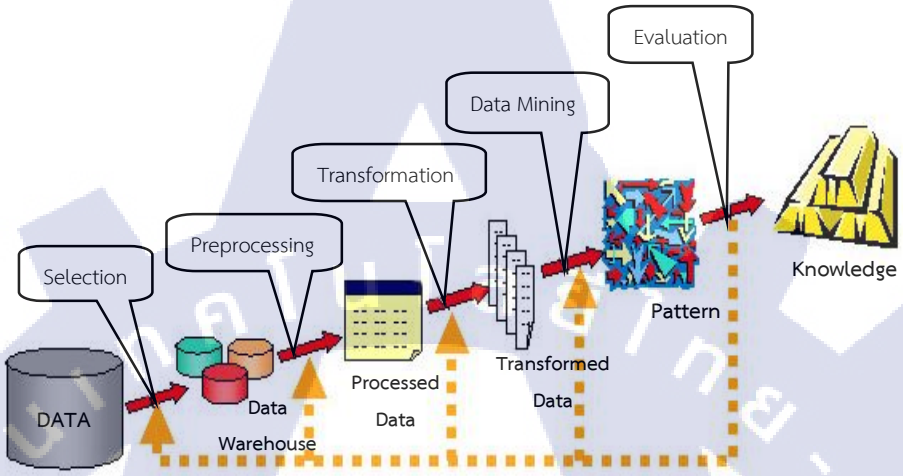
3. การแปลงรูปแบบข้อมูล (Data Transformation) เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้วิเคราะห์ตามอัลกอริทึม (Algorithm) และแบบจำลองที่ใช้ในการทำ Data Mining ต่อไป

4. การทำเหมืองข้อมูล (Data Mining) การใช้เทคนิคเพื่อทำเหมืองข้อมูลโดยทั่วไปประเภทของงานตามลักษณะของแบบจำลองที่ใช้ในการทำ Data Mining นั้นสามารถแบ่งกลุ่มได้เป็น 2 ประเภทใหญ่ๆ คือ

4.1 Predictive Data Mining คือ เป็นการทำนายพฤติกรรมจากข้อมูลที่เกิดขึ้นในอดีต สร้างต้นแบบ ที่สามารถใช้ทำนายกับข้อมูลใหม่ หรือข้อมูลที่เกิดขึ้นภายหลัง

4.2 Descriptive Data Mining คือ เป็นแบบจำลองนี้ไม่ได้มีจุดประสงค์เพื่อการทำนาย แต่เพื่อใช้อธิบายรูปแบบของข้อมูล เพื่อใช้เป็นแนวทางในการตัดสินใจ

5. การวิเคราะห์และประเมินผลลัพธ์ที่ได้ (Result Analysis and Evaluation) เป็นขั้นตอนการแปลความหมายและการประเมินผลลัพธ์ที่ได้ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่โดยทั่วไปควรมีการแสดงผลในรูปแบบที่สามารถเข้าใจได้โดยง่าย



รูปที่ 2.1 ขั้นตอนกระบวนการทำงาน KDD Process

2.4 การเลือกข้อมูล (Data Selection)

จุดประสงค์คือการระบุแหล่งของข้อมูลที่มีและ ทำการดึงเอาข้อมูลออกมาใช้สำหรับการวิเคราะห์เบื้องต้นในการเตรียมตัวสำหรับการที่จะทำการเหมืองข้อมูลในขั้นต่อๆ ไป การเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจ ที่ได้กำหนดไว้ตั้งแต่ต้นและการเลือกข้อมูลก็ยังถูกกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องถูกทำ ความเข้าใจว่าตัวแปรแต่ละตัวหมายความว่าอะไร ประกอบด้วยอะไร ไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้น แต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล, ค่าที่เป็นไปได้, แหล่งกำเนิดของข้อมูล, รูปแบบของข้อมูล และลักษณะอื่นๆ จะมีตัวแปร 2 ชนิดคือ

2.4.1 ตัวแปรแบบ Categorical

1. Nominal Variable กล่าวถึงชนิดนี้ของ Object ที่มันอ้างถึงแต่ไม่มีลำดับในค่าที่เป็นไปได้ (Possible Value) ตัวอย่างเช่น สถานะ การแต่งงาน (โสด, แต่งงาน, หย่า, ไม่ทราบ), เพศ (ชาย, หญิง), ระดับการศึกษา (ปริญญาโท, ปริญญาตรี, ม.ปลาย, ปวช)

2. Ordinal Variable มีลำดับสำหรับค่าที่เป็นไปได้ตัวอย่างเช่นลำดับของลูกค้า (ดี, ปานกลาง, ไม่ดี)

2.4.2 ตัวแปรแบบ Quantitative ซึ่งมีการวัดความแตกต่างระหว่างค่าที่เป็นไปได้

1. Continuous (ค่าที่ต่อเนื่อง) เช่นรายได้, เฉลี่ยจำนวนครั้งที่ซื้อ, รายได้
2. Discrete (ค่าเป็นจำนวนเต็ม) เช่นจำนวนพนักงาน, เวลาปี (เดือน, ฤดู, ไตรมาส)

ตัวแปรของข้อมูลมีหลายตัวมากแต่ตัวแปรที่ถูกเลือกสำหรับทำ Data Mining นั้นถูกเรียกว่า “Active Variable” เพราะว่ามันจะถูกใช้สร้างความแตกต่างของกลุ่มย่อยต่างๆ และสามารถถูกนำมาทำนายผลได้ เมื่อคุณทำการเลือกข้อมูลจะต้อง พิจารณาอายุของข้อมูลด้วย เพราะว่าสถานการณ์ภายนอกเปลี่ยนแปลงตลอดเวลาซึ่งจะทำให้ประสิทธิภาพของการทำเหมืองข้อมูลลดลงตัวอย่าง รสนิยมการใช้ชีวิตการเปลี่ยนงาน

2.5 การกลั่นกรองข้อมูล (Data Preprocessing)

จุดประสงค์ก็เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นเหมาะสม ข้อมูลที่สมบูรณ์เป็นเครื่องประกันว่าการทำ Data Mining จะสำเร็จ ในขั้นตอนนี้เป็นขั้นตอนที่มีปัญหามากกว่า ในขั้นตอนของการเตรียมข้อมูล เพราะข้อมูลส่วนใหญ่ที่มีในองค์กร ไม่ได้ถูกเตรียมมาเพื่องาน Data Mining โดยเฉพาะข้อมูลจะถูกนำมาจากแหล่งต่างๆ ถูกจัดเก็บไม่ดี ข้อมูลที่ถูกนำมาจากภายนอกแล้วนำมาเพื่อให้เข้ากับข้อมูลภายในที่มีอยู่ ปัญหาหลักของ Data คือ คุณภาพและความถูกต้องของข้อมูล (Data Integrity)

โดยในขั้นตอนนี้ก่อนอื่นจะต้องทำการทบทวนโครงสร้างของข้อมูลใหม่ และวัดคุณภาพของมัน โดยวิธีทางสถิติ หรือสุ่มตัวอย่างเครื่องมือที่ใช้ในการทำการกลั่นกรองข้อมูลมีดังต่อไปนี้

1. ค่าตัวแปรเป็นแบบ Categorical การแบ่งความถี่ของค่าจะเป็นวิธีที่ทำให้เกิดความเข้าใจใน Data Content เครื่องมือทางด้านกราฟฟิกจะเป็นตัวช่วยให้เห็นและกำหนดค่าที่หายไปได้
2. ตัวแปรแบบ Quantitative ตัวแปรประเภทนี้มักมีการใช้การวัดตัวอย่างเช่นค่าสูงสุดค่าต่ำสุดค่าเฉลี่ยค่ากลางค่ามัธยฐานและค่าอื่นๆ ทางสถิติเมื่อนำค่าเหล่านี้มาเข้าสูตรคำนวณก็จะบอกถึงค่าที่ไม่สมบูรณ์หรือค่าที่มีปัญหาระหว่างการทำขั้นตอนการกลั่นกรองข้อมูลจะมีปัญหาบ่อยๆ ที่มักพบได้ ได้แก่

Noisy Data คือ ตัวแปรตัวหนึ่งหรือมากกว่ามีค่าซึ่งเกินกว่าค่าที่เราคาดไว้ ซึ่งอาจจะหมายถึงทั้งข้อดีและข้อเสียก็ได้ ในข้อดีก็คือ มันจะแสดงอย่างชัดเจนถึงโอกาสซึ่งเรากำลังมองหาอยู่ ในข้อเสียคือมันอาจจะเป็นข้อมูลที่ไม่สมบูรณ์สาเหตุที่เกิดขึ้นอาจจะมาจากความไม่รอบคอบของมนุษย์ตัวอย่างเช่น Operator ใส่อายุให้คนเป็น 300 ปีหรือใส่ค่าของรายได้เป็นติดลบค่าเหล่านี้ ควรจะถูกแก้ไขหรือเอาออกจากการวิเคราะห์ควรมีขั้นตอนการตรวจสอบข้อมูลก่อนนำมาใช้

ค่าที่หายไป Missing Value คือ ค่าที่ไม่ได้แสดงในข้อมูลที่เราได้เลือกแล้วหรือค่าที่ไม่สมบูรณ์ที่เราลบออกไประหว่างการทำ Noise Detection ค่าอาจจะหายไปเพราะเกิดจากความไม่รอบคอบของมนุษย์เพราะว่าไม่มีข้อมูลนั้นระหว่างการทำ Input ข้อมูลการจัดการกับค่าที่หายไปนั้นสามารถจัดการได้ด้วยเทคนิคที่ต่างกัน

2.6 การสำรวจและตรวจสอบข้อมูล (Data Cleaning and Exploration)

เมื่อทำการเก็บข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปที่ควรกระทำก็คือการตรวจสอบข้อมูลเหตุที่ต้องทำการตรวจสอบข้อมูลมี 2 ข้อ คือ ข้อแรก นักวิเคราะห์ควรมีความคุ้นเคยกับตัวข้อมูล ไม่ใช่รู้แต่ชื่อของ Attribute และความหมายของมันเท่านั้น แต่ต้องรู้ถึงเนื้อหา (Content) หรือความมุ่งหมายที่แท้จริงของข้อมูลด้วย ข้อสอง อาจมีความผิดพลาดของการเก็บสะสมข้อมูลเกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากฐานข้อมูลหลายๆ แหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้ในการวิเคราะห์ ซึ่งนักวิเคราะห์ที่ดีจะต้องทำการตรวจสอบข้อมูลเหล่านี้ให้ถูกต้องตัวอย่างของความผิดพลาดที่เกิดขึ้นได้แก่ความผิดพลาดในการเก็บข้อมูลจาก Attribute ที่ไม่ต้องการซึ่งเกิดจากความสับสนในการตั้งชื่อ Attribute นั้น (Mislabeling of Field) เช่นเราต้องการเก็บค่าของระดับการศึกษาของผู้สมัครเข้าศึกษาต่อซึ่งในความเป็นจริงถูกเก็บไว้ใน Attribute ที่ชื่อ “LEVEL_EDU” แต่ในฐานข้อมูลนั้นบังเอิญมี Attribute อีกตัวหนึ่งชื่อ “EDUCATION” ซึ่งเก็บระดับการศึกษาที่ผู้สมัครต้องการเข้าศึกษาซึ่งถ้าเราไม่ได้ตรวจสอบความสัมพันธ์และความมุ่งหมายที่แท้จริงของแต่ละ Attribute แล้วก็อาจเกิดการสับสนโดยเก็บข้อมูลของ Attribute “EDUCATION” ไปแทนก็ได้ซึ่งเมื่อนำข้อมูลที่ได้ไปทำ Data Mining ผลลัพธ์ที่ได้ก็จะผิดพลาดด้วย

2.7 การแปลงข้อมูล (Data Transformation)

ระหว่างขั้นตอนของการแปลงข้อมูล ข้อมูลที่ได้กลั่นกรองแล้วจะถูกแปลงให้เป็นรูปแบบของข้อมูลที่พร้อมจะถูกวิเคราะห์ รูปแบบของข้อมูลที่พร้อมจะถูกวิเคราะห์ คือรูปแบบของข้อมูลที่ไม่มีความขัดแย้งถูกจัดระเบียบมาอย่างเรียบร้อยกลั่นกรองมาจากแหล่งข้อมูลภายนอกและภายใน

ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมากเนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่า นักวิเคราะห์ ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของ Input อย่างไร ตัวอย่างเช่นหลักการรูปแบบของข้อมูลถูกกำหนดแล้วข้อมูลที่ถูกกลั่นกรองจะเหมาะสมกับรูปแบบเฉพาะสำหรับแต่ละกรรมวิธีของ Data Mining ที่จะถูกใช้การแปลงข้อมูลยังรวมไปถึงการทำ Data Recording และ Data Format Conversion เช่นการแปลงวันที่เป็นต้น

ทางสถิติการทำการแปลงข้อมูลยังมีเทคนิคของ Data Reduction จุดประสงค์เพื่อที่จะลดตัวแปรสำหรับการทำการ Process โดยการนำเอาตัวแปรตั้งแต่ 2 ตัวขึ้นไปมารวมกันแล้วทำการ Process ข้อดีก็คือลดจำนวนของตัวแปรลง และยังสามารถจัดการได้ง่ายขึ้น

อีกเทคนิคเรียกว่า Discretization โดยการแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยการแบ่งค่าของตัวแปรที่จะเป็น Input ให้เป็นช่วงๆ เช่นการแปลงเงินเดือนอายุ

อีกเทคนิคเรียกว่า One of N โดยการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric ตัวอย่างเช่น ชนิดของรถ Ford, Lincoln, Nissan ให้เป็น 100, 010, 001 ปกติแบบนี้มักจะเป็น Input ของพวก Neural Networks

2.8 Data Mining Task

1. Classification

ตัวอย่างนี้จะสร้างความเข้าใจใน Classification Study ซึ่งกรณีของตัวอย่างนี้พบได้ทั่วไปในวงการธุรกิจ นักวิเคราะห์ในองค์กรที่ทำธุรกิจเกี่ยวกับการสื่อสารแห่งหนึ่งต้องการเข้าใจว่าทำไมลูกค้าบางกลุ่มถึงยังคงซื่อสัตย์และมี Brand Loyalty สูงกับสินค้าขององค์กร แต่ในขณะเดียวกันลูกค้าอีกกลุ่มกลับไปหาคู่แข่งแทน ท้ายที่สุดนักวิเคราะห์จึงต้องการจะทำนายลักษณะและนิสัยของลูกค้าที่องค์กรจะต้องเสียไปให้คู่แข่ง

เนื่องจากขณะนี้นักวิเคราะห์มีเป้าหมายในใจเรียบร้อยแล้ว ดังนั้นนักวิเคราะห์จึงสามารถสร้าง Model ที่ข้อมูลต่างๆ ได้มาจากข้อมูลในอดีตของลูกค้าที่มีความซื่อสัตย์ต่อองค์กรและกลุ่มลูกค้าที่ไม่มีความซื่อสัตย์ต่อองค์กรด้วย Model ที่สมบูรณ์ถูกต้องจะสามารถทำให้องค์กรเข้าใจและทำนายลักษณะของธุรกิจที่จะเกิดขึ้นได้

จากตัวอย่างเหตุการณ์จะสามารถอธิบายขั้นตอนของการกำหนดการศึกษาได้การศึกษาจะกำหนดขอบเขตของกิจกรรมของ Data Mining ได้ นอกจากนี้การศึกษาจะสามารถกำหนดจุดประสงค์และข้อมูลที่ต้องการใช้ได้ทั้งหมดด้วยการกำหนดปัญหาทางธุรกิจนั่นก็เป็นสิ่งที่บอกให้นักวิเคราะห์ทราบได้เลยว่าขั้นตอนในการทำ Data Mining จะทำอย่างไรและจุดประสงค์ของการทำคืออะไร

ในการศึกษาต้องการหัวข้อในการศึกษาหัวข้อในการศึกษาอาจหมายได้ถึง Data Element ของ Object ที่เราต้องการจะศึกษา เช่น เราต้องการจะศึกษาถึง Object “ลูกค้า” ซึ่งมี Data Element ที่เกี่ยวข้องคือชนิดของลูกค้าแนวโน้มการซื้อสินค้าระยะเวลาที่เป็นลูกค้าขององค์กรและอื่นๆ ซึ่ง Data Element จะเป็นตัวกำหนดลักษณะและชนิดของลูกค้าการทำ Classification Studies นั้นเราสามารถกำหนดโครงร่างลักษณะเฉพาะหรืออุปนิสัยของลูกค้า ได้โดยดูได้จากตารางที่ 2.1

ตารางที่ 2.1 Data Element ของข้อมูล

ชื่อคอลัมน์	ชนิดของข้อมูล	ค่าที่ได้	คำอธิบาย
เบอร์โทร	Int	ค่าเฉพาะ	ตัวกำหนดเฉพาะสำหรับลูกค้า
ระยะเวลา	Int	จำนวนเต็ม	จำนวนที่ลูกค้าอยู่กับองค์กร
แนวโน้ม	Varchar	เพิ่มขึ้น, เหมือนเดิม, ลดลง	ตัวบ่งชี้แนวโน้มการใช้สินค้า 6 เดือนล่าสุด
สถานะ	Varchar	สูง, กลาง, ต่ำ, ไม่ทราบ	การสำรวจผลความพอใจของลูกค้า
ประเภทลูกค้า	Varchar	ซื่อสัตย์, ไม่ซื่อสัตย์	ลูกค้ายังคงอยู่กับองค์กรหรือเสียให้คู่แข่งไปแล้ว

จากตัวอย่างข้างต้นเรากำหนดให้ชนิดของลูกค้าเป็น Output หรือ Dependent Variable ซึ่งถูกใช้เป็นพื้นฐานในการศึกษาว่าอะไรคือสาเหตุที่ทำให้ลูกค้าซื่อสัตย์กับองค์กรและทำไมลูกค้าถึงจากองค์กรไปและเราจะใช้ Data Element ตัวอื่นๆ มาช่วยในการอธิบายสิ่งที่เกิดขึ้นเรากำหนดให้ชนิดของลูกค้าเป็น Training Data ถ้าเราเปลี่ยน Data Element ตัวอื่นมาเป็น Output จุดประสงค์ของการศึกษาก็จะเปลี่ยนไปด้วย

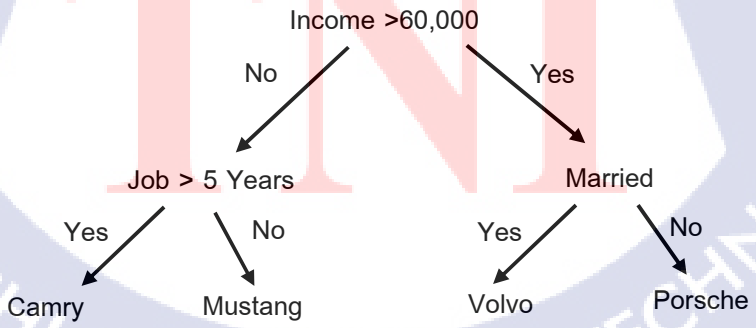
มีเทคนิคของ Data Mining จำนวนมากที่ใช้สำหรับปัญหาแบบ Classification และ Regression และแต่ละเทคนิคก็มี Algorithm มากมายแต่ละ Algorithm ก็ให้ผลลัพธ์ที่แตกต่างกัน ไปสิ่งที่แยกปัญหา Classification ออกจากแบบ Regression คือปัญหา Classification จะให้ผลลัพธ์เป็นค่าที่แน่นอนเช่น “ใช่”, “ไม่ใช่” หรือ “สูง”, “กลาง” และ “ต่ำ” เป็นต้นตัวอย่างเช่นแบบจำลองอาจทำนายว่า “นาย A จะตอบรับข้อเสนอของทางบริษัท” ในขณะที่ผลลัพธ์ที่จะได้จากปัญหาแบบ Regression เป็นค่าเฉพาะที่แน่นอนแต่ค่านี้จะไม่จำกัดคืออาจจะเป็นค่าอะไรก็ได้ตัวอย่างเช่นจากแบบจำลองที่ได้จากการทำ Data Mining แบบ Regression แบบจำลองอาจทำนายว่า “นาย A จะได้รับผลกำไร 500 บาท เป็นต้น”

โดยทั่วไปแล้วปัญหาในแบบ Regression จะสามารถเปลี่ยนเป็นปัญหาแบบ Classification ได้โดยการแบ่งค่าที่ต้องการทำนายให้เป็นกลุ่มของค่าที่ไม่ต่อเนื่องกัน (Discrete Categories) และปัญหาแบบ Classification ก็สามารถเปลี่ยนเป็นแบบ Regression ได้โดยการทำนายค่าหรือความน่าจะเป็นสำหรับแต่ละกลุ่มและกำหนดค่าของช่วงของค่าหรือความน่าจะเป็นที่ทำนายได้

เทคนิคของ Data Mining ที่ใช้ในการแก้ปัญหาแบบ Classification และ Regression เทคนิคที่ใช้ในการทำ Data Mining แบบ Classification และ Regression ที่ใช้กันในผลิตภัณฑ์ด้าน Data Mining ในปัจจุบัน ได้แก่

• Decision Tree เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้าเงื่อนไขแล้วผลลัพธ์” Decision Tree เป็นเทคนิคที่ค่อนข้างแพร่หลายเนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่ายเทคนิค Decision Tree จะจำกัดข้อมูลที่เป็นตัวแปรตาม (Dependent Variable) 1 ตัวต่อ 1 แบบจำลอง ถ้าต้องการทำนายตัวแปรตามหลายๆ ตัว จะต้องสร้างแบบจำลองสำหรับตัวแปรตามแต่ละตัว Algorithm ของเทคนิคแบบ Decision Tree ส่วนใหญ่ไม่รองรับข้อมูลแบบต่อเนื่อง (Continuous Data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) เสียก่อน Algorithm ที่เหล่านั้นได้ก่อน Chi-Squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 และ C5.0 Algorithm เหล่านี้ส่วนมากมักเหมาะกับปัญหาแบบ Classification Algorithm บางตัวปรับให้ใช้ได้กับปัญหาแบบ Regression เช่น Classification and Regression Trees (CART) ซึ่งรองรับทั้งปัญหาในแบบ Classification และ Regression นอกจากนี้ยังรองรับข้อมูลในแบบที่ต่อเนื่องด้วย

Decision Trees เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ ซึ่งมีการทำงานแบบ Supervised Learning คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดได้ก่อนล่วงหน้าที่เรียกว่า Training Set ได้อัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ด้วยรูปแบบของ Tree จะประกอบด้วย Node แรกสุดที่เรียกว่า Root Node จาก Root Node ก็จะแตกออกเป็น Node ลูก และที่ Node ลูกก็จะมีลูกของตัวเองซึ่ง Node ที่ระดับสุดท้ายจะเรียกว่า Leaf Node



รูปที่ 2.2 ตัวอย่างของ Decision Tree

จะเห็นว่าจาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้นซึ่งเส้นทางนี้จะอธิบายถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่มซึ่งในแต่ละ Leaf Node นั้นอาจเป็นกลุ่มเดียวกันซึ่งเกิดจากเหตุผลที่แตกต่างกันได้

อัลกอริทึมที่ใช้ในการสร้างแบบจำลองต้นไม้การตัดสินใจ

ID3 คือ วิธีการสร้างต้นไม้ตัดสินใจโดยมีพื้นฐานจากเทคนิค Divide-and-Conquer ที่ใช้ในการสร้างต้นไม้หรือที่เรียกว่า Top-Down Induction พัฒนาโดย J.Ross Quinlan (1975) แห่งมหาวิทยาลัย Sydney, Australia การสร้างต้นไม้ด้วย ID3 ใช้หลักของการใช้ Information Entropy ค่าที่วัดได้จะนำมาใช้ตัดสินใจเลือกตัวแปรในการสร้างเงื่อนไขในต้นไม้ตัดสินใจ

C4.5 เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยนำเอา ID3 มาปรับปรุงให้มีความสามารถมากขึ้นโดยใช้วิธีการ Information Gain เพิ่มเติมการจัดการกับข้อมูล ตัวเลขข้อมูลที่ขาดหายไปและไม่สมบรูณ์ (Missing Values, Noisy Data) และการ Prune ด้วยการแทน Branch ที่ไม่ช่วยในการตัดสินใจด้วย Leaf Node ที่ตัดสินใจได้ดีกว่าซึ่งสามารถหลีกเลี่ยงการสร้างโครงสร้างต้นไม้ที่ใหญ่เกินไป เนื่องจากมีข้อมูลจำนวนมาก อย่างไรก็ตามขึ้นอยู่กับการกำหนดความลึกอย่างไรเมื่อมีการเจริญเติบโตของ Decision Tree

C5.0/See5 คืออัลกอริทึมที่ J.Ross Quinlan พัฒนาเพิ่มจาก C4.5 มีความสามารถมากกว่าในด้านความเร็วในการประมวลผล การจัดการกับหน่วยความจำมีประสิทธิภาพมากขึ้น ต้นไม้ตัดสินใจมีขนาดเล็กลง ความถูกต้องของแบบจำลองต้นไม้มากขึ้นและมีการใช้ค่าถ่วงดุลน้ำหนักของข้อมูล แต่สำหรับอัลกอริทึมนี้เป็นลักษณะ Commercial และ Closed-Source Production ที่ไม่ได้เปิดเผย Source Code ดังเช่น ID3, C4.5 [6]

วิธีการที่ใช้สร้าง Decision Tree การนำข้อมูลมาสร้าง Tree มีขั้นตอนดังนี้

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

ข้อจำกัดของ Decision Tree

- การแบ่งกลุ่มแบบ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่องเช่นข้อมูลรายได้ข้อมูลราคาต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน

- เมื่อ Algorithm เลือกว่าจะใช้ค่าไหนเป็นตัวแบ่งกลุ่มแล้วก็จะไม่สนใจค่าอื่นที่อาจมีความสำคัญเช่นเดียวกัน
- การจัดการกับข้อมูลที่ไม่ทราบค่าอาจมีผลกระทบกับผลลัพธ์ของ Decision Tree
- Tree ที่มีระดับชั้นมากเกินไปจะทำให้ข้อมูลที่ผ่าน Node แตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้นจะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์
- ปัญหาเรื่อง Overfitting/Over Taining เกิดจากการที่แบบจำลองได้เรียนรู้เข้าไปถึงรายละเอียดของข้อมูลมากเกินไปจะทำให้เกิด Node ที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูลที่ใช้ในการเรียนรู้ซึ่งจะต้องหาวิธีการในการตัดกิ่งนี้ออกไป

งานวิจัยของ Jehn-Yih Wong; and Pi-Heng Chung [7]ที่ได้นำเทคนิคต้นไม้การตัดสินใจหรือ Decision Tree โดยใช้อัลกอริทึม C4.5 เข้ามาช่วยวิเคราะห์หาผลจากข้อมูลลูกค้าที่มีอยู่ โดยจะเน้นในการหาจากกลุ่มลูกค้าที่มีการเชื่อใจในสายการบินเพื่อศึกษาพฤติกรรมและการบริโภคการบริการ เพื่อนำมาสร้างแนวทางในการจูงใจและรักษากลุ่มลูกค้าดังกล่าวไว้ได้ ซึ่งศึกษาจากการทำการทดสอบการให้บริการพฤติกรรมและการบริโภคของผู้โดยสารที่มีความภักดีที่สนามบินไต้หวันจากสถานที่ที่แตกต่างกัน ส่วนใหญ่ประกอบด้วยผู้โดยสารจาก 11 กลุ่มและ 15 พนักงานภาคสาธารณะอย่างยิ่งเน้นบริการจองห้องพักและมีความพึงพอใจอย่างยิ่งในการให้ข้อมูลและความละเอียดการร้องเรียนและพวกเขาได้รับข้อมูลสายการบินโดยไม่ต้องหน่วยงานเพื่อพฤติกรรมการบริโภค

การค้นหาความสัมพันธ์ข้อมูลจากฐานข้อมูลของลูกค้าเป็นสิ่งสำคัญสำหรับการตัดสินใจเพื่อให้ได้คุณภาพสูงสุดทำให้ การศึกษาครั้งนี้การพัฒนาทำเหมืองข้อมูลเพื่อประเมินความจงรักภักดีผู้โดยสารและระบุ 16 กลุ่มของผู้โดยสารจงรักภักดีและซื่อสัตย์โดยต้นไม้ตัดสินใจ C4.5 การทำเหมืองแร่ การค้นหาความสัมพันธ์ข้อมูลเหล่านี้ให้การอ้างอิงที่มีประโยชน์มากสำหรับการวิจัยต่อไปเพื่อขยายการประยุกต์ใช้การจัดการข้อมูลผู้โดยสารระบบ (MIS) ในอุตสาหกรรมการบิน

โดยการศึกษาของงานวิจัยนี้มีข้อจำกัดบางอย่าง อย่างแรกคือขาดข้อมูลของคู่แข่งอื่นๆ เนื่องจากข้อมูลการวิจัยถูกทำให้แค่เพียงสายการบินไต้หวันเพียงอย่างเดียว อย่างไรก็ตามผลการวิจัยยังให้ ข้อมูลที่เป็นประโยชน์อื่นๆ กับสายการบินไต้หวันในการจัดการลูกค้าที่ภักดีเนื่องจาก 2 เหตุผล: กระบวนการการดำเนินงานเป็นคล้ายคลึงกันในหมู่สายการบินไต้หวันและสายการบินอื่นๆซึ่งเป็นมาตรฐานบริษัท ในไต้หวันการขนส่งทางอากาศตลาด ประการที่สองข้อมูลอาจจะขาดรายการที่สำคัญต่างๆ ที่อาจเกิดขึ้นในฐานข้อมูลของสายการบิน, เช่นรายการพฤติกรรมภักดี (เช่น ความถี่การซื้อ) และมันอาจส่งผลให้เกิดความไม่เพียงพอในการประเมินความจงรักภักดีผู้โดยสาร

- Naïve-Bayes เป็นเทคนิคที่ถูกตั้งชื่อตาม Thomas Bayes (1702-1761) เทคนิคแบบ Naïve- Bayes ใช้ทฤษฎี Bayes Theorem ในการคำนวณความน่าจะเป็นซึ่งถูกใช้ในการททำนายผลเมื่อทำการวิเคราะห์กรณีใหม่การทำนายผลทำได้โดยการรวมผลของตัวแปรอิสระ (Independent

Variable) ที่มีต่อตัวแปรตาม (Dependent Variable) Naïve-Bayes เป็นเทคนิคในการแก้ปัญหาแบบ Classification ที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ในทางทฤษฎีแล้วการทำนายผลของ Naïve-Bayes จะถูกต้องถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง ซึ่งในความเป็นจริงแล้วมีไม่มากนักที่ตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน ตัวอย่างเช่น ข้อมูลเกี่ยวกับประวัติบุคคล ซึ่งมักประกอบด้วยรายละเอียดย่อยมากมาย อาทิ น้ำหนัก, การศึกษา, รายได้ เป็นต้น จะเห็นว่ารายละเอียดเหล่านี้มักขึ้นอยู่กับอายุ ในกรณีนี้การใช้ Naïve-Bayes จะต้องคำนึงถึงผลของอายุให้มากๆ นอกจากนี้เทคนิคแบบ Naïve-Bayes ยังไม่รองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (Continuous Data) ด้วย ดังนั้นตัวแปรอิสระหรือตัวแปรตามที่มีค่าเป็นค่าต่อเนื่องจะต้องถูกแบ่งเป็นช่วงเช่น ถ้ามีตัวแปรอิสระที่เป็นค่าของอายุก็อาจแปลงค่าเหล่านั้นให้เป็นช่วงแคบๆ อาทิ “ต่ำกว่า 20 ปี”, “20-40ปี”, “40 ปีขึ้นไป” เป็นต้น ซึ่งการแบ่งช่วงนั้น ถ้าแบ่งไม่เหมาะสม ก็จะมีผลต่อคุณภาพของแบบจำลองที่สร้างขึ้น แต่ถ้าไม่คำนึงถึงข้อจำกัดนี้แล้ว เทคนิคแบบ Naïve-Bayes สามารถให้ผลลัพธ์ที่ดีและรวดเร็วได้ ความง่ายและความเร็วทำให้เทคนิคนี้เป็นเครื่องมือที่ดีในการสร้างแบบจำลองและหารูปแบบความสัมพันธ์ที่ไม่ซับซ้อน

- K-Nearest Neighbor (K-NN) เป็นเทคนิคที่เหมาะกับปัญหาแบบ Classification เทคนิคนี้แตกต่างจากเทคนิคอื่นตรงที่มันไม่ได้ใช้ข้อมูลฝึกหัด (Training Data) ในการสร้างแบบจำลองแต่จะใช้ข้อมูลนั้นมาเป็นตัวแบบจำลองเลย ในการใช้งาน K-NN Algorithm นั้นเราต้องระบุค่าตัวเลขจำนวนเต็มบวกให้กับ k ด้วย ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ Algorithm แบบ K-NN ได้แก่ 1-NN, 2-NN, 3-NN, ... K-NN โดยที่ k แทนเลขจำนวนเต็มบวก เช่น 4-NN หมายถึง Algorithm นี้จะค้นหา 4 กรณี ที่มีลักษณะใกล้เคียงกับกรณีใหม่ (4 Nearest Cases) ในการทำนายกรณีใหม่
- K-Mean Algorithm เทคนิคการจัดกลุ่มข้อมูลโดยอัลกอริทึมนั้นได้มีการพัฒนาและนำเสนอโดย Mac Queen ในปี 1967 ซึ่งแสดงให้เห็นถึงอัลกอริทึมในการจัดกลุ่มที่สมาชิกภายในแต่ละกลุ่มนั้น จะมีระยะใกล้จุดศูนย์กลางหรือตัวแทนของกลุ่ม (Mean) โดยขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Means นั้นจะประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อมูลแต่ละตัวเข้ากลุ่ม และสุดท้ายการปรับตัวแทนกลุ่มในแต่ละกลุ่มปัญหาที่พบ คือ ปัญหาการกำหนดจำนวนจุดเริ่มต้น ซึ่งส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มนั้นมีจำนวนสมาชิกน้อยเกินหรือในกลุ่มเลย จึงมีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดไปนั้นจะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา [8]

ขั้นตอนการจัดกลุ่มโดยอัลกอริทึม K-Means

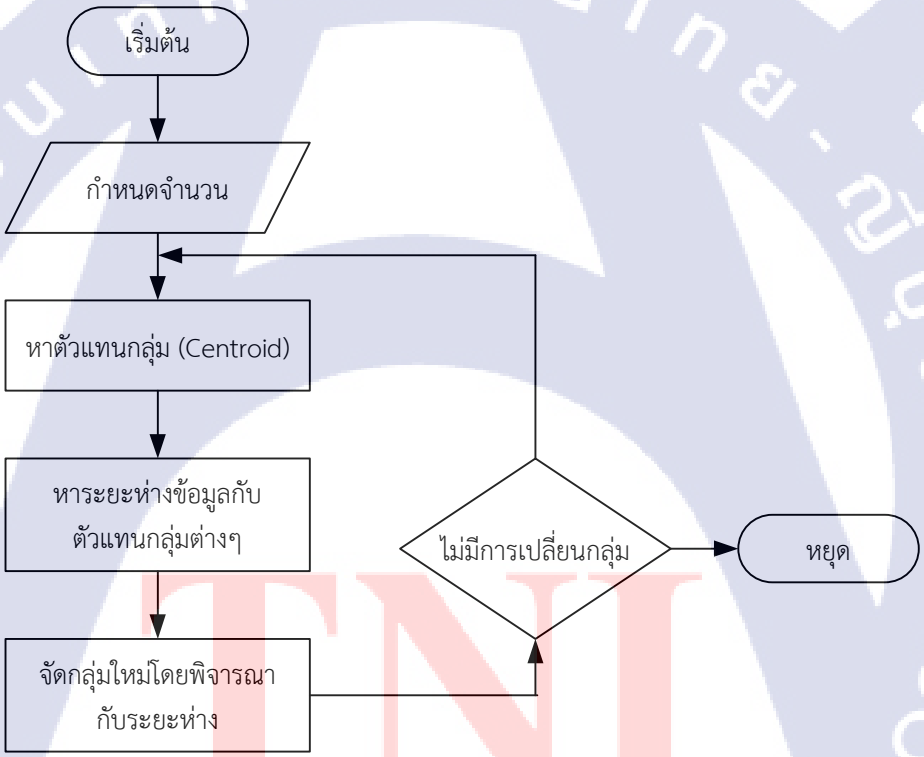
ขั้นตอนการจัดกลุ่มโดยเทคนิคอัลกอริทึม K-Means นั้นประกอบด้วย 4 ขั้นตอนดังต่อไปนี้

1. กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับ จำนวนกลุ่ม โดยที่ในทุกๆ กลุ่มนั้นจะต้องมีสมาชิกภายในกลุ่มเสมอ

2. คำนวณหาตัวแทนกลุ่ม หรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม

3. จัดกลุ่มข้อมูลใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อมูลในกลุ่มกับตัวแทนของกลุ่มต่างๆ ว่าข้อมูลนั้นสำควรที่จะอยู่กลุ่มใด

4. ทำการปรับปรุงการจัดกลุ่มโดนย้อนกลับไปทำข้อ 2 และจะหยุดเมื่อข้อมูลสมาชิกในกลุ่มแต่ละกลุ่มนั้นไม่มีการเปลี่ยนแปลงแล้ว



รูปที่ 2.3 ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Means

ตัวอย่าง การจัดกลุ่มโดยอัลกอริทึม K-Means โดยสมมุติให้มีข้อมูลอยู่ 4 รายการคือ A, B, C และ D ซึ่งแต่ละชุดข้อมูลมี 2 คอลัมน์ คือ X และ Y จากนั้นกำหนดให้มีการจัดกลุ่มข้อมูลเป็น 2 กลุ่ม (K=2) ดังตัวอย่างข้อมูลตารางที่ 2.2

ตารางที่ 2.2 ตัวอย่างข้อมูลเพื่อนำมาแบ่งกลุ่มโดย K-Means

ข้อมูล	X	Y
A	5	3
B	-1	1
C	1	-2
D	-3	-2

1. ให้สุ่มเลือกกลุ่มโดยจับคู่ ดังตัวอย่างจะจับคู่ A กับ B และ C กับ D หลังจากนั้นก็ทำการหาค่าเฉลี่ยจุดกึ่งกลาง จะได้ค่าตัวอย่างตารางที่ 2.3

ตารางที่ 2.3 ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม

กลุ่ม	ค่าเฉลี่ยจุดกึ่งกลาง X'	กลุ่ม Y'
AB	(5 + (1)) / 2 = 2	AB
CD	(1 + (-3)) / 2 = -1	CD

2. หาค่าระยะห่างระหว่างข้อมูลกับกลุ่มโดยใช้สูตาการหาระยะห่างแบบ Euclidean Distance ดังตัวอย่างต่อไปนี้

$A = (X_{1a}, X_{2a})$ และ $B = (X_{1b}, X_{2b})$

$$d^2(A,B) = (X_{1a} - X_{1a})^2 + (X_{2a} - X_{2b})^2$$

3. นำสูตามาคำนวณค่าระยะห่างระหว่างจุดข้อมูลกับกลุ่ม จากนั้นจึงจัดข้อมูลให้อยู่ในกลุ่มที่ซึ่งคำนวณค่าได้น้อยกว่า

$$d^2(A,(AB)) = (5 - 2)^2 + (3 - 2)^2 = 10$$

$$d^2(A,(CD)) = (5 + 1)^2 + (3 + 2)^2 = 61$$

4. จากนั้นจึงคำนวณสูตรกับข้อมูลที่เหลือให้ครบทั้งหมด

$$d^2(B,(AB)) = (-1 - 2)^2 + (1 - 2)^2 = 10$$

$$d^2(B,(CD)) = (-1 + 1)^2 + (1 + 2)^2 = 9$$

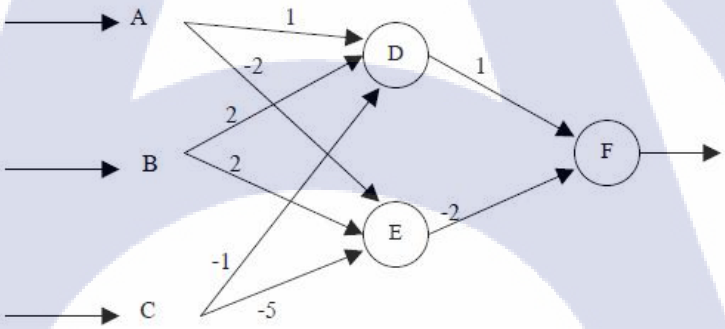
5. เมื่อได้กลุ่มข้อมูลใหม่แล้วก็ให้คำนวณหาค่าเฉลี่ยของแต่ละกลุ่มใหม่ดังตารางที่ 2.4

ตารางที่ 2.4 ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม

กลุ่ม	ค่าเฉลี่ยจุดกึ่งกลาง X’	กลุ่ม Y’
AB	5	A
BCD	-1	BCD

6. ทำซ้ำขั้นตอนที่กล่าวมาข้างต้นจนกว่าข้อมูลจะไม่มีการเปลี่ยนแปลงสุดท้ายแล้วจะได้กลุ่มข้อมูล 2 กลุ่ม โดยแบ่งเป็นกลุ่มที่ 1 คือ A และกลุ่มที่ 2 คือ BCD

- Neural Networks มีพื้นฐานมาจากแบบจำลองการทำงานของสมองมนุษย์ และก็สามารถใช้ได้ดีกับปัญหา Classification, Regression และ Clustering เทคนิคนี้มักถูกเรียกว่า “Black Box” เนื่องจากการทำงานของมันมีความซับซ้อนมากกว่าเทคนิคอื่นๆ ค่อนข้างมากผลลัพธ์ที่ได้ก็ยากต่อการทำความเข้าใจ



รูปที่ 2.4 ผลลัพธ์ของการใช้เทคนิคแบบ Neural Networks

เช่น ในรูปแสดงผลลัพธ์ของการใช้เทคนิคแบบ Neural Networks ในการวิเคราะห์ปัญหาความเสี่ยงของการให้กู้เงิน ซึ่งประกอบด้วยจุด 6 จุด A-F โดยที่ A, B, C เป็นจุดที่เป็นข้อมูลเข้า ซึ่งแทนตัวแปรอิสระ หนี้สิน (Debt), รายได้ (Income) และสถานภาพสมรส (Married) ในขณะที่จุด F เป็นผลลัพธ์ของการวิเคราะห์ แทนตัวแปรตามคือ ความเสี่ยง (Risk) และตัวเลขที่กำกับอยู่ตามเส้นลูกศรคือค่าถ่วงน้ำหนัก (Weight) เป็นต้น

ถึงแม้ว่าเทคนิค Neural Networks นี้จะทำงานได้ดีกับปัญหา Classification, Regression และ Clustering ก็ตามแต่มันเป็นเทคนิคที่ค่อนข้างซับซ้อนกว่าเทคนิคอื่น ซึ่งความซับซ้อนและการที่ไม่สามารถอธิบายได้ของผลลัพธ์ มักทำให้ผู้ใช้หลีกเลี่ยงเทคนิคนี้ อย่างไรก็ตาม เทคนิคนี้ก็มีข้อดีที่

สำคัญที่ไม่มีในเทคนิคอื่นๆ ก็คือ เทคนิคนี้ไม่มีข้อจำกัดเกี่ยวกับชนิดของความสัมพันธ์ เช่น สามารถสร้างแบบจำลองความสัมพันธ์ระหว่างตัวแปรตามกับสัดส่วนของตัวแปรอิสระ 2 ตัวได้ ซึ่งทำได้ยากถ้าใช้เทคนิคแบบ Decision Tree นอกจากนี้ เทคนิคแบบ Neural Networks ยังไม่มีปัญหากับความสัมพันธ์ที่เป็นแบบ Trigonometric หรือ Logarithmic โดยในการใช้งานจริงนั้นเทคนิคแบบ Decision Tree อาจให้ผลลัพธ์ที่ถูกต้องเพียงพอกับความต้องการ แต่ถ้าต้องการความแม่นยำมากๆ แล้วเทคนิคแบบ Neural Networks อาจเป็นหนทางที่ดีที่สุด ทางเดียวที่จะรู้ว่าควรใช้เทคนิคแบบ Neural Networks หรือไม่ ก็คือการเปรียบเทียบความเที่ยงตรงของแบบจำลองกับเทคนิคอื่น (Decision Tree) ถ้าไม่ได้ดีกว่ากันอย่างเห็นได้ชัด ก็ควรเลือกเทคนิคอื่น แต่ถ้าผลลัพธ์ที่ได้จากแบบจำลองของเทคนิค Neural Networks มีความเที่ยงตรงกว่าอย่างเห็นได้ชัด นั่นอาจหมายถึง เราต้องทำการปรับปรุงแบบจำลองของเทคนิค Decision Tree หรือบางทีการใช้เทคนิคแบบ Neural Networks อาจเหมาะสมสำหรับปัญหานี้มากที่สุดก็ได้

งานวิจัยของ Lisa Pritscher, L.; and Feyen [9] ที่ได้นำเทคนิคโครงข่ายประสาทเทียม หรือ Neural Network เข้ามาวิเคราะห์ในด้านการจัดการความสัมพันธ์ของลูกค้า (CRM) เพื่อใช้ในการกำหนดกลยุทธ์ทางการตลาด ซึ่งจาการศึกษาดังกล่าวได้ศึกษาไปถึงพฤติกรรมของกลุ่มลูกค้าที่ทำการใช้บริการบ่อยๆ โดยเลือกเปรียบเทียบกับลูกค้าทั้งหมดของสายการบิน ซึ่งจะค้นพบกลุ่มลูกค้าเพื่อนำมาใช้ในการวิเคราะห์ ตามภูมิภาคที่กำเนิด, ตามตลาดที่ลูกค้าต้องเดินทางบ่อยๆ, ตามความชอบในการเดินทางของลูกค้า หรือตามพฤติกรรมการเดินทาง และแสดงให้เห็นประโยชน์อย่างมากในการนำเทคนิคเหมืองข้อมูลมาช่วยในการตัดสินใจในธุรกิจสายการบิน

- Association Rule เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูล โดยหลักการทำงานของวิธีนี้ คือการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ หรือจาการวิเคราะห์การซื้อสินค้าของลูกค่าเรียกว่า “Market Basket Analysis” ซึ่งประเมินจากข้อมูลที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของบริษัท Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้ จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน เช่น Buys (X, Database) -> Buys (X, Data Mining) [80%, 60%] หมายความว่า เมื่อซื้อหนังสือ Database แล้วมีโอกาสที่จะซื้อหนังสือ Data Mining ด้วย 60% และมีการซื้อทั้งหนังสือ Database และหนังสือ Data Mining พร้อมๆ กัน 80%

พื้นฐานการหากฎความสัมพันธ์นั้นมีดังต่อไปนี้

- เซตไอเท็ม (I) คือเซตที่มีไอเท็มทั้งหมดเป็นสมาชิก ซึ่งในที่นี้อาจเป็นชื่อสินค้าหรือหน่วยพื้นฐานที่จะนำมาใช้

- รายการธุรกรรม (T) เป็นเซตย่อยของไอเท็ม โดยที่ $T \subseteq I$

- เซตข้อมูล (D) คือ เซตที่รายการธุรกรรมทุกตัวเป็นสมาชิก

โดยสามารถนิยามความสัมพันธ์ได้ว่า

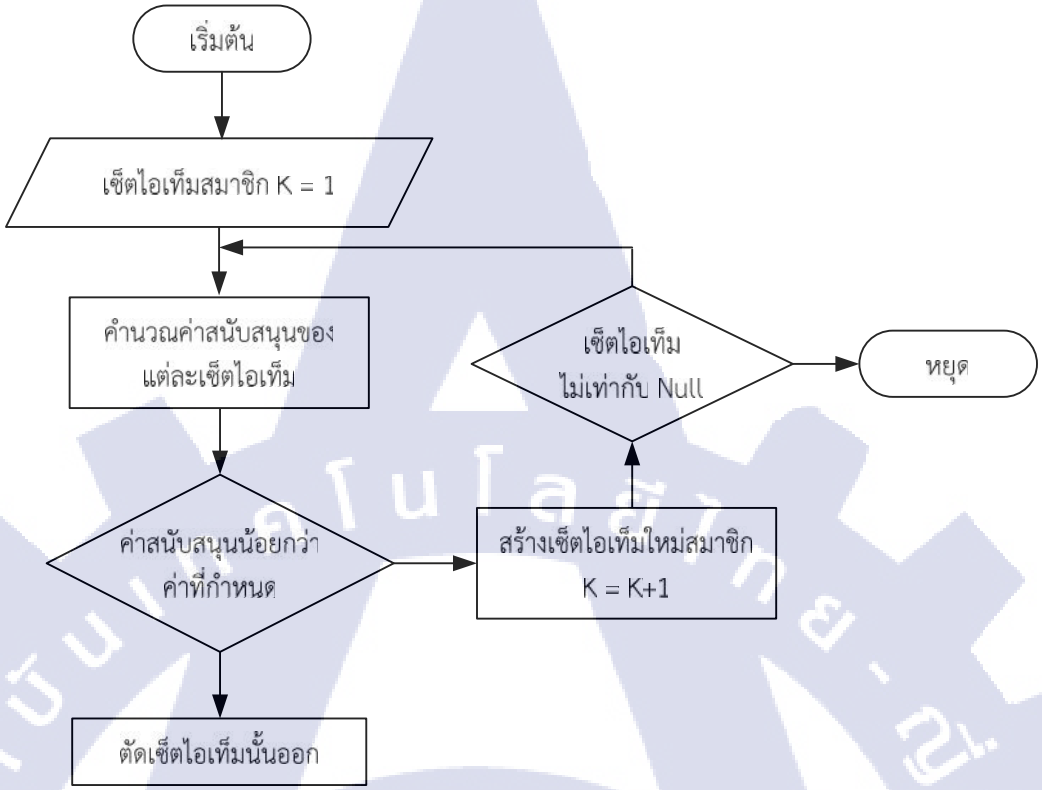
$$X \rightarrow Y \text{ เมื่อ } X \subset I,\ Y \subset I \text{ และ } X \cap Y = \phi$$

นอกจากนี้กฎความสัมพันธ์ทุกกฎจะประกอบด้วยค่าสนับสนุน (Support) และค่าความมั่นใจ (Confidence) ซึ่งมีนิยามดังนี้

- กฎความสัมพันธ์ $X \rightarrow Y$ มีค่าสนับสนุนเท่ากับ s ในเซตข้อมูล D ก็ต่อเมื่อ s% ของรายการธุรกรรมใน D บรรจุ $X \cup Y$

- กฎความสัมพันธ์ $X \rightarrow Y$ มีค่าความเชื่อมั่นเท่ากับ c ในเซตข้อมูล D ก็ต่อเมื่อ c% ของรายการธุรกรรมใน D ที่บรรจุ $X_{\text{บรรจุ}}Y$ ด้วย

• Apriori Algorithm [10] เป็นอัลกอริทึมพื้นฐานที่แพร่หลายในการหาความสัมพันธ์ของข้อมูล โดยใช้หลักการค้นหาแบบวงกว้างก่อนนับรายการธุรกรรม ซึ่งจะสร้างและตรวจสอบเซตไอเท็มทีละชั้น โดยเริ่มจากเซตไอเท็มที่มีจำนวนสมาชิกเท่ากับหนึ่ง ถ้าเซตไอเท็มใดมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนที่กำหนดก็จะตัดเซตไอเท็มนั้นออก ไม่นำไปสร้างเซตไอเท็มในชั้นต่อไป การทำงานของอัลกอริทึมจะวนไปเรื่อยๆ จนกระทั่งได้ทุกระดับชั้น หรือไม่เหลือเซตไอเท็มที่จะสร้างเซตไอเท็มในชั้นต่อไป



รูปที่ 2.5 ขั้นตอนการหาความสัมพันธ์โดยใช้ Apriori Algorithm

ข้อดีของ Apriori Algorithm

เป็นอัลการิทึมที่รู้จักกันดีในการหากฎความสัมพันธ์ เนื่องจากเป็นอัลกอริทึมที่ง่าย เป็นพื้นฐานในการพัฒนาอัลกอริทึมใหม่ๆ เพื่อลดเวลาในการหากฎความสัมพันธ์ และสามารถหากลุ่มข้อมูลปรากกฎบ่อยที่เชื่อถือได้ โดยข้อจำกัดของ Apiori Algorithm ในการหาค่าสนับสนุนของแต่ละ Candidate Itemsets จะต้องทำการอ่านข้อมูลในแต่ละรายการธุรกรรมทุกครั้ง เพื่อให้ได้กลุ่มข้อมูลจำนวน k ครั้ง (เมื่อ k คือ ขนาดของไอเท็มเซตที่ใหญ่ที่สุดที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ) เพื่อตรวจสอบค่าสนับสนุนของ Candidate Itemsets

- FP-Growth Algorithm เป็นโครงสร้างข้อมูลของ FP-Tree ประกอบด้วย 2 ส่วน คือ Header Table เป็นส่วนที่ใช้เก็บไอเท็มที่ผ่านค่าสนับสนุนขั้นต่ำโดยมีการเรียงลำดับตามค่าสนับสนุนจากมากไปหาน้อยและส่วนที่สอง คือ โครงสร้างต้นไม้ที่ใช้เก็บข้อมูลจากรายการธุรกรรม FP-Growth ได้ถูกนำเสนอขึ้นเพื่อใช้หากลุ่มข้อมูลปรากฏบ่อยและเพิ่มประสิทธิภาพในการทำงาน 3 เทคนิค คือ

1. การนำฐานข้อมูลขนาดใหญ่มาย่อเก็บในโครงสร้างข้อมูลที่มีขนาดเล็ก เพื่อหลีกเลี่ยงการอ่านรายการธุรกรรมหลายรอบ ซึ่งโครงสร้างข้อมูลดังกล่าวถูกกล่าวเรียกว่า Frequent Pattern Tree หรือ FP-Tree เพื่อเก็บข้อมูลและจำนวนจอง Frequent Patterns โดยแต่ละโหนดประกอบด้วย Frequent 1-Items ค่าสนับสนุนและพอยน์เตอร์เชื่อมโยงไปยังไอเท็มที่เหมือนกัน ซึ่งเรียงลำดับค่าสนับสนุนจากมากไปน้อย เนื่องจากโหนดที่เกิดบ่อยจะมีโอกาสเป็นโหนดที่มีความสัมพันธ์ร่วมกับโหนดอื่นมากกว่าโหนดที่เกิดขึ้นน้อยครั้ง

2. ใช้วิธีแบบ Pattern Fragment Growth Mining เพื่อหลีกเลี่ยงการสร้าง Candidate Itemsets จำนวนมากโดยเริ่มจากการหา Frequent 1-Items แล้วตรวจสอบเฉพาะ Conditional Pattern Base ซึ่งเป็นส่วนของรายการธุรกรรมที่ประกอบด้วยเฉพาะ Frequent Items มาสร้างเป็น FP-Tree แล้วใช้สร้าง Pattern Growth ซึ่งเกิดจากการสร้าง Condition FP-Tree มาต่อกัน

3. ใช้เทคนิคการแบ่งส่วนข้อมูล (Partitioning-Based) หรือ Devide-and-Conquer Method โดยแบ่งงานออกเป็นส่วนย่อยๆ ในขั้นตอนของการทำงาน เพื่อลดขนาดของ Condition Pattern Base ซึ่งถูกสร้างขึ้นเพื่อใช้ในการค้นหาในระดับต่างๆ ตามขนาดของ Conditional ที่เกี่ยวข้องใน FP-Tree และเป็นการแปลงปัญหาการหา Frequent Patterns ที่มีขนาดยาว โดยการหาจาก Pattern ที่มีขนาดสั้นแล้วนำมาต่อกัน

ขั้นตอนของ FP-Growth มี 2 ขั้นตอน คือ ขั้นที่ 1 สร้าง FP-Tree ประกอบด้วย 2 ขั้นตอนย่อย ดังนี้

1. การอ่านรายการธุรกรรมครั้งที่ 1 เพื่อหากลุ่มของข้อมูลที่ปรากฏบ่อยและ หาค่าสนับสนุนของกลุ่มข้อมูลที่ปรากฏบ่อย 1-Itemset และเรียงลำดับค่าสนับสนุนที่มีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำสุดจากมากไปหาน้อย

2. การอ่านรายการธุรกรรมครั้งที่ 2 เพื่อสร้าง FP-Tree โดยเริ่มสร้างโหนดรากของ FP-Tree และกำหนดค่าให้เป็น Null สำหรับแต่ละรายการธุรกรรมให้เลือกไอเท็มที่ผ่านค่าสนับสนุนขั้นต่ำสุด แต่ละรายการธุรกรรมทำการเรียงลำดับตามค่าสนับสนุนจากมากไปน้อยก่อนเพิ่มเขาไปในโครงสร้าง FP-Tree

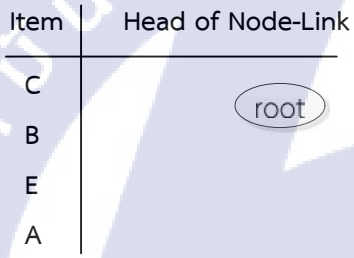
ตารางที่ 2.5 ตัวอย่างการหากลุ่มข้อมูลปรากฏบ่อยขนาด 1-Itemset

TID	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E
5	A, B, C, E

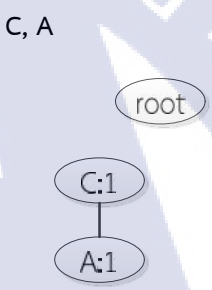
Scan Database →

TID	Items	Frequent Items
1	A, C	C, A
2	B, C, E	B, C, E
3	A, B, C, E	B, C, E, A
4	B, E	B, E
5	A, B, C, E	B, C, E, A

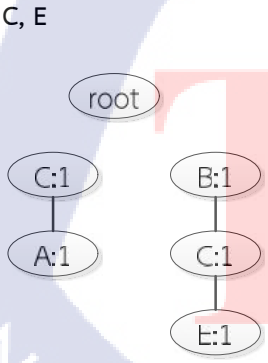
จากตารางที่ 2.5 สามารถหากลุ่มข้อมูลที่ปรากฏบ่อยขนาด 1-Itemset ได้ดังนี้ {C:4} {B:4} {E:4} และ {A:3} โดยกลุ่มข้อมูลที่ปรากฏบ่อยจะถูกเรียงลำดับตามค่าสนับสนุนจากมากไปหาน้อย และทำการจัดเรียงไอเท็มในแต่ละรายการธุรกรรมตามค่าสนับสนุนจากมากไปหาน้อย



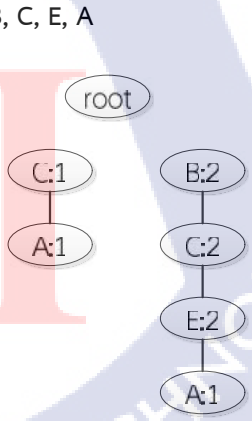
(ก) สร้าง Header



(ข) อ่าน TID 1 และเพิ่มไอเท็มCA

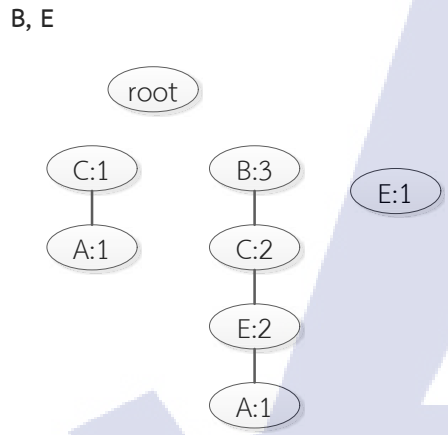


(ค) อ่าน TID 2 และเพิ่มไอเท็ม BCE

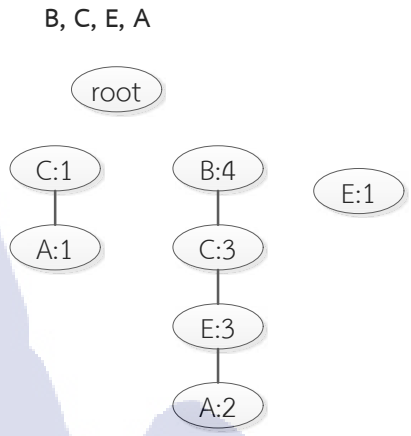


(ง) อ่าน TID 3 และเพิ่มไอเท็ม BCEA

รูปที่ 2.6 ตัวอย่างการสร้าง FP-Tree



(จ) อ่าน TID 4 และเพิ่มไอเท็ม BE



(ฉ) อ่าน TID 5 และเพิ่มไอเท็ม BCEA

รูปที่ 2.6 ตัวอย่างการสร้าง FP-Tree (ต่อ)

ขั้นตอนที่ 2 การหากลุ่มข้อมูลปรากฏบ่อยที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำจาก FP-Tree ที่สร้างจากขั้นตอนที่ 1

สำหรับการหา Frequent Patterns จากโครงสร้าง FP-Tree มีอินพุตของอัลกอริทึมคือ FP-Tree ที่สร้างตามอัลกอริทึมคือ เซตของ Frequent Construction และสัญลักษณ์ ξ หมายถึง ค่าสนับสนุนขั้นต่ำสุด เอาต์พุตของอัลกอริทึมคือ เซตของ Frequent Patterns ทั้งหมด α หมายถึง ไอเท็มเซตในทรานแซตชั่น และ β หมายถึง ไอเท็มเซตใน Conditional Pattern Base ของ α (α's Conditional Pattern Base) ตัวอย่างการหากลุ่มข้อมูลปรากฏบ่อยด้วยอัลกอริทึม FP-Growth โดยใช้ FP-Tree จากรูปที่ 2.6 จะเริ่มหา Pattern จาก Frequent Item ที่อยู่ด้านล่างของ Header Table ซึ่งจากอัลกอริทึม FP-Growth จะได้ Condition Pattern Base และ Condition FP-Tree ของแต่ละไอเท็มตามตารางที่ 2.6

ตารางที่ 2.6 ตัวอย่างการหา Frequent Patterns

Item	Conditional Pattern Base	Conditional FP-Tree
A	{(C:1),(B;2,C:2,E:2)}	{(B:2,C:3,E2)}|A
E	{(B:3,C:3),(B:1)}	{(B:4,C:3)}|E
C	{(B:3)}	{(B:3)}|C
B	Ø	Ø

เมื่อพิจารณาแต่ละไอเท็มในตารางที่ 2.6 สามารถหากลุ่มข้อมูลปรากฏบ่อยทั้งหมดได้ดังนี้ {A:3} {EA:2} {CA:3} {BA:2} {CEA:2} {BEA:2} {BCA:2} {BCEA:2} {E:4} {BE:4} {CE:3} {BCE:3} {C:4} {BC:3} และ {B:4}

ข้อดีของ FP-Growth Algorithm ใช้โครงสร้างของ FP-Tree สามารถเก็บข้อมูลในการหา Frequent Pattern ได้อย่างกระชับและเล็กกว่ารายการธุรกรรมเริ่มต้นมาก ทำให้ประหยัดเวลาในการตรวจสอบรายการธุรกรรม และ FP-Growth เป็นวิธีการหา Frequent Pattern ในรายการธุรกรรมที่จำนวนมากได้อย่างมีประสิทธิภาพ หลักเลี่ยงการสร้างและทดสอบ Candidate Itemsets

อัลกอริทึม Apriori, FP-Growth ได้เสนอวิธีการค้นหาความสัมพันธ์จากฐานข้อมูล ซึ่งตามกฎความสัมพันธ์ที่ได้จากการทำอัลกอริทึมเหล่านี้จำนวนมากมาย จนไม่สามารถวิเคราะห์กฎทั้งหมดได้ จึงมีการเสนอค่าวัดผลต่างๆ สำหรับการเลือกกฎที่มีความสำคัญ ถูกต้องน่าเชื่อถือ หรือหน่าสนใจจริงๆ ไปใช้ ในที่นี่จะสนใจค่าวัดผลดังต่อไปนี้ [11]

1. ค่าสนับสนุน (Support)
2. ค่าความเชื่อมั่น (Confidence)
3. ค่าคอนวิคชั่น (Conviction)
4. ค่าลิฟท์ (Lift)

ตารางที่ 2.7 ตัวอย่างข้อมูลรายการธุรกรรม

Transaction ID	Items
1	A B
2	B Y
3	C
4	A B Y
5	B

จากตารางที่ 2.7 กฎความสัมพันธ์แบบมีคลาสจะถูกกำหนดโดยรูปแบบกฎดังตัวอย่างข้างล่าง

A -> Y

เมื่อ

A = {A, B, C}

Y = {Y}

เมื่อคำนวณค่าวัดผลสำหรับกฎความสัมพันธ์แบบมีคลาส ค่าวัดผลแต่ละแบบจะคำนวณค่าอยู่บนความน่าจะเป็นตามสมการข้างล่าง

$$P(A) = Count(A)/|D|$$

$$P(Y) = Count(Y)/|D|$$

$$P(AY) = Count(AY)/|D|$$

โดยที่

1. Count(A) คือ จำนวนรายการธุรกรรม (Transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม A

2. Count(Y) คือ จำนวนรายการธุรกรรม (Transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม Y

3. Count(XY) คือ จำนวนรายการธุรกรรม (Transaction) หรือเร็คคอร์ดในฐานข้อมูลที่ประกอบไปด้วยไอเท็ม A และ Y

4. |D| คือ จำนวนรายการธุรกรรม (Transaction) หรือเร็คคอร์ดในฐานข้อมูล

ค่าสนับสนุน (Support)

ค่าสนับสนุนคือเปอร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูลเขียนอยู่ในรูปสมการดังนี้

$$Support(A) = P(A)$$

จากตารางที่ 2.7เราสามารถคำนวณค่าสนับสนุนได้ เช่น

$$P(A) = 2/5 = 0.4$$

$$P(B) = 4/5 = 0.8$$

$$P(C) = 1/5 = 0.2$$

$$P(Y) = 2/5 = 0.4$$

ค่าความเชื่อมั่น (Confidence)

ค่าความเชื่อมั่นคือเปอร์เซ็นต์ของจำนวน Itemsets ของกฎที่เกิดขึ้นในฐานข้อมูล ต่อจำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ เขียนอยู่ในรูปสมการดังนี้

Confidence (A->Y) = P(A and Y)/P(A)

จากตารางที่ 2.7 เราสามารถคำนวณค่าความเชื่อมั่นได้ เช่น

P(A->Y) = (1/5)/(2/5) = 0.5

P(B->Y) = (2/5)/(2/5) = 1

P(AB->Y) = (1/5)/(2/5) = 0.5

ค่าคอนวิคชั่น (Conviction)

ค่าคอนวิคชั่น คือ ผลคูณระหว่างเปอร์เซ็นต์ของจำนวน Itemsets ทางด้านซ้ายมือของกฎ และเปอร์เซ็นต์ของจำนวน Itemsets ที่เป็นคลาสแอททริบิวต์ยกเว้น Itemsets ที่เป็นคลาสแอททริบิวต์ทางด้านขวามือของกฎต่อเปอร์เซ็นต์ของ Itemsets ทางด้านซ้ายมือของกฎที่ไม่มีคลาสแอททริบิวต์ทางด้านขวามือของกฎประกอบอยู่ด้วยจะเขียนอยู่ในรูปสมการดังนี้

$$\text{Conviction } (A \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(X)P(Y)}$$

จากตารางที่ 2.7 เราสามารถคำนวณค่าคอนวิคชั่นได้ เช่น

P(A->Y) = (2/5)*(3/5)/(1/5) = 1.2

P(B->Y) = (3/5)*(3/5)/(1/5) = 1.8

P(AB->Y) = (2/5)*(3/5)/(1/5) = 1.2

ค่าลิฟท์ (Lift)

ค่าลิฟท์ คือ เปอร์เซ็นต์ของจำนวน Itemsets ของกฎที่เกิดขึ้นในฐานข้อมูล ต่อจำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ และจำนวน Itemsets ที่เกิดขึ้นทางด้านขวามือของกฎเขียนอยู่ในรูปสมการ

$$\text{Lift}(X->Y) = \frac{P(X \text{ and } Y)}{P(X)P(Y)}$$

จากตารางที่ 2.7 เราสามารถคำนวณค่าความลิฟท์ได้ เช่น

P(A->Y) = (1/5)/((2/5)*(2/5)) = 0.5

P(B->Y) = (2/5)/((4/5)*(2/5)) = 1.25

P(AB->Y) = (1/5)/((2/5)*(2/5)) = 0.2

2. Estimation/Prediction

ลักษณะของ Classification นั้นคำนึงถึงผลกำหนดที่ออกมาชัดเจนว่าคุณสมบัติดังกล่าวจะอยู่ในชั้นใดแต่ Estimation เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติดังกล่าวให้ชัดเจนเป็นการจัดการกับค่าที่มีผลในการวัดที่ต่อเนื่องตัวอย่างเช่น

- การประเมินรายได้ของครอบครัว
- การประเมินความสูงของบุคคลในครอบครัว
- การประเมินจำนวนของเด็กๆ ในครอบครัว

Prediction เหมือนกับ Classification และ Estimation ยกเว้นว่า Record ที่ถูกแยกจัดลำดับนั้นเกิดขึ้นตามการทำนาย พฤติกรรมในอนาคตหรือการทำนายค่าที่จะเกิดขึ้นในอนาคตข้อมูลในอดีตจะถูกสร้างเป็น Model ขึ้นมาเพื่อทำนายหรืออธิบายสิ่งที่จะเกิดขึ้นในอนาคตตัวอย่างเช่น

- การทำนายว่าลูกค้ากลุ่มใดที่องค์กรจะสูญเสียไปภายใน 6 เดือนหน้า
- การทำนายว่ายอดซื้อของลูกค้าจะเป็นเท่าใดถ้าบริษัทลดราคาสินค้า 10%

3. Description/Visualization

Description จุดประสงค์ของการทำ Data Mining การหาคำอธิบายถึงสิ่งที่จะเกิดขึ้นโดยอาศัยข้อมูลจากฐานข้อมูล ตัวอย่างเช่น ผู้หญิงจะสนับสนุนพรรคการเมืองพรรคหนึ่งมากกว่าผู้ชาย Visualization เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิกการนำเสนอจะสามารถทำได้มากกว่า 2 มิติ ซึ่งจะสร้างความละเอียดของการนำเสนอและสร้างความเข้าใจให้มากขึ้น ตัวอย่างเช่น องค์กรต้องการที่จะหาสถานที่ในการตั้งสาขาขององค์กรในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้รูปแผนที่ที่มีการ Plot ที่ตั้งขององค์กรคู่แข่งที่มีสาขาตั้งอยู่ในเขตนั้น เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด Data Visualization จะใช้มากกับ Data Mining Tools สิ่งที่สำคัญของ Visualization ก็คือไม่สามารถเน้นการวิเคราะห์ข้อมูลที่มีประสิทธิภาพ ในขณะที่แบบแผนทางสถิติและ Confirmatory Analysis เป็นการสร้างการวิเคราะห์ข้อมูลที่แท้จริง

2.9 เครื่องมือในการทำ Data Mining

ในการทำเหมืองข้อมูลนี้จะใช้วิธีของต้นไม้ตัดสินใจในการค้นว่านี้เลือกใช้อัลกอริทึมแบบ K-Mean, C4.5 และ Association Rule โดยทำการสร้างผ่านโปรแกรม Rapid Miner Studio ซึ่งเป็นซอฟต์แวร์ทาง Data Mining ที่ได้รับการโหวตว่ามีผู้ใช้งานมากที่สุดจากเว็บไซต์ KDnuggets.com เมื่อปี 2013

RFM ย่อมาจากคำว่า Recency, Frequency และ Monetary ตามลำดับ โดยทั้งสามตัวแปรนี้จะช่วยวิเคราะห์ลูกค้าในฐานข้อมูลได้อย่างมีประสิทธิภาพและยังช่วยให้นักการตลาดสามารถกำหนดกลยุทธ์ทางธุรกิจในแต่ละกลุ่มลูกค้าได้อย่างถูกต้องอีกด้วย หลักการง่ายๆ อย่าง RFM จะช่วยวิเคราะห์ได้อย่างมหาศาล โดยลูกค้าแต่ละรายจะถูกกำหนดคะแนนในแต่ละตัวแปรทั้งสาม ด้วยคะแนน 1, 2 และ 3 โดย 1 คะแนนหมายถึงค่าต่ำสุดและ 3 คะแนน หมายถึง ค่าสูงสุด โดยค่าตัวแปรทั้ง 3 ประเภท คือ

Recency คือ ระยะเวลาที่ลูกค้ามีการสั่งซื้อล่าสุดจนถึงปัจจุบัน เช่น ลูกค้าที่สั่งซื้อล่าสุดเมื่อเดือนที่แล้วก็จะได้รับคะแนนสูงกว่าลูกค้าที่สั่งซื้อล่าสุดเมื่อปีที่แล้ว เป็นต้น

Frequency คือ จำนวนการสั่งซื้อในแต่ละช่วงเวลา เช่น ลูกค้าที่สั่งซื้อ 6 ครั้ง จะได้รับคะแนนสูงกว่าลูกค้าที่สั่งซื้อเพียง 1 ครั้ง ในช่วงระยะเวลา 3 ปี เหมือนกัน เป็นต้น

Monetary คือ มูลค่าเฉลี่ยต่อการสั่งซื้อแต่ละครั้งในช่วงเวลาเดียวกัน เช่น ลูกค้าที่สั่งซื้อโดยเฉลี่ย 500,000 บาท จะมีคะแนนสูงกว่าลูกค้าที่สั่งซื้อโดยเฉลี่ย 200,000 บาท (มูลค่าการสั่งซื้อรวมในระยะเวลา 3 ปีหารด้วยจำนวนการสั่งซื้อทั้งหมดในระยะเวลา 3 ปี) เป็นต้น

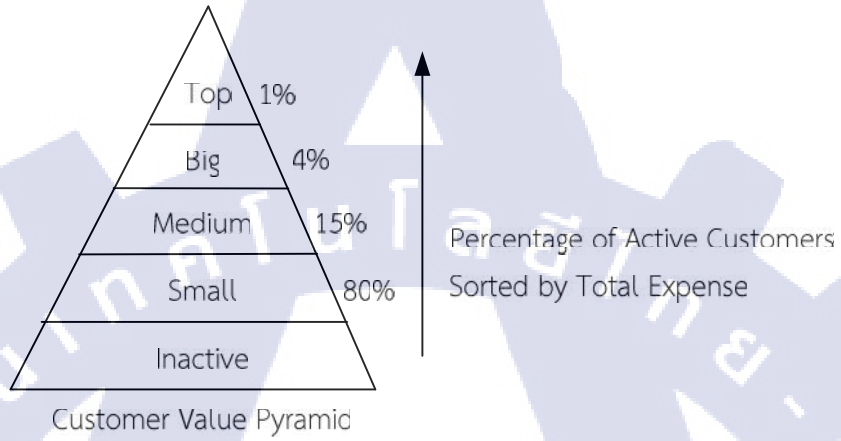
ซึ่งการกำหนดคะแนนจะขึ้นอยู่กับรูปแบบของแต่ละธุรกิจ เช่น ธุรกิจขายรถยนต์ อาจจะให้คะแนน ดังนี้ Recency 1 = ลูกค้าที่สั่งซื้อครั้งล่าสุดนานมากกว่า 48 เดือน, 2 = ลูกค้าที่สั่งซื้อครั้งล่าสุดนานมากกว่า 12 เดือน แต่ไม่ถึง 48 เดือน, 3 = ลูกค้าที่สั่งซื้อครั้งล่าสุดภายใน 12 เดือน

Frequency 1 = ลูกค้าที่สั่งซื้อ 1 ครั้ง ภายใน 60 เดือน, 2 = ลูกค้าที่สั่งซื้อ 2 ครั้ง ภายใน 60 เดือน, 3 = ลูกค้าที่สั่งซื้อมากกว่า 3 ครั้ง ภายใน 60 เดือน

Monetary 1 = ลูกค้าที่สั่งซื้อโดยเฉลี่ย ไม่เกิน 300,000 บาท, 2 = ลูกค้าที่สั่งซื้อโดยเฉลี่ยระหว่าง 300,000-600,000 บาท, 3 = ลูกค้าที่สั่งซื้อโดยเฉลี่ย มากกว่า 600,000 บาท

ดังนั้น จะได้คะแนนลูกค้า (333) คือ ลูกค้าที่สั่งซื้อล่าสุดไม่เกิน 12 เดือน โดยมีจำนวนสั่งซื้อมากกว่า 3 ครั้ง ด้วยมูลค่าเฉลี่ยมากกว่า 600,000 บาทต่อครั้ง ในทางตรงกันข้าม ก็จะมีลูกค้าด้วยคะแนน (111) เช่นกันหลังจากที่กำหนดคะแนนของลูกค้าทั้งหมดในฐานข้อมูลแล้ว จะได้กลุ่มลูกค้าทั้งหมด 27 กลุ่ม จากลูกค้าที่มีคุณค่ามากที่สุด ด้วยคะแนน (333) จนไปถึง ลูกค้าที่มีคุณค่าน้อยที่สุด ด้วยคะแนน (111) [12]

RFM Analysis เป็นเทคนิคหนึ่งในการตลาดซึ่งสามารถใช้ในการแบ่งกลุ่มลูกค้า จัดลำดับความสำคัญและแสดงถึงกลุ่มลูกค้าที่มีความสำคัญ ตามแนวคิด 80/20 ซึ่งหมายถึงร้อยละ 20 ของจำนวนลูกค้าทั้งหมดจะนำมาซึ่งผลกำไร 80% สำหรับการแบ่งกลุ่มนั้นได้มีการนำเสนอพีรามิดโมเดล เป็นโมเดลในการแบ่งกลุ่มโดยพิจารณาจากจำนวนกำไรที่ได้จากลูกค้าดังรูปที่ 2.7



รูปที่ 2.7 พีรามิดโมเดล (Pyramid Model)

<u>ขั้นตอนวิธีการแบ่งกลุ่มข้อมูลด้วย RFM Analysis</u>

การแบ่งกลุ่มทางการตลาดแบบ RFM Analysis มี 4 ขั้นตอนต่อไปนี้

1. จัดเตรียมข้อมูลลูกค้าเพื่อเตรียมการคำนวณ โดยแบ่งข้อมูลเป็น 3 ส่วน ดังนี้

ก. ข้อมูลการเข้ามาทำรายการล่าสุดของลูกค้า (Recency) โดยระบุชื่อลูกค้า หรือชื่อบริษัทพร้อมวันที่ทำรายการ โดยวันที่ทำรายการสำหรับลูกค้าละรายนั้นให้เลือกเป็นวันที่ทำรายการล่าสุด

ข. ข้อมูลความถี่ในการเข้าใช้บริการ (Frequency) โดยระบุชื่อลูกค้าพร้อมจำนวนรายการทั้งหมดที่ใช้บริการภายในช่วงเวลาที่กำหนด

ค. ข้อมูลรายได้รวม (Monetary) จากลูกค้าแต่ละราย โดยให้ระบุชื่อลูกค้าพร้อมจำนวนเงินรวมที่ลูกค้าได้ใช้บริการ

2. ทำการเรียงลำดับข้อมูลแต่ละส่วน โดยเรียงจากมากไปน้อย จากนั้นทำการกำหนดค่าให้แต่ละรายการ โดยการแบ่งกลุ่มเงื่อนไขออกเป็น 5 กลุ่ม เช่นข้อมูลการเข้าทำรายการจะถูกเรียงลำดับจากวันที่ใกล้เคียงปัจจุบันไปจนถึงวันที่ต้นปี ดังนั้นเมื่อเริ่มทำการกำหนดค่า จะได้ว่าลูกค้าที่มัน

วันที่เข้าทำรายการล่าสุดใกล้เคียงกับวันที่ปัจจุบัน จะมีค่าลำดับความสำคัญจากการเข้าใช้บริการ (R Score) เท่ากับ 5 และถัดไปจะเป็น 4, 3, 2 และ 1 ตามลำดับ

3. นำค่า R Score, F Score และ M Score ของแต่ละลูกค้ามารวมกัน เพื่อคำนวณหาค่า RFM Score

RFM Score = R Score + F Score + M Score

4. นำ RFM Score มาเรียงลำดับจากมากไปน้อย ซึ่งค่า RFM Score มากจะบอกถึงว่าลูกค้านั้น คือ ลูกค้าชั้นดี

ตัวอย่าง การคำนวณ RFM Analysis

ตารางที่ 2.8 ข้อมูลรายละเอียดการทำรายการของลูกค้า

Customer Name	Transaction Date	Amount (baht)
Customer A	31/10/2013	15,000.00
Customer B	24/10/2013	500.00
Customer A	24/09/2013	1,250.00
Customer A	20/08/2013	650.00
Customer C	11/11/2013	4,500.00
...	...	...

สมมติข้อมูลการทำรายการของลูกค้ามีดังตารางที่ 2.8 การคำนวณหาค่า RFM Analysis สามารถทำได้ตามขั้นตอนต่อไปนี้

1. จัดเตรียมข้อมูลลูกค้าเพื่อเตรียมการคำนวณ ดังต่อไปนี้

1.1 R Score เลือกวันที่ใช้บริการล่าสุด สำหรับลูกค้าแต่ละราย

1.2 F Score จำนวนความถี่ที่ใช้บริการ โดยนับจากรายการทั้งหมด

1.3 M Score กำไรรวม หรือยอดรวม โดยนับจากมูลค่าเงินของรายการทั้งหมดของลูกค้ารายนั้นๆ จะได้ตารางข้อมูลดังตารางที่ 2.7 ต่อไปนี้

ตารางที่ 2.9 ข้อมูลเพื่อเตรียมคำนวณค่า RFM

Customer Name	Recency	Frequency	Monetary (Baht)
Customer A	31/10/2013	3	16,900.00
Customer B	24/10/2013	1	500.00
Customer C	24/09/2013	2	4,500.00
Customer D	01/11/2013	5	9,000.00
Customer E	25/10/2013	1	1,200.00
...	...	...	...

2. เรียงลำดับข้อมูลแต่ละค่าจากมากไปน้อย โดยให้ทำทีละค่า จากนั้นจึงจำกำหนดค่าให้โดย 20% ของจำนวนข้อมูลจากรายการแรกให้มีค่าเท่ากับ 5 สำหรับ 20% ถัดไปมีค่าเท่ากับ 4, 3, 2 และ 1 ตามลำดับ ดังตารางที่ 2.10, 2.11 และตารางที่ 2.12 ต่อไปนี้

ตารางที่ 2.10 การคำนวณค่า R Score

Customer Name	Recency	R Score
Customer D	01/11/2013	5
Customer A	31/10/2013	5
Customer E	25/10/2013	4
Customer B	24/10/2013	4
Customer C	24/09/2013	3
...	...	...

ตารางที่ 2.11 การคำนวณค่า F Score

Customer Name	Frequency	F Score
Customer D	5	5
Customer A	3	5
Customer C	2	4
Customer E	1	4
Customer B	1	3
...	...	...

ตารางที่ 2.12 การคำนวณค่า M Score

Customer Name	Monetary (baht)	M Score
Customer A	16,900.00	5
Customer D	9,000.00	5
Customer C	4,500.00	4
Customer E	1,200.00	4
Customer B	500.00	3
...	...	...

3. นำค่า R Score, F Score และ M Score มาคำนวณตามสูตรสมการ

RFM Score = R Score + F Score + M Score จะได้ค่า RFM ตามตารางที่ 2.13

ตารางที่ 2.13 วิธีการคำนวณค่า RFM

Customer Name	R Score	F Score	M Score	RFM Score
Customer A	5	5	5	5+5+5 = 15 หรือ 555
Customer	4	3	3	4+3+3 = 10 หรือ 433
Customer	3	4	3	3+4+4 = 10 หรือ 344
Customer	5	5	5	5+5+5 = 15 หรือ 555
Customer	4	4	4	4+4+4 = 12 หรือ 444
...	...	...	...	...

4. นำข้อมูลเรียงลำดับค่า RFM Score ซึ่งจะทำให้ทราบว่าลูกค้ารายใดที่เป็นกลุ่มลูกค้าชั้นดีตามรายละเอียดตารางที่ 2.14 ซึ่งจากตัวอย่างภายหลังการคำนวณค่า RFM จะเห็นว่า Customer A และ Customer D คือ กลุ่มลูกค้าที่มีคุณค่า หรือลูกค้าชั้นดีของบริษัท

ตารางที่ 2.14 ผลการจัดอันดับความสำคัญของข้อมูลตามค่า RFM

Customer Name	RFM Score
Customer A	15
Customer D	15
Customer E	12
Customer B	10
Customer C	10
...	...

Clustering Using RFM ได้นำเทคนิคการจัดกลุ่มหน้าที่ใน การศึกษา RFM, ได้นำเสนอวิธีการใช้ K-Means++ เพื่อหากลุ่มลูกค้าที่มีค่า RFM ที่คล้ายกันของขั้นตอนวิธีการจัดกลุ่มอื่นๆ เช่น K-Means หมายถึง แผนที่จะจัดตนเองเพราะข้อดีของมันในแง่ของคุณภาพของการจัดกลุ่ม

Classification Using RFM ได้ใช้ขั้นตอนวิธีการจัดหมวดหมู่โดยใช้กลุ่มลูกค้าโดยการค้นพบขั้นตอนวิธีการจัดกลุ่มและนำเสนอการค้นพบกฎการจัดหมวดหมู่โดยพิจารณาตัวแปรด้านประชากรศาสตร์ของลูกค้า เช่น อายุ เพศ อาชีพและ สถานะสมรส

Association Rule Mining Using RFM ใช้ในการค้นหาความสัมพันธ์ที่ซ่อนอยู่ในระหว่าง Transection ในฐานข้อมูล โดยใน Association Rule นั้นจะหาข้อมูลที่สอดคล้องกันที่แข็งแกร่งในฐานข้อมูล

งานวิจัยของ Derya Birant [13] ได้มีการเอา RFM เข้ามาใช้กับ Data Mining เพื่อหาค่า Recency, Frequency และ Monetary ในการวิเคราะห์ RFM เป็นเทคนิคการตลาดที่ใช้ในการวิเคราะห์พฤติกรรมของลูกค้าเช่นเมื่อเร็วๆ นี้วิธีการที่ลูกค้าได้ซื้อ (Recency) บ่อยครั้งที่ลูกค้าซื้อ (Frequency) และเท่าใดลูกค้าใช้จ่าย (Monetary) มันเป็นวิธีที่มีประโยชน์ในการปรับปรุงการแบ่งส่วนลูกค้าโดยแบ่งลูกค้าออกเป็นกลุ่มต่างๆ สำหรับการให้บริการส่วนบุคคลในอนาคตและเพื่อระบุลูกค้าที่มีแนวโน้มที่จะตอบสนองต่อการโปรโมชั่น โดยในงานวิจัยได้นำ RFM เข้ามาใช้กับงานของ Data Mining ทั้งรูปแบบ Clustering, Classification, Association Rule ซึ่งในงานวิจัยได้บอกแนวคิด RFM เริ่มต้นโดย Bult and Wansbeekได้รับการพิสูจน์ว่าการนำ RFM เข้าไปใช้มีประสิทธิภาพมากเมื่อนำไปใช้กับฐานข้อมูลการตลาด ในงานวิจัยได้ทำการเริ่มการวิเคราะห์โดยเริ่มจากความใหม่ ความถี่ และค่าของเงิน โดยเรียงลำดับตามความใหม่ของระยะเวลา คือ มีการซื้อล่าสุดเมื่อไหร่ ไปถึงต่ำสุด และให้คะแนนตามเรียงจากด้านบนลงล่างเป็นส่วนๆ ละ 20% 5, 4, 3, 2, 1 และในส่วนของการให้คะแนนค่าของเงินจะเรียงตามลำดับมูลค่าจากมากสุดไปน้อยที่สุด และสุดท้ายนำมาจัดอันดับต่อกัน

บทที่ 3

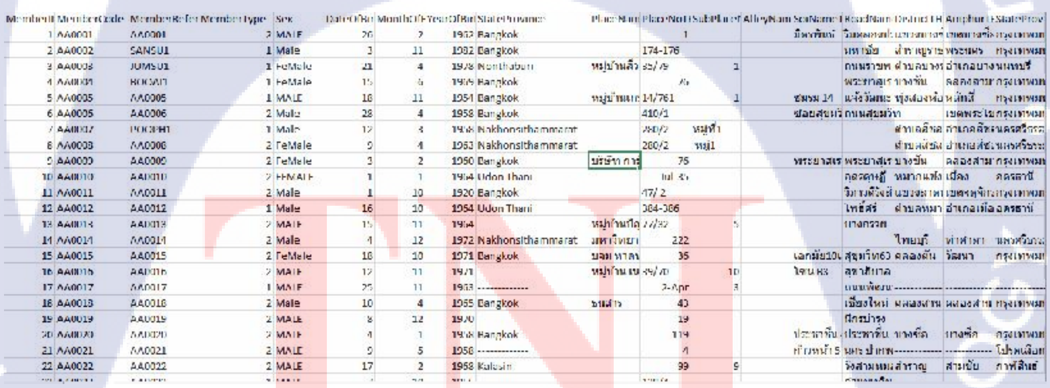
วิธีการดำเนินงานวิจัย

3.1 ความเข้าใจในธุรกิจ (Business Understanding)

การศึกษาครั้งนี้เป็นการศึกษาถึงพฤติกรรมของกลุ่มสมาชิกนกแฟนคลับ โดยข้อมูลจะถูกนำมาทำการวิเคราะห์ถึงพฤติกรรมการใช้บริการของสายการบินซึ่งก็คือการทำเหมืองข้อมูล การทำเหมืองข้อมูลจะเป็นการเสริมสร้างกระบวนการส่งเสริมการขายให้กับสายการบินเพราะทุกขั้นตอนจะต้องมีการตัดสินใจบนพื้นฐานของข้อมูลที่หนักแน่นเพียงพอ ซึ่งเป็นเรื่องที่การศึกษาในครั้งนี้ให้ความสำคัญในการทำการศึกษา เนื่องจากส่วนงานด้านการหาความสัมพันธ์ระหวางบริการ มีเป้าหมายในการสร้างความจงรักภักดี (Loyalty) ของลูกค้าที่มีต่อสายการบิน สามารถนำเสนอบริการที่ตรงกับความต้องการได้ เพื่อลูกค้าเกิดความพึงพอใจและกลับมาใช้บริการซ้ำ

ข้อมูลที่ใช้ในการทดสอบทำเหมืองข้อมูลนี้นำมาจากฐานข้อมูลของ Nok Airlines Public Company Limited โดยคัดเลือกข้อมูลมาจากสมาชิกกลุ่มนกแฟนคลับโดยจะเน้นศึกษาจากข้อมูลสมาชิกที่มีอยู่แล้วจำนวน 290,000 ราย

แหล่งที่มาของข้อมูลตัวอย่างที่ใช้ในการศึกษานำมาจาก ข้อมูลการใช้บริการของสมาชิกกลุ่มนกแฟนคลับจากฐานข้อมูลนกแฟนคลับทั้งปี 2012-2013 จำนวน 149,831 ราย



MemberI	MemberCode	MemberRefer	MemberType	Sex	DateOfBir	MonthOfB	YearOfBir	StateProvince	PlaceNam	PlaceNo	SubPlace	AlleyNam	SoiName	RoadNam	DistrictTH	AmphurT	StateProv
1	AA0001	AA0001	2	MALE	26	2	[illegible]	Bangkok		1			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	AA0002	SANSU1	1	Male	3	11	1982	Bangkok		174-176				มหาชัย	[illegible]	พระนคร	กรุงเทพม
3	AA0003	JUMSU1	1	FeMale	21	4	1978	Nonthaburi	หมู่บ้าน[illegible]	[illegible]	1			[illegible]	[illegible]	[illegible]	[illegible]
4	AA0004	[illegible]	1	FeMale	15	6	[illegible]	Bangkok		[illegible]				[illegible]	[illegible]	[illegible]	[illegible]
5	AA0005	AA0005	1	MALE	18	11	1954	Bangkok	หมู่บ้าน[illegible]	14/761	1		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	AA0006	AA0006	2	Male	28	4	1958	Bangkok		410/1			[illegible]	ถนนสุขุมวิท		[illegible]	กรุงเทพม
7	AA0007	[illegible]	1	Male	12	3	[illegible]	Nakhonsithammarat		[illegible]	หมู่ที่1				[illegible]	[illegible]	[illegible]
8	AA0008	AA0008	2	FeMale	9	4	1963	Nakhonsithammarat		280/2	หมู่1				[illegible]	[illegible]	[illegible]
9	AA0009	AA0009	2	FeMale	3	2	1960	Bangkok	บริษัท การ	75			พระยาสุเร	พระยาสุเร	บางชัน	คลองสาม	กรุงเทพม
10	AA0010	AA0010	2	FEMALE	1	1	1964	Udon Thani		[illegible]				[illegible]	[illegible]	[illegible]	[illegible]
11	AA0011	AA0011	2	Male	1	10	1920	Bangkok		47/ 2				[illegible]	[illegible]	[illegible]	[illegible]
12	AA0012	AA0012	1	Male	16	10	1954	Udon Thani		304-306				[illegible]	[illegible]	[illegible]	[illegible]
13	AA0013	AA0013	2	MALE	15	11	1964		หมู่บ้าน[illegible]	[illegible]	5			บางกรวย			
14	AA0014	AA0014	2	Male	4	12	1972	Nakhonsithammarat	[illegible]	222					[illegible]	[illegible]	[illegible]
15	AA0015	AA0015	2	FeMale	18	10	1971	Bangkok	[illegible]	35			[illegible]	[illegible]	คลองตัน	วัฒนา	กรุงเทพม
16	AA0016	AA0016	2	MALE	12	11	1971		หมู่บ้าน[illegible]	[illegible]	10		[illegible]	สุขาภิบาล			
17	AA0017	AA0017	1	MALE	25	11	1963	------------		3-Apr	3			[illegible]	------------	------------	------------
18	AA0018	AA0018	2	Male	10	4	1965	Bangkok	[illegible]	43				[illegible]	[illegible]	[illegible]	[illegible]
19	AA0019	AA0019	2	MALE	8	12	1970			19				[illegible]			
20	AA0020	AA0020	2	MALE	4	1	[illegible]	Bangkok		114			[illegible]	ประชาชื่น	[illegible]	[illegible]	กรุงเทพม
21	AA0021	AA0021	2	MALE	9	5	1958	------------		4			[illegible]	[illegible]	------------	------------	[illegible]
22	AA0022	AA0022	2	MALE	17	2	1958	Kalasin		99	9			[illegible]	สำราญ	สามชัย	กาฬสินธุ์

รูปที่ 3.1 ตัวอย่างข้อมูลสมาชิกกลุ่มนกแฟนคลับ

MemberID	MemberType	Sex	Age	StateProvince	Seat	TicketPromotion	Week	AvgAdvanceBookingDate	LastFlight
AA0001	2	0	52	Bangkok	Window	0	1	13	4/20/2012
AA0005	1	0	60	Bangkok	Aisle	0	0	102	12/24/2013
AA0006	2	0	56	Bangkok	Aisle	0	0	2	12/24/2013
AA0008	2	1	51	Nakhon Si Thammarat	Window	0	1	0	5/18/2012
AA0010	1	1	50	Udon Thani	Aisle	1	1	84	3/16/2013
AA0011	1	0	94	Bangkok	Window	0	0	2	11/22/2012
AA0012	2	0	50	Udon Thani	Window	0	0	14	12/21/2013
AA0013	2	0	50	Nonthaburi	Window	0	0	7	12/24/2013
AA0014	2	0	42	Nakhon Si Thammarat	Aisle	1	0	23	3/29/2012
AA0015	2	1	43	Bangkok	Aisle	0	1	1	2/1/2013
AA0016	2	0	43	Bangkok	Window	0	0	48	10/22/2012
AA0017	2	0	51	Nakhon Si Thammarat	Aisle	0	1	38	12/15/2013
AA0018	2	0	49	Bangkok	Aisle	0	0	6	4/21/2013
AA0019	2	0	44	Bangkok	Window	0	0	32	12/22/2013

รูปที่ 3.2 ตัวอย่างข้อมูลการใช้บริการของกลุ่มสมาชิกนกแฟนคลับ

3.2 ความเข้าใจเกี่ยวกับข้อมูล (Data Understanding)

การเก็บรวบรวมข้อมูลขั้นต้น (Collect Initial Data)

เริ่มต้นจากการดึงข้อมูลจากฐานข้อมูล โดยข้อมูลที่ใช้ในการสร้างแบบจำลองจะมาจากประวัติการใช้บริการของของลูกค้าในแต่ละครั้งที่ใช้บริการ (Transaction Data) ของกลุ่มสมาชิกนกแฟนคลับที่ประกอบด้วยข้อมูลลูกค้าและข้อมูลการใช้บริการ จากนั้นสั่งให้โปรแกรมส่งข้อมูลออกมาในรูปแบบข้อความ (Text File) จากฐานข้อมูลเพื่อนำข้อมูลที่เป็นข้อมูลดิบเหล่านี้มาเตรียมพร้อมสำหรับการทำเหมืองข้อมูล

การอธิบายข้อมูล (Describe Data) หลังจากที่ได้ข้อมูลดิบที่นำออกมาจากฐานข้อมูลแล้วพบว่า ข้อมูลทั้งหมดประกอบด้วยข้อมูลการใช้บริการของลูกค้าสมาชิกลุ่มนกแฟนคลับทั้งปี 2012-2013

การนำข้อมูลที่ถูกเก็บไว้เป็นจำนวนมากมาใช้งานมักมีข้อมูลที่ผิดพลาด เช่น ข้อมูลไม่สมบูรณ์ ข้อมูลผิดปกติจากความเป็นจริง สาเหตุอาจเกิดจากความผิดพลาดของผู้ใช้งาน ความบกพร่องของระบบ หรือจากเครื่องคอมพิวเตอร์ขัดข้องก็ตาม ซึ่งล้วนแล้วแต่ส่งผลต่อความถูกต้องและความน่าเชื่อถือของแบบจำลอง เพื่อให้แบบจำลองมีความน่าเชื่อถือมากที่สุดข้อมูลที่ไม่มีความจำเป็นจะต้องถูกตัดทิ้ง ข้อมูลที่มีความผิดพลาดจะต้องถูกแก้ไข ในขั้นตอนการเตรียมข้อมูลนี้จึงมีความสำคัญมาก

การทำความสะอาดข้อมูล (Cleansing Data) เมื่อทำการคัดเลือกข้อมูลเหมาะสมในการทำเหมืองข้อมูลแล้ว พบว่าจากข้อมูลที่ได้มา มีส่วนของข้อมูลที่ไม่สามารถนำมาใช้ทำการจำแนก (Classification) ได้ เช่น ข้อมูลที่ขาดหาย (Missing Value) ข้อมูลที่มีค่าผิดปกติ ข้อมูลที่ขาดหายไปได้มีการแทนค่าให้สามารถนำไปใช้ในการทดสอบได้ และข้อมูลบางส่วนไม่สามารถนำมาใช้ได้ตรงๆ

ยกตัวอย่างเช่นข้อมูลวันเกิดของลูกค้าจะถูกแปลเป็นอายุเพื่อนนำมาหาลักษณะเด่นของกลุ่มลูกค้าแต่ละกลุ่ม แต่เมื่อแปลงแล้วกลับพบว่าค่าที่ได้น้อยกว่าความเป็นจริง ซึ่งสาเหตุอาจเกิดจากความผิดพลาดในการกรอกข้อมูล ดังนั้นจะตองทำการคัดกรองข้อมูลเหล่านี้ทิ้งไป และยังมีข้อมูลในส่วนของตัวบุคคลที่ระบุเป็นแบบเจาะจง ผู้ศึกษาจึงไม่ได้เลือกนำมาใช้ โดยขั้นตอนนี้ได้ทำควบคู่กับขั้นตอนที่ 1 ในบางส่วนและยังคงทำต่อหลังจากที่ขั้นตอนที่ 1 เสร็จสิ้นไปแล้ว

จากรูปที่ 3.1 และ 3.2 จะเห็นได้ว่าคอลัมน์ Age มีค่าเป็น 0 และค่าว่าง คอลัมน์ State Province จะมีการกรอกข้อมูลที่ผิดโดยการกรอกยกตัวอย่างเช่น “-------“ หรือ ค่าว่าง ทำให้เกิดข้อมูลที่ผิด จากสองคอลัมน์นั้นทางผู้ศึกษาได้ทำการ Cleansing คอลัมน์ Age โดยการให้โปรแกรม Rapid Mier ทำการแทนค่าโดยใช้ค่าเฉลี่ยของกลุ่มข้อมูลทั้งหมดแทนค่าลงไปของข้อมูลที่มีค่าเป็น 0 และที่เป็นค่าว่าง ส่วนคอลัมน์ State Province จะทำการ Cleansing โดยการตัด Row ที่มีการกรอกข้อมูลเข้ามาผิดออกไป ไม่นำ Row นั้นเข้ามาใช้ในการค้นหาข้อมูล

3.3 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูล การเลือกข้อมูลสำหรับกลุ่มสมาชิกนกแฟนคลับที่ใช้บริการสายการบินในช่วงระยะเวลาระหว่างปี 2012-2013 โดยกลุ่มสมาชิกนั้นเป็นกลุ่มสมาชิกที่มีจำนวนไม่มาก หากแต่ปริมาณธุรกิจค่อนข้างสูง ซึ่งในที่นี้การเลือกข้อมูลต้องการจะทำการเลือกข้อมูลจากข้อมูลสมาชิก ข้อมูลการใช้บริการ ข้อมูลรายการต่างๆ ที่กลุ่มสมาชิกได้เข้ามากระทำในระบบ จากนั้นนำข้อมูลมาแยกส่วนที่ซ้ำซ้อนออกไปพร้อมจัดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมโดยการแปลงให้อยู่ในรูปแบบไฟล์ .csv (Comma Separated Values) Format โดยจะได้ข้อมูลออกมาในรูปแบบไฟล์ที่มีเครื่องหมาย Comma คั่นระหว่างแอตทริบิวต์

```
MemberID,MemberType,Sex,Age,StateProvince,Seat,TicketPromotion,Week,AvgAdvanc
eBookingDate,LastFlight,Recency,Frequency,Monetary,RFM
AA0001,2,1,52,Bangkok,0,0,1,13,4/20/2012,621,3,6467,111
AA0005,1,1,60,Bangkok,1,0,0,102,12/24/2013,8,104,137125,521
AA0006,2,1,56,Bangkok,1,0,0,2,12/24/2013,8,5,8126,511
AA0008,2,2,51,Nakhon Si Thammarat,0,0,1,0,5/18/2012,593,4,13234,111
AA0010,1,2,50,Udon Thani,1,1,1,84,3/16/2013,291,39,36083,211
AA0011,1,1,94,Bangkok,0,0,0,2,11/22/2012,405,2,4187,111
AA0012,2,1,50,Udon Thani,0,0,0,14,12/21/2013,11,28,34439,511
```

AA0013,2,1,50,Nonthaburi,0,0,0,7,12/24/2013,8,9,16619,511

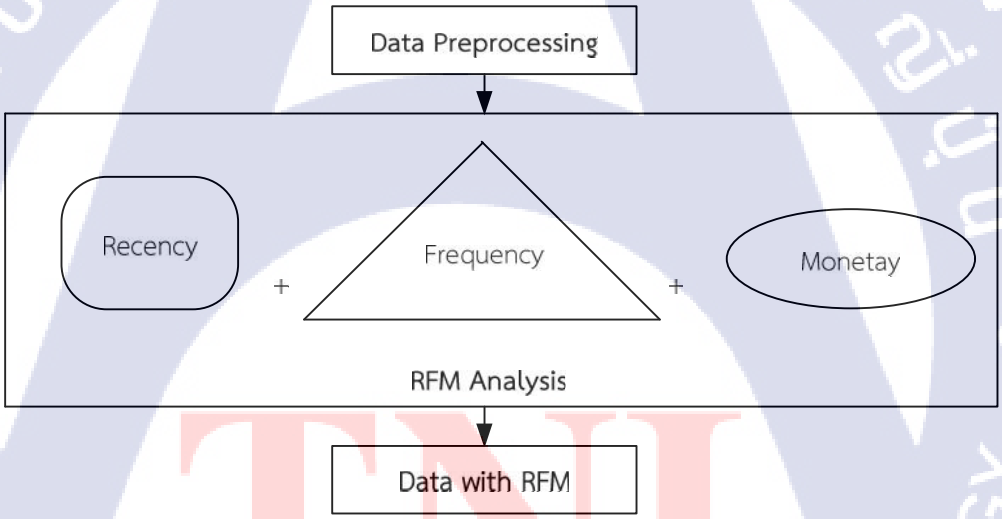
AA0014,2,1,42,Nakhon Si Thammarat,1,1,0,23,3/29/2012,643,3,3729,111

3.4 การขุดค้นข้อมูล (Data Mining)

ในขั้นตอนการขุดค้นข้อมูลนี้ ผู้ศึกษาเลือกเทคนิคการจัดกลุ่มข้อมูล (Clustering) เข้ามาใช้ในการทดลองนำข้อมูลแต่ละกลุ่มที่ได้จากการขุดค้นข้อมูลมาทำการหาความสัมพันธ์ในแต่ละฟังก์ชั่นการใช้งาน โดยใช้เทคนิค Association Rule นำข้อมูลที่ได้ไปวิเคราะห์หาแนวโน้มความต้องการของลูกค้า เพื่อนำมาใช้ในการวิเคราะห์ถึงฟังก์ชันที่ควรนำมาพัฒนาในระบบเพิ่มเติม ช่วยสร้างความพึงพอใจในการใช้บริการให้แก่ลูกค้าต่อไป

ทำการวิเคราะห์และสรุปผลทั้งในด้านพฤติกรรมการใช้งาน และความสัมพันธ์ในการใช้บริการ

การคำนวณ RFM โดยใช้เทคนิค RFM Analysis



รูปที่ 3.3 การคำนวณ RFM โดยใช้เทคนิค RFM Analysis

RFM Analysis การวิเคราะห์ทางการตลาดโดยใช้ RFM Analysis เป็นการนำข้อมูลมาคำนวณเพิ่มในส่วนค่า RFM ซึ่งเป็นค่าที่จะสามารถวัดความสำคัญของลูกค้าแต่ละรายโดยวิเคราะห์จาก Recency, Frequency และ Monetary ในการทำรายการเข้ามาในส่วนประกอบหนึ่งในข้อมูลส่วนตัวของลูกค้ารายนั้นๆ

จากขั้นตอนการเตรียมข้อมูล นำข้อมูล Recency, Frequency และ Monetary ของลูกค้าแต่ละรายมาทำการคำนวณค่า RFM เพื่อเป็นข้อมูลหนึ่งในการแบ่งกลุ่มข้อมูล โดยใช้สูตรการคำนวณ RFM Score = R+F+M

เนื่องจากข้อมูลในงานวิจัยนี้เป็นข้อมูลการใช้บริการในปี 2012-2013 ผู้ศึกษาจึงได้ให้คะแนน RFM ดังตารางต่อไปนี้

Recency

ตารางที่ 3.1 การให้คะแนน Recency

ระยะเวลาการใช้บริการล่าสุด	คะแนน
10/2013 - 12/2013	5
07/2013 - 09/2013	4
04/2013 - 06/2013	3
01/2013 - 03/2013	2
01/2012 - 12/2012	1

Frequency

ตารางที่ 3.2 การให้คะแนน Frequency

จำนวนการใช้บริการ	คะแนน
มากกว่า 250 ครั้ง	5
มากกว่า 200 ครั้ง	4
มากกว่า 150 ครั้ง	3
มากกว่า 100 ครั้ง	2
ต่ำกว่า 100 ครั้ง	1

Monetary

ตารางที่ 3.3 การให้คะแนน Monetary

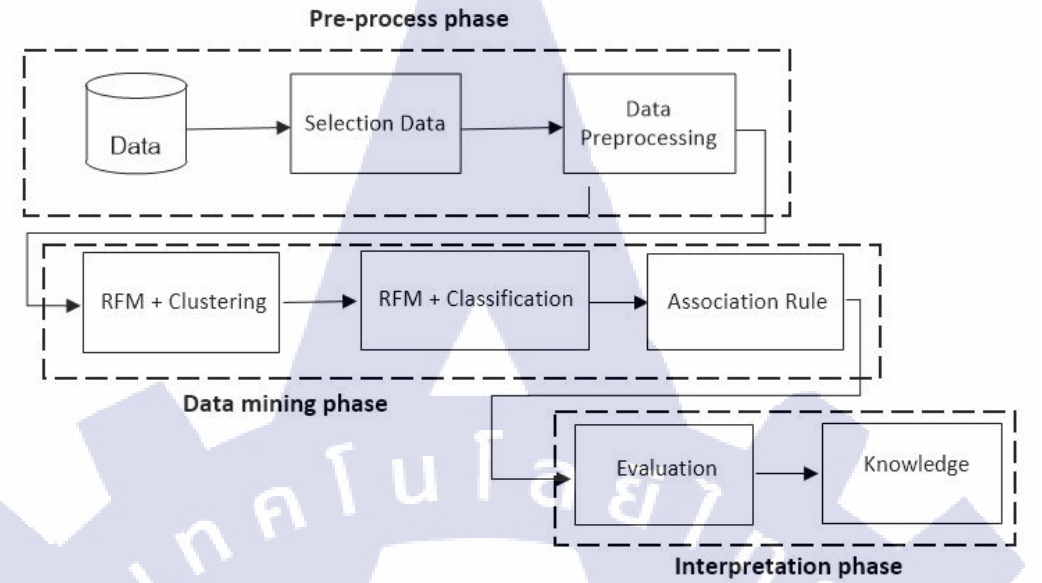
มูลค่าเงิน	คะแนน
มากกว่า 800,000 บาท	5
มากกว่า 600,000 บาท	4
มากกว่า 400,000 บาท	3
มากกว่า 200,000 บาท	2
น้อยกว่า 100,000 บาท	1

MemberID	MemberType	Sex	Age	StateProvince	Seat	TicketPromotion	Week	AvgAdvanceBookingDate	LastFlight	Recency	Frequency	Monetary	RFM
AA0001	2	0	52	Bangkok	Window	0	1	13	4/20/2012	671	3	6467	111
AA0005	1	0	60	Bangkok	Aisle	0	0	102	12/24/2013	8	104	137125	521
AA0006	2	0	56	Bangkok	Aisle	0	0	2	12/24/2013	8	5	8125	511
AA0008	2	1	51	Nakhon Si Thammarat	Window	0	1	0	5/18/2012	503	4	13234	111
AA0010	1	1	50	Udor Thani	Aisle	1	1	84	3/16/2013	291	39	36083	211
AA0011	1	0	54	Bangkok	Window	0	0	2	11/22/2012	405	2	4187	111
AA0012	2	0	50	Udor Thani	Window	0	0	14	12/21/2013	11	28	34459	511
AA0013	2	0	50	Nonthaburi	Window	0	0	7	12/24/2013	8	9	16619	511
AA0014	2	0	42	Nakhon Si Thammarat	Aisle	1	0	23	3/29/2012	643	3	3729	111
AA0015	2	1	43	Bangkok	Aisle	0	1	1	2/1/2013	334	1	1860	211
AA0016	2	0	43	Bangkok	Window	0	0	43	10/22/2012	436	2	3625	111
AA0017	2	0	51	Nakhon Si Thammarat	Aisle	0	1	38	12/15/2013	17	29	52853	511
AA0018	2	0	49	Bangkok	Aisle	0	0	5	4/21/2013	255	8	16581	311
AA0019	2	0	44	Bangkok	Window	0	0	32	12/22/2013	10	58	93801	511

รูปที่ 3.4 การให้คะแนน RFM

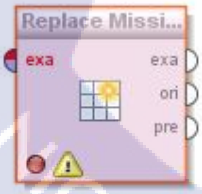
3.5 การพัฒนาแบบจำลอง (Modeling)

เลือกเทคนิคการสร้างแบบจำลองผู้ศึกษาได้เลือกใช้กฎการแบ่งกลุ่มแบบ Clustering เพื่อทำการค้นหาผลการจัดกลุ่มในที่นี้ได้นำ K-Means Algorithm มาใช้ในการแบ่งกลุ่มและใช้เทคนิค Classification ในการจำแนกประเภทเพื่อให้เห็นภาพว่าสมาชิกแบบไหน ประเภทไหนตกอยู่ในกลุ่มไหน เพื่อนำมาให้ซึ่งข้อมูลที่จะนำไปใช้ในการวิเคราะห์หาความสัมพันธ์การใช้บริการโดย Association Rule ต่อไป



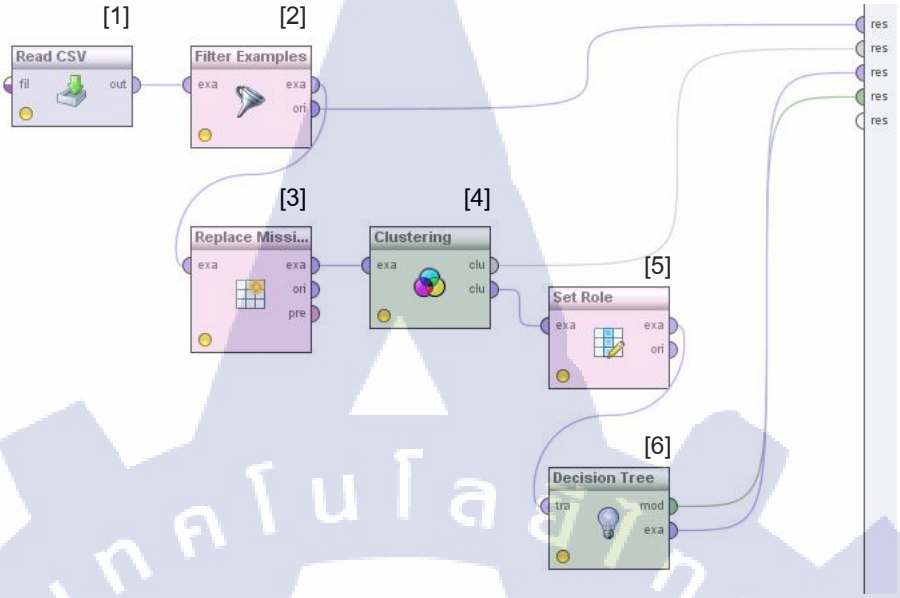
รูปที่ 3.5 Mining Process

ตารางที่ 3.4 โอเปอเรเตอร์ต่างๆ ในโปรแกรม RapidMiner Studio ที่ใช้ในการทำการศึกษา

โอเปอเรเตอร์		คำอธิบาย
	Read CSV	ใช้สำหรับดึงข้อมูลออกมาจาก File csv
	Filter Examples	ใช้สำหรับกรองข้อมูลตามเงื่อนไขที่กำหนด
	Replace Missing Value	ใช้สำหรับแทนที่ข้อมูลที่ต้องการเปลี่ยนค่า

ตารางที่ 3.4 โอเปอเรเตอร์ต่างๆ ในโปรแกรม RapidMiner Studio ที่ใช้ในการทำการศึกษา (ต่อ)

โอเปอเรเตอร์		คำอธิบาย
	Discretize	ใช้แปลงข้อมูลแบบ Discretization
	Numerical to Binominal	ใช้สำหรับแปลงข้อมูลให้เป็น Binominal
	K-Means	ใช้สำหรับการแบ่งกลุ่มด้วยเทคนิค K-Means อัลกอริทึม
	Decision Tree	ใช้สำหรับสร้างโมเดล Decision Tree
	FP-Growth	ใช้สำหรับค้นหารูปแบบการใช้บริการที่เกิดขึ้นบ่อยๆ
	Create Association Rules	ใช้สำหรับสร้างกฎความสัมพันธ์จากข้อมูลรูปแบบที่หาได้



รูปที่ 3.6 การสร้าง K-Means Process โดย RapidMiner Studio

จากรูปที่ 3.6 อธิบายการทำงานได้ดังนี้

1. ทำการ Read CSV จาก Data ที่ได้หลังจากการ Cleansing เป็น File CSV จะได้ข้อมูลส่งต่อไปยัง Operator ต่อไป (out)

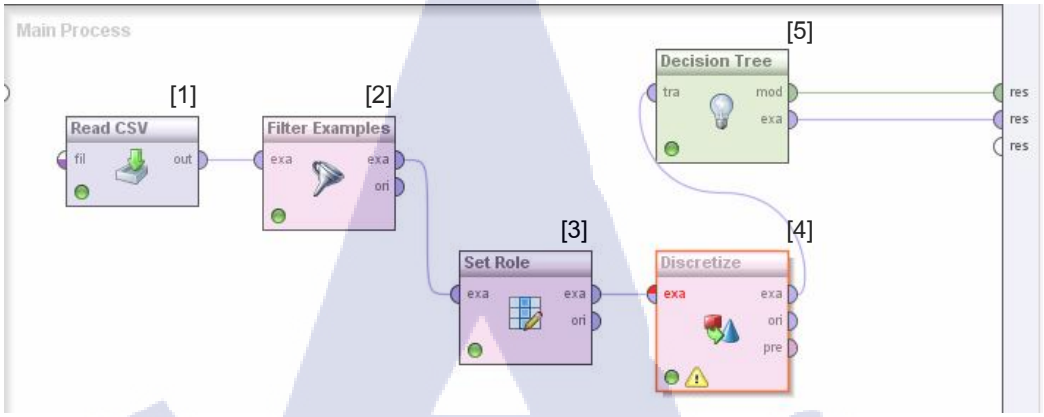
2. ทำการ Filter Examples ข้อมูลช่วงอายุระหว่าง 18-80 จะได้ Example Data (exa) ส่งต่อไปยัง Operator Replace Missing

3. Replace Missing Value เพื่อทำการแทนค่าของอายุของสมาชิกที่ยังว่างอยู่โดยการให้ค่า average ของข้อมูลทั้งหมดทำการส่งค่า Example (exa) ไปยัง Operator Clustering และทำการส่งค่า Original (ori) ออกไปยัง Results

4. เข้าสู่การทำ Clustering Process โดยใช้ K-Means Algorithm กำหนดค่า K = 4, Max Runs = 10 จะได้ Cluster ที่ต้องการทั้งหมด 4 กลุ่ม ส่งค่า Cluster (clu) ไปให้ Results (res) และส่ง Cluster (clu) อีกส่วนหนึ่งไปยัง Operator Set Role

5. Operator Set Role ทำการ Set Attribute Cluster ที่ได้ให้เป็น Label เพื่อที่จะส่งค่า Example (exa) เข้าสู่ Operator Decision Tree

6. Operator Decision Tree เพื่อหาค่าการจำแนกของกลุ่ม Cluster อีกครั้งเพื่อดูผลการแบ่งกลุ่มที่ได้จากการทำ Clustering และทำการส่ง Model (mod) ไป Results และ Example (exa) ของผลลัพธ์ที่ได้ไป Results (res)



รูปที่ 3.7 การสร้าง Classification Process โดย RapidMiner Studio

จากรูปที่ 3.7 อธิบายการทำงานได้ดังนี้

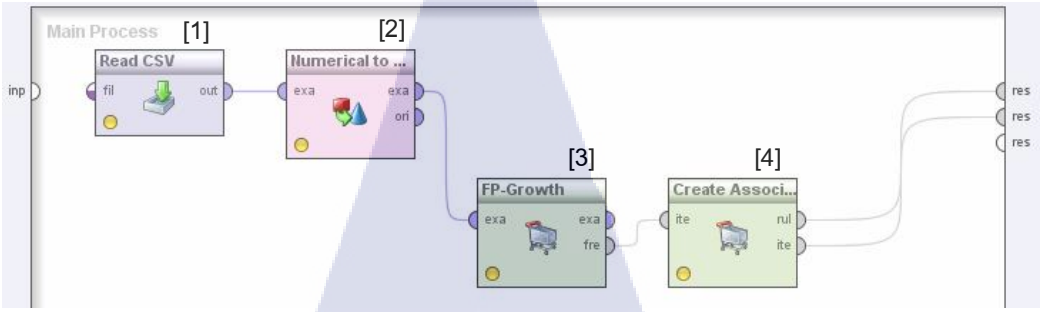
1. ทำการ Read CSV จาก Data ที่ได้หลังจากการ Cleansing เป็น File CSV จะได้ข้อมูลส่งต่อไปยัง Operator ต่อไป (out)

2. ทำการ Filter Examples ข้อมูลช่วงอายุระหว่าง 18-80 จะได้ Example Data (exa) ส่งต่อไปยัง Operator Set Role

3. ทำการ Set Role โดยกำหนดให้ Role RFM เป็น Lable ข้อมูลและส่ง Example (exa) ไปยัง Operater Discretize

4. Operator Discretize ทำการแปลงข้อมูลและส่งค่า Example (exa) ไปยัง Operator Decision Tree

5. Operator Decision Tree เพื่อทำการจำแนกข้อมูลต่างๆ ออกมาและสร้างในรูปแบบของโมเดล Decision Tree และทำการส่ง Model (mod) ไป Results และ Example (exa) ของผลลัพธ์ที่ได้ไป Results (res)



รูปที่ 3.8 การสร้าง Association Process โดย RapidMiner Studio

จากรูปที่ 3.8 อธิบายการทำงานได้ดังนี้

1. ทำการ Read CSV จาก Data ที่ได้หลังจากการ Cleansing เป็น File CSV จะได้ข้อมูลส่งต่อไปยัง Operator ต่อไป (out)

2. ทำการแปลงข้อมูลให้เป็น Binominal และส่งผลลัพธ์ Example (exa) ที่ได้ไปยัง Operator FP-Growth

3. Operator FP-Growth ทำการค้นหารูปแบบที่เกิดขึ้นบ่อยๆ และส่งค่า Frequent (fre) ที่ได้จากการค้นหาส่งไป Operator Create Association Rule

4. ทำการ Create Association rule จากรูปแบบที่หาได้จาก FP-Growth ทำการส่ง Rules (rul) และ Items Set (itm) ที่หาได้ไปยังผลลัพธ์ของการทำงาน Results (res)

3.6 การประเมินแบบจำลอง (Asses Model)

จากแบบจำลองที่ได้จากโปรแกรม RapidMiner Studio ในการทำเหมืองข้อมูลแบบจำลองการแบ่งกลุ่ม แบบจำลองแบบจำแนก และแบบจำลองการหาความสัมพันธ์ของบริการกับพฤติกรรมลูกค้าที่ใช้บริการ สามารถตรวจสอบความแม่นยำ และประเมินหาความน่าเชื่อถือได้ ด้วยการนำผลลัพธ์ของข้อมูลเรียนรู้มาเปรียบเทียบค่าความถูกต้องกับข้อมูลทดสอบ ซึ่งถือว่าเป็นการประเมินทางด้านเทคนิคเกี่ยวกับปัจจัยในด้านต่างๆ ที่มีผลในการสร้างแบบจำลอง และนอกจากนี้จะต้องประเมินถึงความสามารถในการตอบสนองเป้าหมายทางธุรกิจ โดยเปรียบเทียบผลลัพธ์กับวัตถุประสงค์ทางธุรกิจที่ได้ตั้งเอาไว้ตอนเริ่มต้นการศึกษา ได้แก่

1. แบบจำลองการแบ่งกลุ่มลูกค้าย่อย สามารถนำไปประยุกต์ใช้ในการออกแผนการตลาดหรือนำเสนอบริการโปรโมชั่นได้ตรงกับความต้องการของกลุ่มลูกค้า สร้างความพึงพอใจ และช่วยทำให้องค์กรสามารถ สร้างรายได้เพิ่มมากขึ้น

2. แบบจำลองกฎความสัมพันธ์ของบริการแต่ละประเภทกับลูกค้าที่มาใช้บริการ สามารถนำผลลัพธ์ที่ได้ทั้งหมด ไปเป็นข้อมูลอีกส่วนหนึ่งในการกำหนดแผนการตลาด หรือนำไปประกอบการพิจารณาตัดสินใจในการขาย

3. เมื่อนำแบบจำลองทั้งสองจะช่วยให้องค์กรสามารถวิเคราะห์ข้อมูลได้อย่างรวดเร็ว สามารถดึงเอาความรู้ที่ถูกซ่อนอยู่ในข้อมูลออกมา

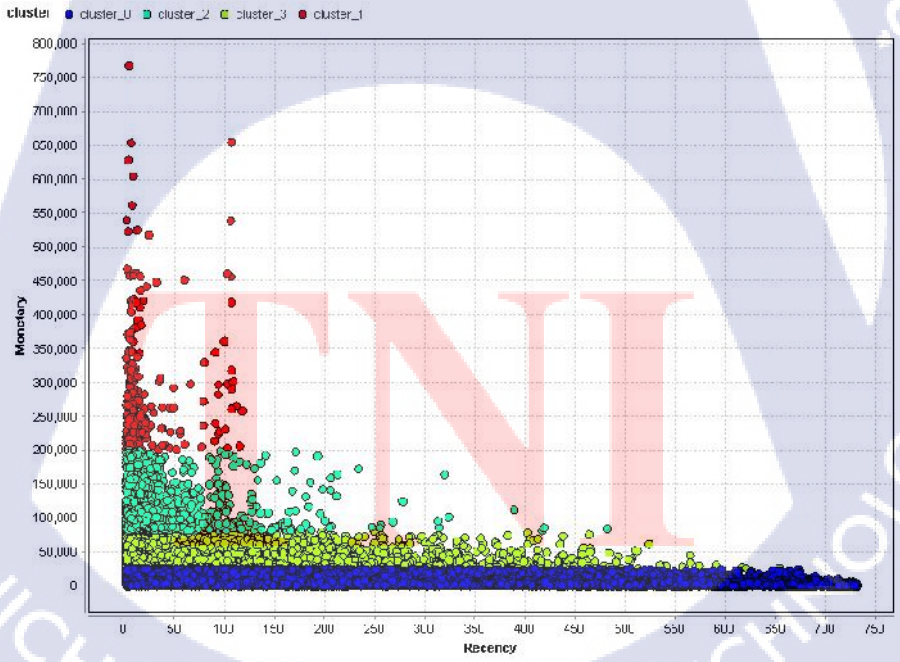
3.7 การนำแบบจำลองไปใช้ (Deployment)

การนำผลการศึกษาไปประยุกต์ใช้ จะต้องนำแบบจำลองที่ได้จากการทำเหมืองข้อมูลในรูปแบบการแบ่งกลุ่มของลูกค้า การจำแนกกลุ่มของลูกค้า และการทำแบบจำลองความสัมพันธ์ของการใช้บริการของลูกค้า โดยหลังจากการวิเคราะห์แล้วจะทำให้ทราบถึงพฤติกรรมการใช้บริการของลูกค้าแต่ละกลุ่ม เพื่อให้นำไปใช้งานได้อย่างสะดวกที่สุด ฝ่ายขายสามารถนำผลลัพธ์ไปประยุกต์ใช้กับระบบที่มีอยู่แล้ว การนำไปใช้ที่ชัดเจนที่สุดคือ การออกโปรโมชั่นให้กับลูกค้าโดยอาศัยความสัมพันธ์จากแบบจำลองที่ได้ เช่น ถ้าหากลูกค้ามาจากกลุ่มย่อยที่ 1 ฝ่ายขายจะต้องส่งข้อมูลบริการ รวมถึงข้อมูลโปรโมชั่นที่ลูกค้ากลุ่มนั้นให้ความสนใจและมีโอกาสที่จะใช้บริการอื่นๆ ที่ต่อเนื่องกันให้กับลูกค้าในกลุ่มนั้น ซึ่งจะเป็นการเพิ่มโอกาสในการให้บริการประเภทอื่นๆ

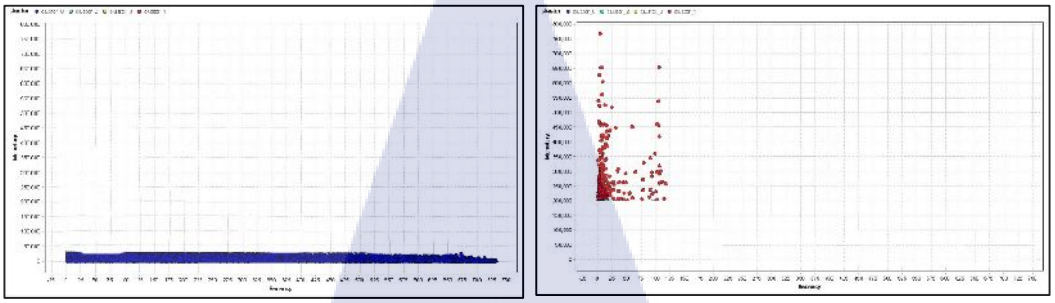
บทที่ 4
ผลการวิเคราะห์ข้อมูล

4.1 การวิเคราะห์ข้อมูล

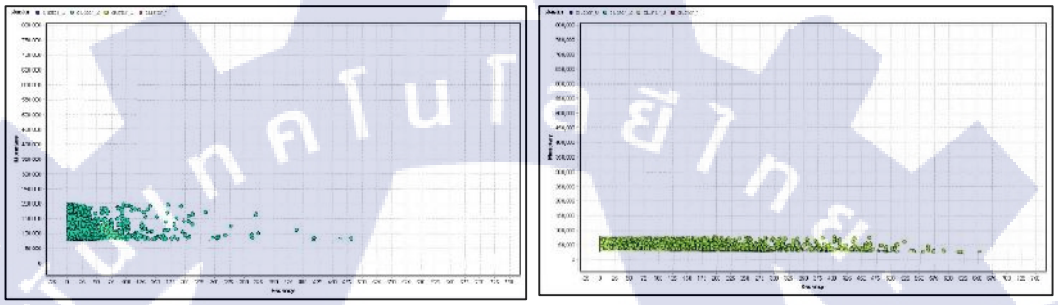
จากการทดลองกับข้อมูลกลุ่มสมาชิกนกแฟนคลับที่ใช้บริการในปี 2012-2013 จำนวน 149,831 ราย หลังจากการ Cleansing ข้อมูลแล้วจำนวนกลุ่มที่สมาชิกที่นำมาทดลองทั้งหมด 62,496 ราย จากจำนวนสมาชิกอายุระหว่าง 18-80 ปี ผลที่ได้จากการทำการทดลองแบบการจัดกลุ่มโดยใช้ K-Means อัลกอริทึม ออกมาเป็น 4 กลุ่ม โดยผู้ศึกษาขอทำการแทนค่ากลุ่มดังนี้ Cluster 0 = Group1, Cluster 1 = Group2, Cluster 2 = Group 3, Cluster 3 = Group 4 โดยการแบ่งกลุ่มในปัจจัยทางด้านราคาของตั๋วเครื่องบินในการใช้บริการนั้นมีอัตราที่ใกล้เคียงกันจึงทำให้ผลของการจัดกลุ่มออกมาในรูปแบบการเรียงลำดับของข้อมูล ทำให้เห็นว่า Group 2 ที่มีค่า Monetary ในกลุ่มสูงที่สุดจะอยู่บนสุดซึ่งมูลค่าการใช้เงินสูงสุดในกลุ่มนั้นมีมูลค่าถึง 750,000 บาท และ Group 3 เป็นกลุ่มสองที่มีค่า Monetary ออกมาเป็นลำดับสอง โดยจะเห็นได้ว่ากลุ่ม Group 1 เป็นกลุ่มล่างสุดและมีจำนวนสมาชิกมากที่สุด ซึ่งเป็นกลุ่มที่มีค่า Recency, Frequency, Monetary ต่ำสุดนั่นเอง



รูปที่ 4.1 ผลจากการจัดกลุ่มด้วย K-Means อัลกอริทึมโดยแกน x คือ Recency และแกน y คือ Monetary

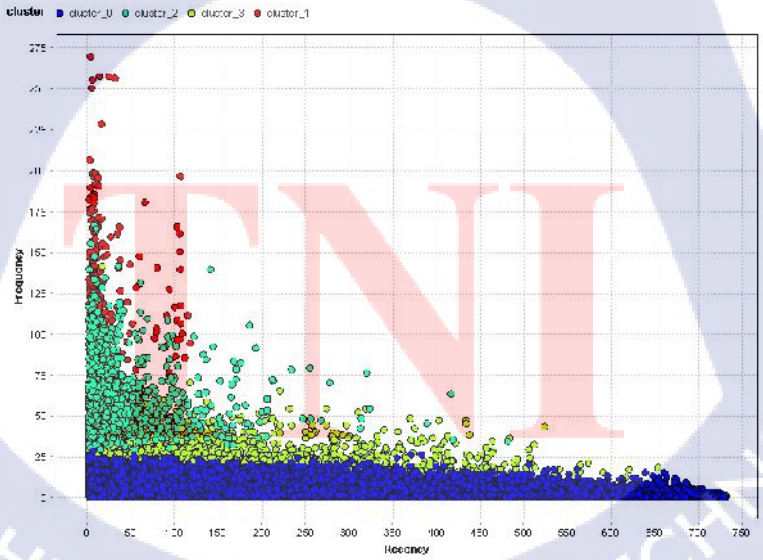


Group 1 Group 2



Group 3 Group 4

รูปที่ 4.2 ผลจากการจัดกลุ่มด้วย K-Means อัลกอริทึม Cluster 0-3 (Group 1-4)



รูปที่ 4.3 ผลจากการจัดกลุ่มด้วย K-Means อัลกอริทึมโดยแกน x คือ Recency และแกน y คือ ค่า Frequency

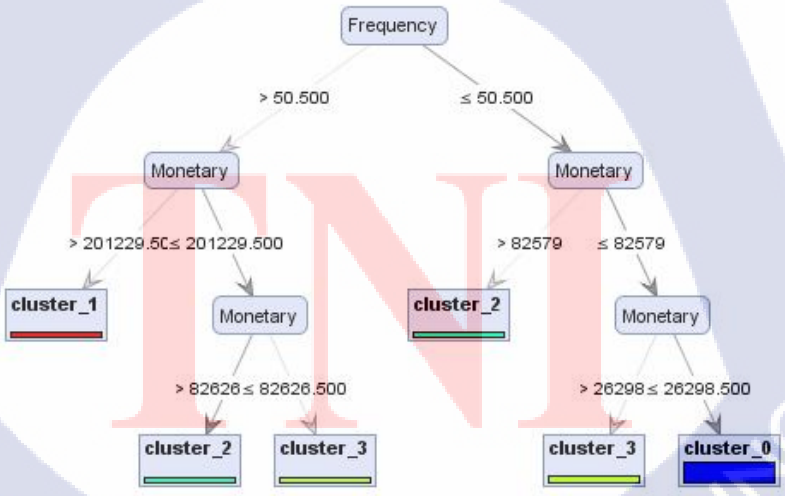
Cluster Model

```
Cluster 0: 50935 items
Cluster 1: 275 items
Cluster 2: 1930 items
Cluster 3: 9356 items
Total number of items: 62496
```

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
MemberType	1.989	1.185	1.583	1.916
Sex	1.372	1.127	1.177	1.274
Age	41.364	52.636	49.336	46.116
Seat	0.607	0.625	0.593	0.599
TicketPromo	0.263	0.022	0.047	0.115
Week	0.360	0.269	0.301	0.332
AvgAdvanceE	27.988	11.891	16.062	20.589
Recency	253.287	19.578	26.551	62.239
Frequency	4.906	126.625	63.663	25.924
Monetary	7741.387	282350.076	120446.248	44895.893

รูปที่ 4.4 ตารางข้อมูล Centroid Table ที่หาได้จากการทำ K-Means อัลกอริทึม

จากปัญหาของการจัดกลุ่มผู้ศึกษาจึงได้นำเทคนิค Decision Tree เข้ามาช่วยในการตอบปัญหาของการจัดกลุ่มว่าเหตุใดกลุ่มสมาชิกที่นำมาทดลองในการศึกษาถึงการตกไปอยู่ในกลุ่มต่างๆ



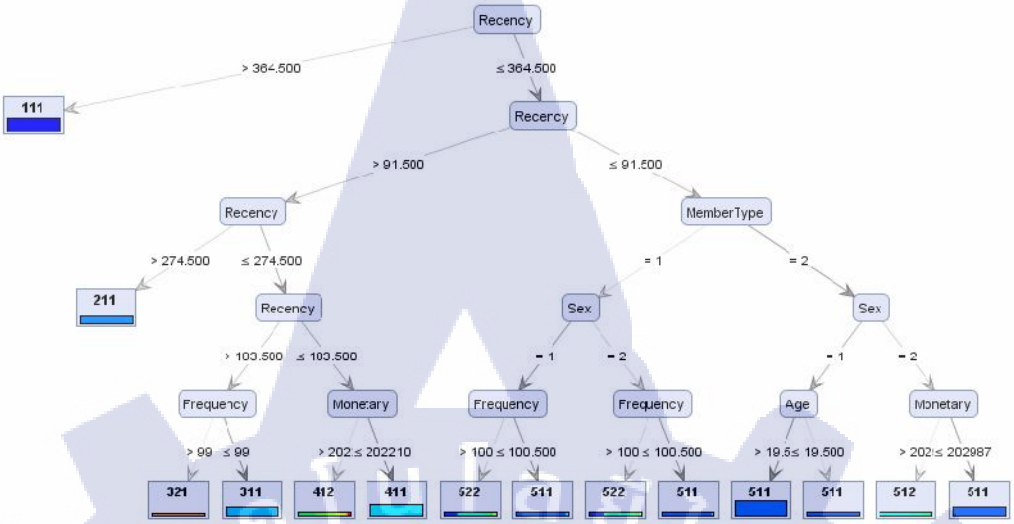
รูปที่ 4.5 ผลลัพธ์ที่ได้จากการใช้ Decision Tree หลังจากการทำ K-Means

Tree

```
Frequency > 50.500
|   Monetary > 201229.500: cluster_1 {cluster_0=0, cluster_2=0,
cluster_3=0, cluster_1=275}
|   Monetary ≤ 201229.500
|   |   Monetary > 82626.500: cluster_2 {cluster_0=0, cluster_2=1378,
cluster_3=0, cluster_1=0}
|   |   Monetary ≤ 82626.500: cluster_3 {cluster_0=0, cluster_2=0,
cluster_3=186, cluster_1=0}
Frequency ≤ 50.500
|   Monetary > 82579: cluster_2 {cluster_0=0, cluster_2=552,
cluster_3=0, cluster_1=0}
|   Monetary ≤ 82579
|   |   Monetary > 26298.500: cluster_3 {cluster_0=0, cluster_2=0,
cluster_3=9170, cluster_1=0}
|   |   Monetary ≤ 26298.500: cluster_0 {cluster_0=50935,
cluster_2=0, cluster_3=0, cluster_1=0}
```

ผลลัพธ์จากการทำ K-Means และ Decision Tree จะเห็นได้ว่ากลุ่มสมาชิกที่ตกอยู่ในกลุ่ม Cluster ที่ 1 และ 2 มีความน่าสนใจที่สุดเนื่องจากเป็นกลุ่มสมาชิกที่มี Recency และ Monetary ในการใช้บริการเยอะ โดยที่ Group 2 มี Centroid Recency ในการใช้บริการอยู่ที่ 19.57 Frequency ในการใช้บริการอยู่ที่ 126.62 และ Centroid ของ Monetary ในการใช้บริการอยู่ที่ 282,235 บาท และ Group 3 มี Centroid Recency ในการใช้บริการอยู่ที่ 26.55 Frequency ในการใช้บริการอยู่ที่ 63.66 และ Centroid ของ Monetary ในการใช้บริการอยู่ที่ 120446.24 บาท

หลังจากการทดลองแบ่งแยกการจัดกลุ่ม เพื่อให้เห็นถึงการจัดกลุ่มโดยการให้คะแนน RFM ผู้ศึกษาจึงได้นำข้อมูลสมาชิกที่ใช้ในการทดลองหลังจากการให้คะแนน RFM เรียบร้อยมาเข้าสู่การทำ Decision Tree อีกครั้งหนึ่งเพื่อให้เห็นถึงลักษณะการตกไปอยู่ในกลุ่มนั้นๆ ของสมาชิก



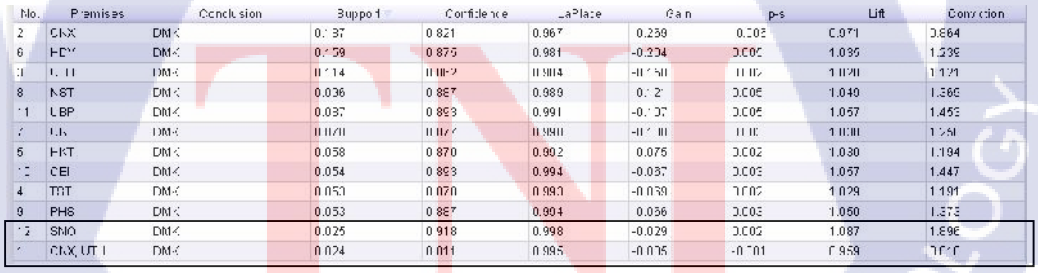
รูปที่ 4.6 ผลลัพธ์ที่ได้จากการใช้ Decision Tree

Tree

```
Recency > 364.500: 111 {111=14499, 521=0, 511=0, 211=0, 311=0, 411=0,
512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0, 552=0, 412=0, 534=0,
523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0, 423=0}
Recency ≤ 364.500
|   Recency > 91.500
|   |   Recency > 274.500: 211 {111=0, 521=0, 511=0, 211=5361, 311=0,
411=0, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0, 552=0, 412=0,
534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0, 423=0}
|   |   Recency ≤ 274.500
|   |   |   Recency > 183.500
|   |   |   |   Frequency > 99: 321 {111=0, 521=0, 511=0, 211=0,
311=0, 411=0, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0, 552=0,
412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=1, 421=0, 554=0, 423=0}
|   |   |   |   Frequency ≤ 99: 311 {111=0, 521=0, 511=0, 211=0,
311=7191, 411=0, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0,
552=0, 412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0,
423=0}
|   |   |   Recency ≤ 183.500
|   |   |   |   Monetary > 202210: 412 {111=0, 521=0, 511=0, 211=0,
311=0, 411=0, 512=0, 522=0, 532=0, 531=0, 422=7, 533=0, 553=0, 552=0,
412=8, 534=0, 523=0, 433=3, 434=1, 543=0, 321=0, 421=0, 554=0, 423=1}
|   |   |   |   Monetary ≤ 202210: 411 {111=0, 521=0, 511=0, 211=0,
311=0, 411=9759, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0,
552=0, 412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=1, 554=0,
423=0}
|   Recency ≤ 91.500
|   |   MemberType = 1
|   |   |   Sex = 1
|   |   |   |   Frequency > 100.500: 522 {111=0, 521=51, 511=0,
211=0, 311=0, 411=0, 512=0, 522=91, 532=28, 531=2, 422=0, 533=10,
553=3, 552=2, 412=0, 534=3, 523=5, 433=0, 434=0, 543=2, 321=0, 421=0,
554=0, 423=0}
```

```
|   |   |   |   Frequency ≤ 100.500: 511 {111=0, 521=0, 511=1203,
211=0, 311=0, 411=0, 512=46, 522=0, 532=0, 531=0, 422=0, 533=0,
553=0, 552=0, 412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0,
554=0, 423=0}
|   |   |   Sex = 2
|   |   |   |   Frequency > 100.500: 522 {111=0, 521=10, 511=0,
211=0, 311=0, 411=0, 512=0, 522=15, 532=3, 531=3, 422=0, 533=0,
553=0, 552=0, 412=0, 534=0, 523=1, 433=0, 434=0, 543=0, 321=0, 421=0,
554=0, 423=0}
|   |   |   |   Frequency ≤ 100.500: 511 {111=0, 521=0, 511=271,
211=0, 311=0, 411=0, 512=4, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0,
552=0, 412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0,
423=0}
|   |   MemberType = 2
|   |   |   Sex = 1
|   |   |   |   Age > 19.500: 511 {111=0, 521=23, 511=16129, 211=0,
311=0, 411=0, 512=8, 522=24, 532=3, 531=0, 422=0, 533=0, 553=0,
552=0, 412=0, 534=0, 523=1, 433=0, 434=0, 543=0, 321=0, 421=0, 554=1,
423=0}
|   |   |   |   Age ≤ 19.500: 511 {111=0, 521=1, 511=100, 211=0,
311=0, 411=0, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0, 552=0,
412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0, 423=0}
|   |   |   Sex = 2
|   |   |   |   Monetary > 202987: 512 {111=0, 521=0, 511=0, 211=0,
311=0, 411=0, 512=5, 522=2, 532=2, 531=0, 422=0, 533=0, 553=0, 552=0,
412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0, 423=0}
|   |   |   |   Monetary ≤ 202987: 511 {111=0, 521=4, 511=7608,
211=0, 311=0, 411=0, 512=0, 522=0, 532=0, 531=0, 422=0, 533=0, 553=0,
552=0, 412=0, 534=0, 523=0, 433=0, 434=0, 543=0, 321=0, 421=0, 554=0,
423=0}
```

จากผลลัพธ์การทำเหมืองข้อมูลทั้งสามแบบนั้นสามารถสรุปลักษณะทั่วไปและพฤติกรรมการใช้งานของสมาชิกได้ดังต่อไปนี้



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
2	CNX	DM=1	0.137	0.821	0.967	0.259	0.103	0.971	0.864
6	HDY	DM=1	0.159	0.875	0.981	-0.204	0.005	1.035	1.239
11	[illegible]	DM=1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	NST	DM=1	0.036	0.857	0.989	0.127	0.005	1.040	1.369
11	LBP	DM=1	0.037	0.893	0.991	-0.137	0.005	1.057	1.453
7	[illegible]	DM=1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	HKT	DM=1	0.058	0.870	0.992	0.075	0.002	1.030	1.194
12	CEI	DM=1	0.054	0.893	0.994	-0.037	0.005	1.057	1.447
4	TRT	DM=1	0.053	0.870	0.993	-0.039	0.002	1.029	1.191
9	PHS	DM=1	0.053	0.857	0.994	0.036	0.005	1.050	1.373
13	SNO	DM=1	0.025	0.918	0.998	-0.029	0.005	1.087	1.896
1	CNX, UTT	DM=1	0.024	0.911	0.995	-0.015	-0.001	0.959	0.000

รูปที่ 4.7 ผลลัพธ์จากการทำ Association Rules Cluster 0

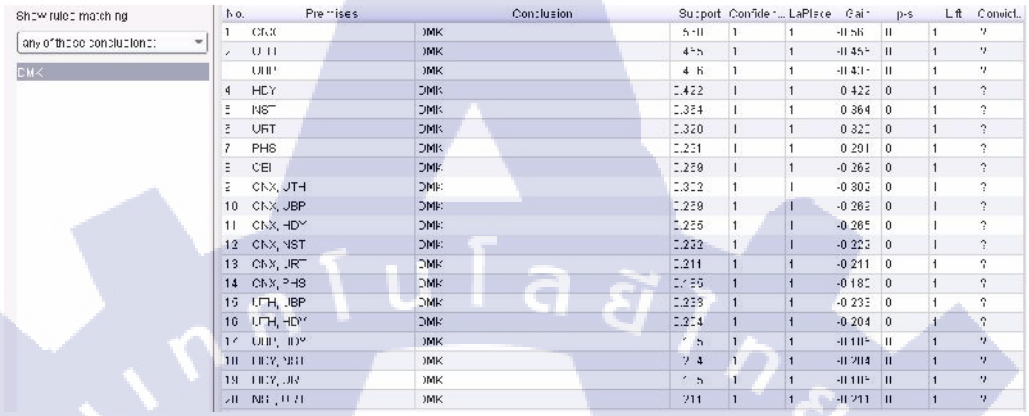
รูปที่ 4.8 Statistic ของกลุ่มที่ 1 Cluster 0

<u>กลุ่มที่ 1 Group 1 (Cluster 0) จำนวน 50,935 ราย</u>

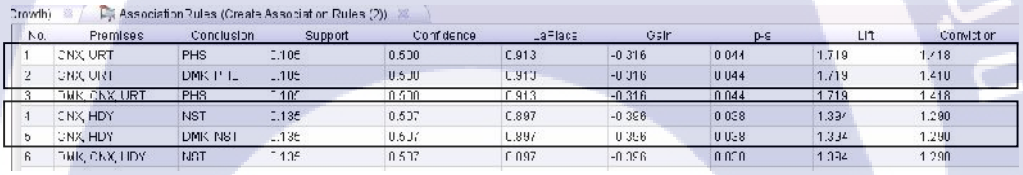
ลักษณะทั่วไป : จะเป็นกลุ่มสมาชิกที่อยู่ในกลุ่มสมาชิก Nok SmilePlus

พฤติกรรม : สมาชิกมีค่า RFM ในเกณฑ์ที่ต่ำ คือมี Recency Frequency และ Monetary ในการใช้บริการซึ่งถือเป็นว่ากลุ่ม Strangersถือเป็นลูกค้าที่มีกำไรต่ำและซื้อสินค้าหรือบริการจากเราในระยะสั้นจากการใช้บริการส่วนใหญ่จะบินในเส้นทางหลักๆ ดอนเมือง และเชียงใหม่จะชอบเลือกที่นั่งริมหน้าต่าง ไม่ได้ซื้อตั๋วในช่วงโปรโมชั่น โดยจากการทำ Association rules ดังรูปที่ 4.6 แสดงให้เห็นถึงตัวอย่างความสัมพันธ์การใช้บริการภายในกลุ่มที่ 1 จำนวน 2 กฎ โดยกำหนดค่าสนับสนุนไว้เพียง 0.1 และค่าความเชื่อมั่นต่ำสุดเท่ากับ 0.5 เนื่องจากกลุ่มที่ 1 นั้นมีการใช้บริการที่ต่างกันน้อยมาก หากกำหนดค่าสนับสนุนและค่าความเชื่อมั่นมากเกินไป จะทำให้หากฎความสัมพันธ์ภายในกลุ่มไม่ได้เลย โดยจากความสัมพันธ์ที่หาได้ จะเห็นได้ว่าสมาชิกในกลุ่มนี้มักจะเดินทางเชียงใหม่ ดอนเมือง ร้อยละ 82.1 จากค่าความเชื่อมั่น 0.821 และเดินทางหาดใหญ่ ดอนเมืองร้อยละ 87.5 จากค่าความ

เชื่อมั่น 0.875 จากทั้งสองความสัมพันธ์นั้นแสดงให้เห็นถึงโอกาสที่จะเกิดกฎทั้งสองนั้นคือค่าสนับสนุนที่ 0.187 และ 0.159 หรือ จากกฎอื่นเช่น หากมีการเดินทางเชียงใหม่ อุดรธานี มีความเป็นไปได้ที่จะเดินทางต่อมาดอนเมือง ร้อยละ 81.1 จากค่าสนับสนุน 0.024



รูปที่ 4.9 ผลลัพธ์จากการทำ Association Rules Cluster 1, Set Support = 0.95, Confidence = 0.8



รูปที่ 4.10 ผลลัพธ์จากการทำ Association Rules Cluster 1, Set Support = 0.1, Confidence = 0.5

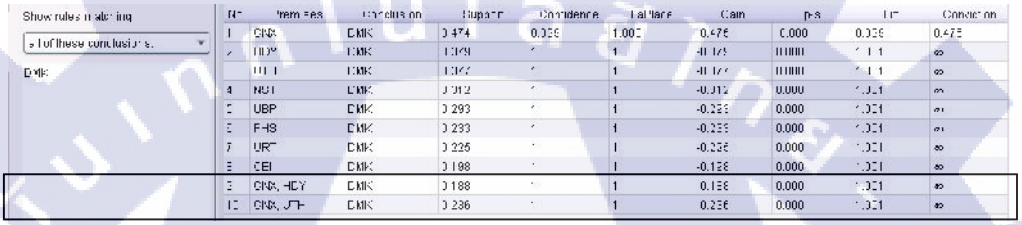
รูปที่ 4.11 Statistic ของกลุ่มที่ 2 Cluster 1

<u>กลุ่มที่ 2 Group 2 (Cluster 1) จำนวน 275 ราย</u>

ลักษณะทั่วไป : จะเป็นสมาชิกที่อยู่ในกลุ่มสมาชิก Nok Smile

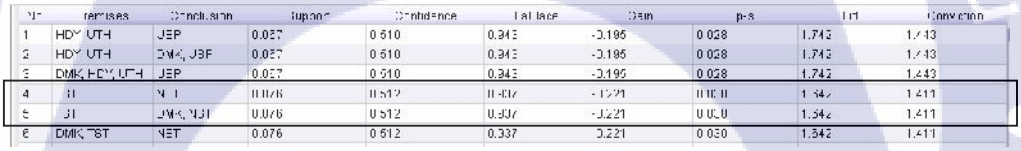
พฤติกรรม : สมาชิกในกลุ่มนี้จะมี Recency, Frequency และ Monetary เยอะกว่ากลุ่มที่ 1 และจะไม่ค่อยได้สนใจซื้อตั๋วในช่วงโปรโมชั่นเท่าไหร่นักสมาชิกกลุ่มนี้จัดอยู่ในกลุ่ม True Friends ลูกค้าที่มีกำไรสูงและซื้อสินค้าหรือบริการจากเราเป็นระยะเวลานาน จากการทำ Association Rule จะเห็นได้ว่า หากมีการเดินทางดอนเมือง เชียงใหม่ ก็เป็นไปได้ว่าจะเดินทางเชียงใหม่ อุดรธานี ถือเป็นกลุ่มที่น่าสนใจที่สุด จากรูปที่ 4.8 แสดงให้เห็นถึงความสัมพันธ์การใช้บริการภายในกลุ่มที่ 2 โดยกฎส่วนใหญ่นั้นจะเกี่ยวข้องกับการเดินทางในเส้นทางหลัก เช่น ดอนเมือง เชียงใหม่ อุดรธานี อุบลราชธานี หาดใหญ่ เป็นต้น จากการทดลองครั้งแรกโดยกำหนดค่าสนับสนุนไว้เพียง 0.95 และค่าความเชื่อมั่นต่ำสุดเท่ากับ 0.8 ไม่ทำให้เห็นค่าความสัมพันธ์ชัดเจนนักและเป็นไปตามที่นักการตลาดคาดการณ์ไว้อยู่แล้วว่าลูกค้าส่วนใหญ่ก็จะบินจากเส้นทางหลักๆ จึงได้ทดลองปรับโดยกำหนดค่าสนับสนุนไว้เพียง 0.1 และค่าความเชื่อมั่นต่ำสุดเท่ากับ 0.5 เพื่อให้เห็นค่าความสัมพันธ์ที่ชัดเจนขึ้นดังรูปที่ 4.9 จะเห็น

กฎที่ 1 ว่าลูกค้าบางคนที่มีการเดินทาง เชียงใหม่ สุราษฎร์ธานีก็อาจจะมีการใช้บริการบินไปยังพิษณุโลกหรือเดินทาง ดอนเมือง เชียงใหม่ สุราษฎร์ธานี พิษณุโลก ด้วยร้อยละ 50 จากค่าความเชื่อมั่นที่เท่ากัน 0.500 จากค่าสนับสนุน 0.105 และเดินทางเชียงใหม่ หาดใหญ่ นครศรีธรรมราช จากค่าความเชื่อมันที่ 0.507 ค่าสนับสนุนที่ 0.135 คือ โอกาสที่จะเกิดกฎนี้เพียงร้อยละ 13.5 หากเกิดแล้วโอกาสที่จะเป็นไปตามกฎนั้นเท่ากับร้อยละ 50.7 ซึ่งจากความสัมพันธ์ที่ได้มีความน่าสนใจตรงที่การเดินทางไปพิษณุโลก ซึ่งไม่ใช่เส้นทางหลักและเห็นได้ว่ามีการเดินทางที่น่าสนใจในอนาคตแม้จะมีโอกาสน้อย แต่คุณภาพของกฎนั้นจัดว่าอยู่ในเกณฑ์ดีสังเกตได้จากค่า Lift ซึ่งคือหน่วยวัดความสัมพันธ์ของกฎในแต่ละกฎนั้นจะมีค่ามากกว่า 1 ขึ้นไป



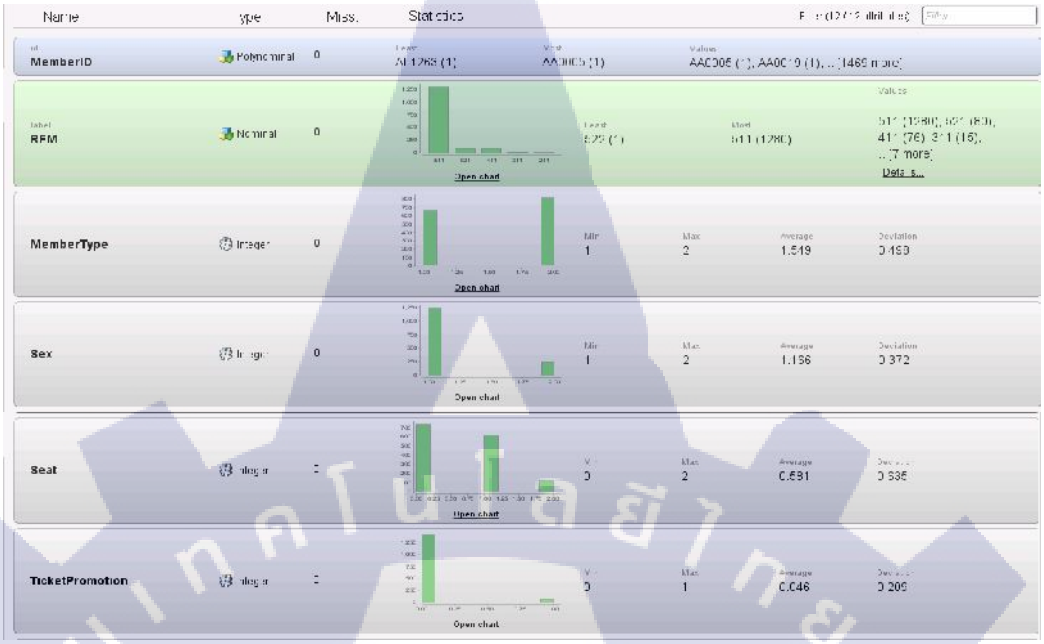
	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	CNX	DMK	0.474	0.838	1.000	0.478	0.000	0.838	0.478
2	HDY	DMK	0.439	[illegible]	1	-0.175	0.000	[illegible]	∞
3	UTH	DMK	0.362	[illegible]	1	-0.175	0.000	[illegible]	∞
4	NST	DMK	0.312	[illegible]	1	-0.312	0.000	[illegible]	∞
5	UBP	DMK	0.293	[illegible]	1	-0.293	0.000	[illegible]	∞
6	PHS	DMK	0.293	[illegible]	1	-0.293	0.000	[illegible]	∞
7	URT	DMK	0.226	[illegible]	1	-0.226	0.000	[illegible]	∞
8	CEI	DMK	0.198	[illegible]	1	-0.198	0.000	[illegible]	∞
9	CNX, HDY	DMK	0.188	[illegible]	1	0.188	0.000	[illegible]	∞
10	CNX, UTH	DMK	0.236	[illegible]	1	0.236	0.000	[illegible]	∞

รูปที่ 4.12 ผลลัพธ์จากการทำ Association Rules Cluster 2, Set Support = 0.95, Confidence = 0.8



	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	HDY, UTH	JEP	0.037	0.510	0.343	-0.195	0.028	1.743	1.443
2	HDY, UTH	DMK, JEP	0.037	0.510	0.343	-0.195	0.028	1.743	1.443
3	DMK, HDY, UTH	JEP	0.037	0.510	0.343	-0.195	0.028	1.743	1.443
4	JT	NST	0.076	0.512	0.337	-0.221	0.050	1.642	1.411
5	JT	DMK, NST	0.076	0.512	0.337	-0.221	0.050	1.642	1.411
6	DMK, JT	NST	0.076	0.512	0.337	-0.221	0.050	1.642	1.411

รูปที่ 4.13 ผลลัพธ์จากการทำ Association Rules Cluster 2, Set Support = 0.1, Confidence = 0.5



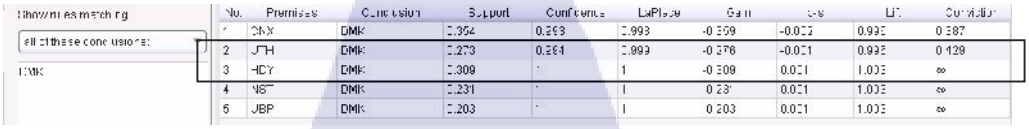
รูปที่ 4.14 Statistic ของกลุ่มที่ 3 Cluster 2

<u>กลุ่มที่ 3 Group 3 (Cluster 2) จำนวน 1,930 ราย</u>

ลักษณะทั่วไป : จะเป็นกลุ่มสมาชิกที่อยู่ในกลุ่มสมาชิก Nok Smile Plus

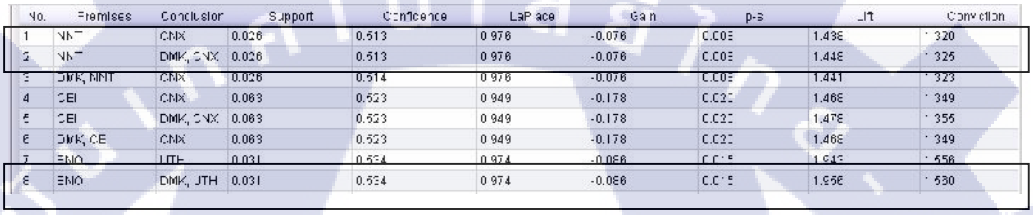
พฤติกรรม : สมาชิกในกลุ่มนี้จะเป็นกลุ่มที่มี Recency, Frequency ในการใช้บริการค่อนข้างบ่อยไม่ค่อยสนใจซื้อตั๋วในช่วงที่มีโปรโมชั่นทำให้เกิด Monetary ที่สูง มักจะเดินทางในช่วงวันธรรมดาจันทร์-ศุกร์ และชอบเลือกที่นั่งริมหน้าต่าง สมาชิกในกลุ่ม Butterflies ลูกค้าที่พร้อมจะเปลี่ยนไปได้ทุกเมื่อ (ลูกค้าที่มีกำไรสูง และซื้อสินค้าหรือบริการจากเราในระยะสั้น) โดยจากรูปที่ 4.13 แสดงให้เห็นถึงความสัมพันธ์การใช้บริการภายในกลุ่มที่ 3 โดยกฎส่วนใหญ่นั้นจะเกี่ยวข้องกับการเดินทางในเส้นทางหลักๆ เช่น ดอนเมือง เชียงใหม่ อุดรธานี อุบลราชธานี หาดใหญ่ เป็นต้น แต่จากความสัมพันธ์ในกลุ่มนี้จะเห็นถึงความน่าสนใจในกฎข้อที่ 9 และ 10 ที่มีการเดินทาง ดอนเมือง เชียงใหม่ อุดรธานี และหาดใหญ่ซึ่งมีความเป็นไปได้ว่าลูกค้าในกลุ่มนี้มีการเดินทางอาจจะซื้อตั๋วเดินทางจากเชียงใหม่ไป อุดร และเดินทางจาก อุดรมาดอนเมืองต่อเลย เช่นเดียวกันกับหาดใหญ่โดยลูกค้าในกลุ่มที่ 3 นี้มีการใช้ Monetary เป็นอันดับ 2 ใน ทั้งหมด 4 กลุ่ม ผู้ศึกษาจึงได้ทดลองกำหนดค่าความเชื่อมั่นที่ 0.5 และค่าสนับสนุนที่ 0.1 เพื่อหาค่าความสัมพันธ์ได้ดังรูปที่ 4.12 ซึ่งจะเห็นได้ว่ามีเส้นทางที่ได้มาอีก 1 เส้นทางในกลุ่มนี้ที่ไม่ค่อยได้เห็นนักในกฎข้องที่ 4 และ 5 เป็นการเดินทาง ตรัง นครศรีธรรมราช ดอนเมือง ร้อยละ 51 จากค่าความเชื่อมั่น 0.512 ค่าสนับสนุนที่ 0.076 แม้โอกาสที่

จะเกิดกฎดังกล่าวจะน้อยแต่ถือว่ามีความน่าสนใจในกลุ่มนี้ โดยมีค่า Lift มากกว่า 1 ซึ่งเป็นหน่วยวัดความสัมพันธ์ของกฎนั้นพบว่ากฎความสัมพันธ์ต่างๆ ที่ได้มาอยู่ในเกณฑ์ดี



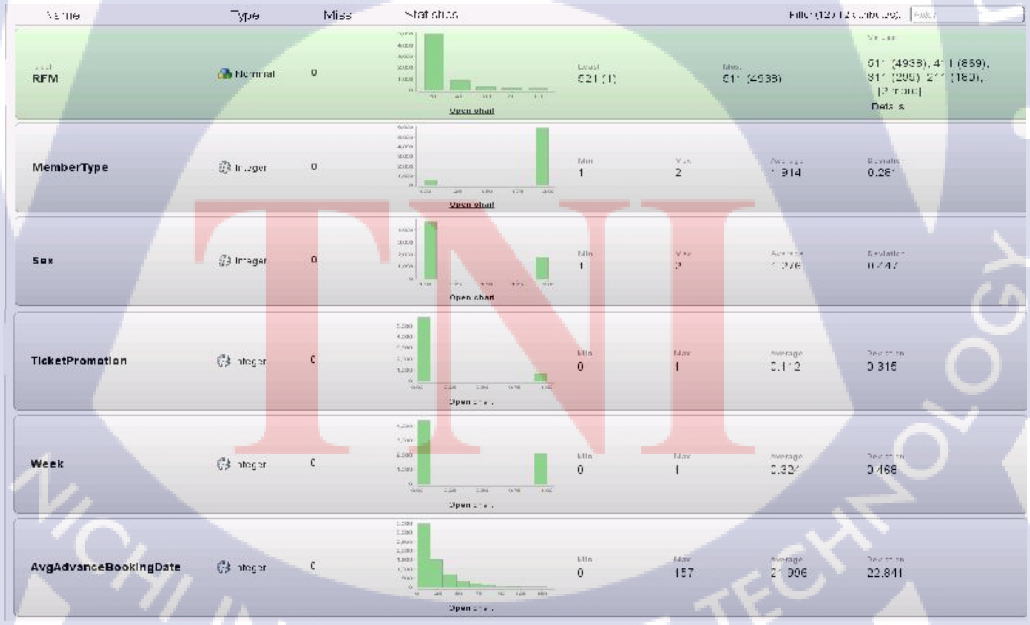
No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	CNX	DMK	0.354	0.993	0.993	-0.359	-0.001	0.995	0.987
2	JTH	DMK	0.273	0.994	0.994	-0.276	-0.001	0.995	0.429
3	HDY	DMK	0.309	1	1	-0.309	0.001	1.003	∞
4	NST	DMK	0.231	1	1	0.231	0.001	1.003	∞
5	UBP	DMK	0.203	1	1	0.203	0.001	1.003	∞

รูปที่ 4.15 ผลลัพธ์จากการทำ Association Rules Cluster 3, Set Support = 0.95, Confidence = 0.8



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	NNT	CNX	0.026	0.513	0.976	-0.076	0.001	1.438	1.320
2	NNT	DMK, CNX	0.026	0.513	0.976	-0.076	0.001	1.448	1.325
3	DMK, NNT	CNX	0.026	0.514	0.976	-0.076	0.001	1.441	1.323
4	CEI	CNX	0.063	0.523	0.949	-0.178	0.021	1.466	1.349
5	CEI	DMK, CNX	0.063	0.523	0.949	-0.178	0.021	1.478	1.355
6	DMK, CE	CNX	0.063	0.523	0.949	-0.178	0.021	1.466	1.349
7	EMO	UTH	0.031	0.534	0.974	-0.088	0.011	1.942	1.556
8	EMO	DMK, JTH	0.031	0.534	0.974	-0.088	0.011	1.956	1.530

รูปที่ 4.16 ผลลัพธ์จากการทำ Association Rules Cluster 3, Set Support = 0.1, Confidence = 0.5



รูปที่ 4.17 Statistic ของกลุ่มที่ 4 Cluster 3

กลุ่มที่ 4 Group 4 (Cluster 3) จำนวน 9,356 ราย

ลักษณะทั่วไป : จะเป็นกลุ่มสมาชิกที่อยู่ในกลุ่มสมาชิก Nok Smile Plus

พฤติกรรม : สมาชิกในในกลุ่มนี้ถือว่าเป็นกลุ่มที่ Barnacles ลูกค้าที่มีกำไรต่ำ เพราะมี Frequency Monetary ในการใช้บริการข้อนข้างต่ำ แต่ยังมี Recency ในการใช้บริการอยู่ โดยจากรูปที่ 4.14แสดงให้เห็นถึงความสัมพันธ์การใช้บริการภายในกลุ่มที่ 4 จำนวน 5 กฎ โดยกฎส่วนใหญ่นั้นจะเป็นการเดินทางใน 5 เส้นทางหลักๆ เชียงใหม่ อุดรธานี นราธิวาส หาดใหญ่ และอุบลราชธานี โดยในกฎที่ 1 การเดินทางเชียงใหม่ ดอนเมืองแม้ว่าจะมีค่าสนับสนุนเพียง 0.354 แต่หากเกิดแล้วจะมีโอกาสเป็นไปตามกฎมากถึงร้อยละ 99.3 จากค่าความเชื่อมันที่เท่ากับ 0.993 นั่นเอง และกฎที่ 2 เดินทางอุดรธานี ดอนเมือง มีค่าสนับสนุนเพียง 0.273 แต่หากเกิดแล้วจะมีโอกาสเป็นไปตามกฎมากถึงร้อยละ 99.4 จากค่าความเชื่อมันที่เท่ากับ 0.994 หากทดลองกำหนดค่าความเชื่อมั่นที่ 0.5 และค่าสนับสนุนที่ 0.1 เพื่อหาค่าความสัมพันธ์ได้ดังรูปที่ 4.15 จะได้กฎความสัมพันธ์ทั้งหมด 80 กฎ จากรูปจะเห็นกฎที่น่าสนใจคือกฎที่ 1, 2 คือการเดินทางน่าน เชียงใหม่ ดอนเมือง โดยค่าความเชื่อมั่นร้อยละ 51.3 และค่าสนับสนุนที่ 0.026 และกฎข้อ 7, 8 เดินทาง สกลนคร สุราษฎร์ธานี ดอนเมือง ค่าความเชื่อมั่นร้อยละ 53.4 และค่าสนับสนุนที่ 0.031 โดยเส้นทาง น่าน และสกลนครเป็นเส้นทางที่ไม่ใช่เส้นทางหลักในการใช้บริการจะเห็นได้ว่าลูกค้ากลุ่มนี้มีการเดินทางไปสกลนคร และน่านบ้าง แต่อย่างไรลูกค้ากลุ่มนี้ก็ไม่เหมาะที่จะทำการตลาดด้วยนักเนื่องจาก Monetary ในการใช้บริการค่อนข้างต่ำ แต่ยังมีการใช้บริการอยู่เรื่อยๆ

บทที่ 5
สรุปผลการศึกษา

จากการศึกษานี้เป็นการศึกษาเพื่อนำมาใช้ในการบอกถึงลำดับความสำคัญของลูกค้าหรือทราบถึงกลุ่มลูกค้าที่มีคุณค่า นอกเหนือจากการทราบถึงพฤติกรรมการใช้บริการ ซึ่งเป็นผลที่ได้จากการจัดกลุ่มโดยทั่วไป โดยใช้เทคนิค Data Mining ในการสร้างความสัมพันธ์กับลูกค้าโดยใช้เทคนิคต่างๆ ไม่ว่าจะเป็น Decision Tree, Clustering, และ Association Rules มาเพื่อที่จะค้นหาความต้องการของลูกค้า อีกทั้งยังสามารถนำข้อมูลมาเพื่อมาปรับปรุงพัฒนาการบริการให้ดีขึ้น ฉะนั้นแล้วการนำเทคโนโลยีเหมืองข้อมูล ซึ่งเป็นเทคโนโลยีที่ได้รับการยอมรับว่าเป็นเครื่องมือที่นำไปสู่การเข้าใจลูกค้า จึงถูกนำมาใช้เพื่อวิเคราะห์พฤติกรรม ความต้องการ ตลอดจนแนวโน้มในการใช้จ่ายของลูกค้า อีกทั้งหลายองค์กรยังหยิบมาใช้เพื่อช่วยในการวิเคราะห์สิ่งต่างๆ เช่น ช่วยในการพยากรณ์การยอดการให้บริการในอนาคต ช่วยหาความสัมพันธ์ของสินค้าหรือการบริการ โดยผลที่ได้จากการวิเคราะห์จะนำข้อมูลไปปรับปรุงการทำงานรวมถึงสร้างเป็นโปรโมชั่นให้ตรงกับความต้องการ หรือนำไปสู่การตัดสินใจในการลงทุนภายในองค์กรต่อไป

โดยผลจากการศึกษาพฤติกรรมอ้างอิงงานวิจัยก่อนหน้า [13] นำมาใช้สมาชิกกลุ่มนกแฟนคลับที่ใช้บริการในปี 2012-2013 จำนวน 62,496 ราย จากการแบ่งช่วงอายุระหว่าง 18-80 ปี ซึ่งถือว่าเป็นบุคคลที่บรรลุภาวะเรียบร้อยแล้ว หลังจากการศึกษาพบว่ามีข้อจำกัดบางอย่างของข้อมูล เช่น การขาดหายของข้อมูลเกิดจากเริ่มระบบสมัครสมาชิกไม่ได้มีการดักไว้ให้กรอกข้อมูลตามจริง ลูกค้าบางคน หรือแม้แต่พนักงานผู้รับสมัครอาจเกิดการกรอกข้อมูลที่ผิดพลาดบ่อย หากมีปริมาณข้อมูลครบถ้วนก็จะส่งผลให้การศึกษาพฤติกรรมมีความชัดเจนมากขึ้น ดังนั้น งานศึกษานี้ ผู้ศึกษาได้ทดลองจากข้อมูลที่มีในปี 2012-2013 และอธิบายถึงพฤติกรรมการใช้บริการ โดยทำการทดลองโดยการให้คะแนน RFM เพื่อนำไปสู่การทำเหมืองข้อมูล นำไปจัดกลุ่ม เพื่อหาค่า Centroid ของข้อมูล จำแนกข้อมูล และการหาความสัมพันธ์

ผลการทดลองหลังจากการจัดกลุ่มด้วย K-Means อัลกอริทึม ผลจากการแบ่งกลุ่มนั้นนอกจากทราบถึงลักษณะพฤติกรรมการใช้งานของลูกค้าแล้ว การประยุกต์ใช้ค่าเทคนิคการแบ่งกลุ่มทางการตลาด หรือ RFM สามารถช่วยให้ทราบถึงอันดับความสำคัญของกลุ่มลูกค้า ซึ่งเป็นประโยชน์มากในการทำธุรกิจ ช่วยสร้างกลยุทธ์ในการให้บริการหรือกลยุทธ์ทางการตลาด ซึ่งสามารถเห็นได้ชัดเจนจากชุดข้อมูลลูกค้าบริษัทจะแบ่งกลุ่มออกมาได้ 4 กลุ่ม โดยผลจากการจัดกลุ่มนั้นจะกลุ่มออกมาในรูปแบบการเรียงลำดับของข้อมูล เนื่องจากปัจจัยด้านราคาของราคาตั๋วเครื่องบินในการใช้บริการ ผลลัพธ์จากการทำ K-Means และ Decision Tree จะเห็นได้ว่ากลุ่มสมาชิกที่ตกอยู่ในกลุ่มที่

2 และ 3 ความน่าสนใจที่สุดโดยที่ Group 2 มี Centroid Recency ในการใช้บริการอยู่ที่ 19.57 Frequency ในการใช้บริการอยู่ที่ 126.62 และ Centroid ของ Monetary ในการใช้บริการอยู่ที่ 282,235 บาท และ Group 3 มี Centroid Recency ในการใช้บริการอยู่ที่ 26.55 Frequency ในการใช้บริการอยู่ที่ 63.66 และ Centroid ของ Monetary ในการใช้บริการอยู่ที่ 120,446.24 บาท ด้านพฤติกรรมการใช้งานทั่วไปนั้นส่วนใหญ่ทั้ง 4 กลุ่มมีลักษณะพฤติกรรมการใช้งานที่ค่อนข้างคล้ายคลึงกันคือเดินทางจากเส้นทางหลักๆ แต่เมื่อวิเคราะห์แล้วมีกลุ่มที่น่าสนใจอยู่ 2 กลุ่มคือ กลุ่มที่ 2 และ 3 ที่เหมาะแก่การทำการตลาดมากที่สุดเนื่องจากเป็นกลุ่มที่ก่อรายได้ให้กับบริษัทมาก ซึ่งเป็นสิ่งที่บริษัทสนใจ

จากการแบ่งกลุ่มข้อมูลข้างต้นนั้นจะเห็นว่าการนำเทคนิคการแบ่งกลุ่มทางการตลาดหรือ RFM เข้ามาช่วยในการแบ่งกลุ่มจะสามารถช่วยให้ทราบถึงลำดับความสำคัญของกลุ่มแต่ละกลุ่มได้อย่างมีประสิทธิภาพ หากแต่การคำนวณค่า RFM นั้นของลูกค้าแต่ละรายนั้นประกอบด้วยปัจจัยต่างๆ อันได้แก่วันที่เข้าใช้บริการล่าสุด เพื่อบอกถึงสถานะ Recency ของผู้ใช้ จำนวนเที่ยวบินทั้งหมดเพื่อให้ทราบถึง Frequency ในการใช้บริการของลูกค้าและ Monetary คือจำนวนเงินเพื่อบ่งบอกถึงรายได้ที่ได้รับจากลูกค้ารายนั้นๆ ดังนั้นโอกาสที่ค่า RFM ของลูกค้าแต่ละรายมีจำนวนเท่ากันหากแต่เกิดจากปัจจัยที่แตกต่างย่อมสามารถเป็นไปได้ หรือการที่มีค่า RFM สูงอาจจะไม่ได้เป็นลูกค้าชั้นดีหรือมีความสำคัญมากเสมอไป ขึ้นกับมุมมอง และการให้ความสำคัญในแต่ละปัจจัยโดยอาจจะแตกต่างไปในแต่ละช่วงเวลา หรือนโยบายของผู้บริหาร

5.1 ประโยชน์ที่ได้รับจากการทำสารนิพนธ์นี้

ทำให้ทราบถึงลักษณะพฤติกรรมการใช้งานของลูกค้าแล้ว การประยุกต์ใช้ค่าเทคนิคการแบ่งกลุ่มทางการตลาด หรือ RFM สามารถช่วยให้ทราบถึงอันดับความสำคัญของกลุ่มลูกค้า ซึ่งเป็นประโยชน์มากในการทำธุรกิจ ช่วยสร้างกลยุทธ์ในการให้บริการหรือกลยุทธ์ทางการตลาดแต่ในทางกลับกันการจัดอันดับความสำคัญของกลุ่มโดยอาศัยค่า RFM นั้น ก็อาจไม่ถูกต้องหรือตรงต่อความต้องการของธุรกิจเสมอไปเมื่อค่า RFM ที่ได้เกิดจากปัจจัยที่มีความสำคัญน้อยสำหรับกลุ่มธุรกิจนั้นๆ

5.2 ข้อเสนอแนะ

1. การศึกษานี้ได้เลือกเกณฑ์ศึกษาข้อมูลจากกลุ่มสมาชิกนกแฟนคลับมาเพื่อศึกษาพฤติกรรมการใช้บริการโดยทั้งนี้อายุการเป็นสมาชิกนกแฟนคลับจะมีอายุอยู่เพียงแค่ 2 ปี และจะต้องต่ออายุสมาชิก ดังนั้นควรมีการทดลองทำการศึกษาพฤติกรรมของสมาชิกทุกๆ 2 ปี

2. รูปแบบการของข้อมูลมีความผิดพลาดเยอะซึ่งข้อมูลที่ได้มาเกิดจากการกรอกข้อมูลของสมาชิกตั้งแต่ต้น ทำให้เกินความผิดพลาดในการศึกษา ดังนั้นการศึกษาครั้งต่อไปควรศึกษาวิธีที่จะทำให้ได้มาซึ่งข้อมูลที่ถูกต้องครบถ้วน

3. การศึกษานี้เป็นการทำการศึกษาเฉพาะพฤติกรรมในกลุ่มสมาชิกนกแฟนคลับเท่านั้น ควรมีการนำข้อมูลของกลุ่มลูกค้าที่ไม่ได้เป็นสมาชิกนกแฟนคลับเข้ามาเพื่อศึกษาพฤติกรรมการเดินทาง เพื่อให้ได้ซึ่งข้อมูลซึ่งสามารถเป็นประโยชน์แก่การออกโปรโมชั่นสำหรับกลุ่มลูกค้าที่ไม่ได้เป็นสมาชิกนกแฟนคลับหรือเพื่อเชิญกลุ่มลูกค้าเข้ามาเป็นสมาชิกนกแฟนคลับได้มากขึ้น

4. ในอนาคตระบบสมาชิกนกแฟนคลับจะมีการ Link ข้อมูลกับ Facebook จะทำให้การเก็บข้อมูลของสมาชิกมีความครบถ้วนมากขึ้น

บรรณานุกรม

บรรณานุกรม

[1] Dick Alan and BasuKunal, “Customer Loyalty: Toward an Integrated Conceptual Framework,” *Journal of the Academy of Marketing Science*, vol.37, no.2, pp.170-180, 1994.

[2] Werner J. Reinartz and V. Kumar, “The mismanagement of customer loyalty,” *Harvard business review*, vol.80, no.7, pp.86-95,2002.

[3] Oliver Richard L and William O Bearden, *Satisfaction: A Behavioral Perspective on the Consumer*, New York: Irwin/McGraw, 1997.

[4] Tor WallinAndreassenand BodilLindestad, “Customer Loyalty and Complex Services: The Impact of Corporate Image on Quality, Customer Satisfaction and Loyalty for Customers with Varying Degrees of Service Expertise,” *International Journal of Service Industry Management*, vol.9, no.1, pp.7-23, 1998.

[5] Arjun Chaudhuri and Morris B. Holbrook,“The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty,” *Journal of Marketing*, vol.65, no.2, pp.81-93, 2001.

[6] ปรัชญา ผ่องอักษร, *การค้นหารูปแบบช่วงของข้อมูลโดยใช้คลาสเพื่อลดปริมาณช่วงของรูปแบบที่ได้,* กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์, 2552.

[7] Jehn-Yih Wong and Pi-Heng Chung, “Retaining Passenger Loyalty through Data Mining: A Case Study of Taiwanese Airlines,” *Transportation Journal (American Society of Transportation & Logistics Inc)*, vol.47, pp.17-29, 2008.

[8] วามิณี นิยากาศ, *การวิเคราะห์พฤติกรรมการใช้บริการของลูกค้าอินเทอร์เน็ตแบงก์กิ้งด้วยเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนและ RFM Analysis*, กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์, 2549.

[9] Lisa Pritscher and Hans Feyen. “*Data Mining and Strategic Marketing in the Airline Industry,*” [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.124.7062. [Accessed: September 7, 2001].

[10] Agrawal and Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” *Proceedings of the 20th International Conference on Very Large Data Bases,* pp.487-499, 1994.

[11] กฤษฎากร กิ่งอุบล, *การใช้เกณฑ์ความต่างลำดับในการปรับปรุงกฎความสัมพันธ์จำแนกประเภทข้อมูล*, กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์, 2552

[12] นาวิก นำเสียง, “*หลักการ อาร์เอฟเอ็ม (RFM),*” [Online]. Available: http://www.crminaction.com/article/7/198/. [Accessed: June 18, 2002].

[13] DeryaBirant, “*Data Mining Using RFM Analysis,*” [Online]. Available: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/data-mining-using-rfm-analysis. [Accessed: January 21, 2011].

ภาคผนวก

ภาคผนวก ก.

ขั้นตอนการใช้งาน RapidMiner Studio 6

ขั้นตอนการใช้งาน RapidMiner Studio 6

1. การติดตั้งโปรแกรม

สามารถดาวน์โหลดข้อมูลได้จาก http://rapidminer.com เลือกที่เมนู Products และ RapidMiner Studio ในตอนเริ่มต้นใช้งานซอฟต์แวร์จะถามว่าเราจะใช้ RapidMiner Studio ด้วย License แบบไหนซึ่งมี 2 แบบ คือ Starter และ Professional โดยแบบแรกจะใช้งานได้ฟรีไม่ต้องเสียค่าใช้จ่ายแต่จะมีข้อจำกัดในการดึงข้อมูลจากฐานข้อมูลแต่แบบหลังจะต้องเสียค่าใช้จ่ายจะใช้แบบ Starter ซึ่งมีฟีเจอร์เหมือนกันกับระบบ Professional ดังนั้นเลือก Continue Using Starter

รูปที่ ก.1 หน้าต่าง Welcome ของโปรแกรม RapidMiner Studio 6

2. เริ่มต้นใช้งานโปรแกรม RapidMiner Studio 6

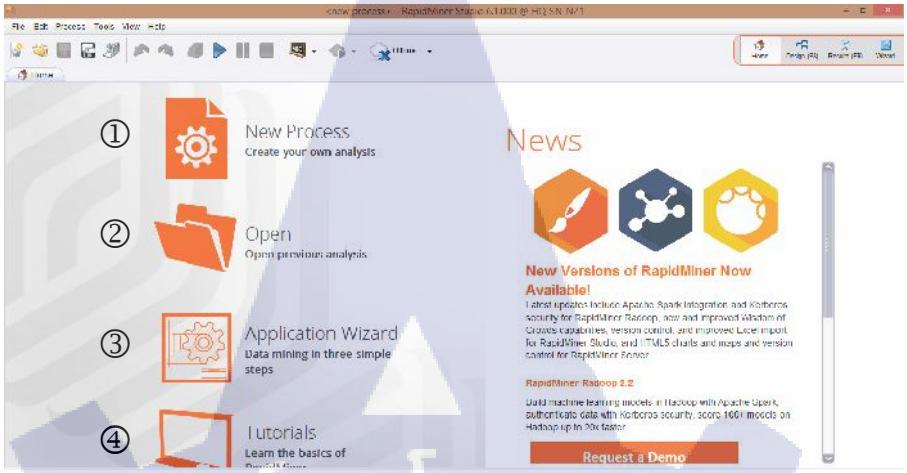
เมื่อคลิกที่เมนูดังกล่าวแล้วจะพบกับหน้าจอเริ่มการใช้งานซอฟต์แวร์ดังในรูปที่ ก.2 ซึ่งประกอบด้วยเมนูต่างๆ ดังนี้

2.1 New Process เมนูนี้ใช้สำหรับการสร้างโพรเซส (Process) ใหม่ขั้นตอนการทำงานของ RapidMiner Studio เราจะเรียกว่าโพรเซสซึ่งประกอบด้วยการนำโอเปอเรเตอร์ (Operator) ต่างมาเชื่อมต่อกัน

2.2 Open เมนูนี้ใช้สำหรับการเลือกไฟล์โพรเซสที่ได้สร้างไว้แล้วกลับมาใช้งานใหม่อีกครั้ง

2.3 Application Wizard เมนูนี้ใช้สำหรับเปิดโพรเซสที่ RapidMiner Studio เตรียมไว้ให้

2.4 Tutorials เมนูนี้ใช้เปิดระบบการสอนที่ RapidMiner Studio เตรียมไว้ให้



รูปที่ ก.2 หน้าต่างเริ่มต้นการใช้งานของ RapidMiner Studio 6

3. องค์ประกอบของ RapidMiner Studio 6

เมื่อเราคลิกที่เมนู New Process แล้วจะเข้าสู่หน้าจอการออกแบบ (Design Perspective) ของซอฟต์แวร์ RapidMiner Studio 6 ดังรูปที่ ก.3 ซึ่งสามารถแบ่งเป็นส่วนต่างๆ ได้ 5 ส่วน ดังนี้

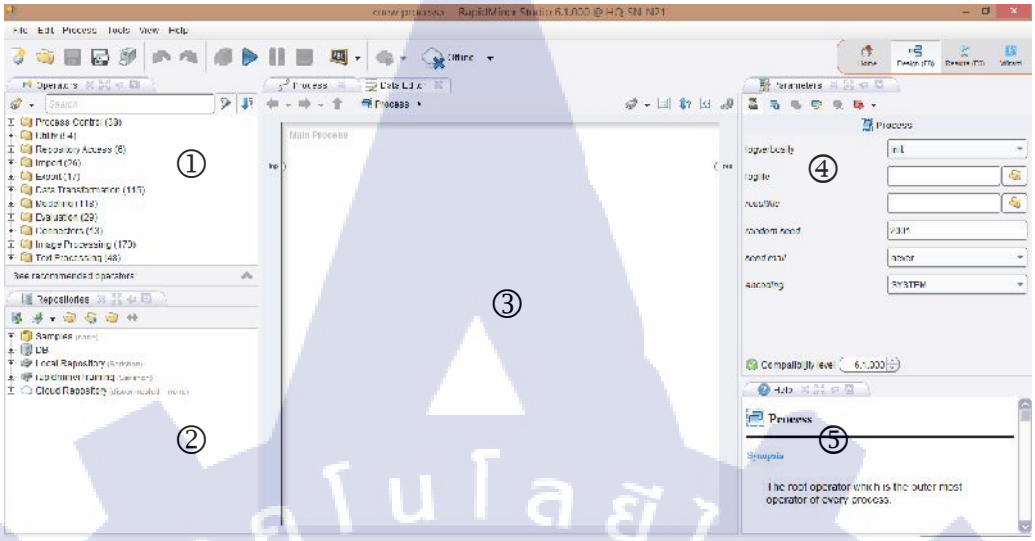
3.1 Operators ส่วนนี้จะเก็บโอเปอเรเตอร์ในการใช้งานต่างๆ ไว้เป็นกลุ่มๆ ตามหน้าที่ที่คล้ายคลึงกันเช่นโอเปอเรเตอร์สำหรับการอ่านข้อมูลจากไฟล์ประเภท CSV จะอยู่ในหมวด Import และหมวดย่อย Data นอกจากนี้ในส่วนของโอเปอเรเตอร์นี้ยังมีส่วนสำหรับใช้ในการค้นหาชื่อของโอเปอเรเตอร์ต่างๆ ได้ซึ่งจะแสดงตัวอย่างให้ดูในหัวข้อ “ตัวอย่างการสร้างโมเดล Decision Tree”

3.2 Repositories ส่วนนี้จะใช้ในการจัดการไฟล์ต่างๆ หลักการของ RapidMiner Studio นี้จะเก็บไฟล์ข้อมูลหรือโพรเซสต่างๆ ไว้ในโฟลเดอร์เพื่อความสะดวกในการเรียกใช้งานครั้งถัดไปโฟลเดอร์ที่ใช้เก็บไฟล์เหล่านี้จะเรียกว่าเป็น Repository

3.3 Process ส่วนนี้เป็นอีกส่วนที่สำคัญของ RapidMiner Studio เพราะหลักการทำงานของซอฟต์แวร์นี้คือการนำโอเปอเรเตอร์ต่างๆ มาประกอบกันให้เป็นโพรเซสขึ้นมา รายละเอียดดูได้จากหัวข้อ “ตัวอย่างการสร้างโมเดล Decision Tree”

3.4 Parameters ส่วนนี้จะเป็นส่วนที่แสดงพารามิเตอร์ (Parameter) ที่เกี่ยวข้องกับแต่ละโอเปอเรเตอร์เช่นโอเปอเรเตอร์ Read CSV สำหรับอ่านไฟล์ CSV จะมีพารามิเตอร์ที่เกี่ยวข้องเช่นตำแหน่งของไฟล์ CSV เป็นต้น

3.5 Help ส่วนนี้จะเป็นส่วนที่แสดงข้อความช่วยเหลือหรือรายละเอียดของโอเปอเรเตอร์ที่เลือกใช้งานอยู่ซึ่งประกอบด้วยรายละเอียดเบื้องต้นความหมายของแต่ละพารามิเตอร์เป็นต้น



รูปที่ ก.3 องค์ประกอบต่างๆ ของ RapidMiner Studio 6

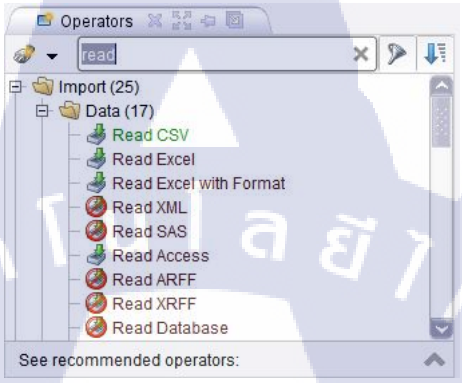
นอกจากทั้ง 5 ส่วนใหญ่ๆที่ได้อธิบายไปแล้วยังมีส่วนเมนูหลักๆ ด้านบนเพิ่มเติมดังนี้

รูปที่ ก.4 เมนูส่วนหลักๆ เพิ่มเติมของ RapidMiner Studio 6

1. เมนูสำหรับการสร้างโพรเซสใหม่
2. เมนูสำหรับสั่งให้โพรเซสทำงาน (Run Process)
3. เมนูสำหรับการเรียกดู Tutorial
4. แสดงหน้าจอกลับไปที่หน้าจอเริ่มต้น (Home)
5. แสดงหน้าจอการออกแบบ (Design Perspective)
6. แสดงหน้าจอผลลัพธ์การทำงาน (Results Perspective)
7. แสดงหน้าจอตัวอย่างที่กำหนดไว้ (Wizard Perspective)

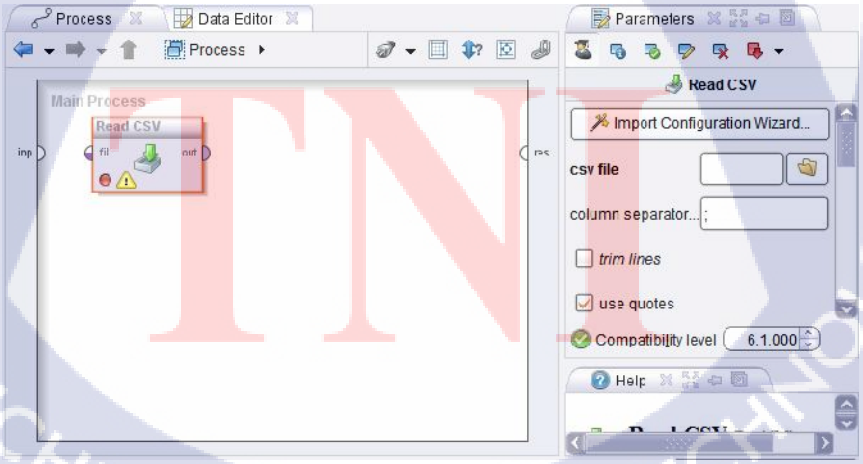
4. ทดลองสร้างโมเดล Decision Tree

4.1 เปิดหน้าจอการออกแบบและเลือกโอเปอเรเตอร์ชื่อว่า Read CSV ที่ใช้สำหรับการอ่านไฟล์ประเภท CSV (Comma Separated Values) ซึ่งคือไฟล์ที่มีเครื่องหมาย Comma คั่นระหว่างแอตทริบิวต์โอเปอเรเตอร์นี้อยู่ในหมวด Import และ Data ดังในรูปที่ ก.5



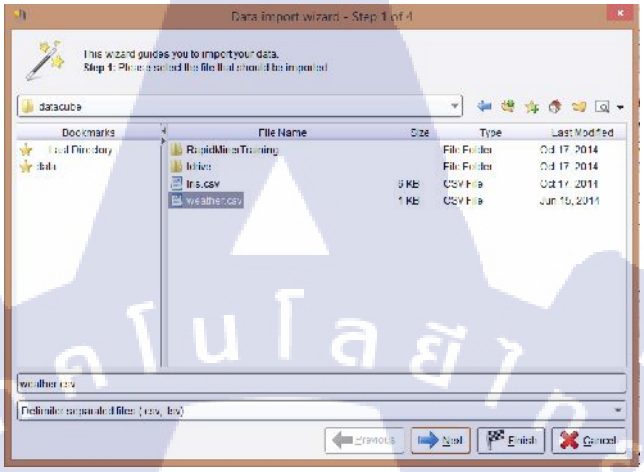
รูปที่ ก.5 เลือกโอเปอเรเตอร์ Read CSV สำหรับอ่านไฟล์ประเภท CSV

4.2 ลากโอเปอเรเตอร์ Read CSV มาวางในส่วน Main Process เมื่อไปคลิกที่โอเปอเรเตอร์นี้จะเห็นว่าที่ส่วนของพารามิเตอร์ที่อยู่ทางด้านขวาจะเปลี่ยนไปแสดงพารามิเตอร์ที่เกี่ยวข้องกับโอเปอเรเตอร์ Read CSV



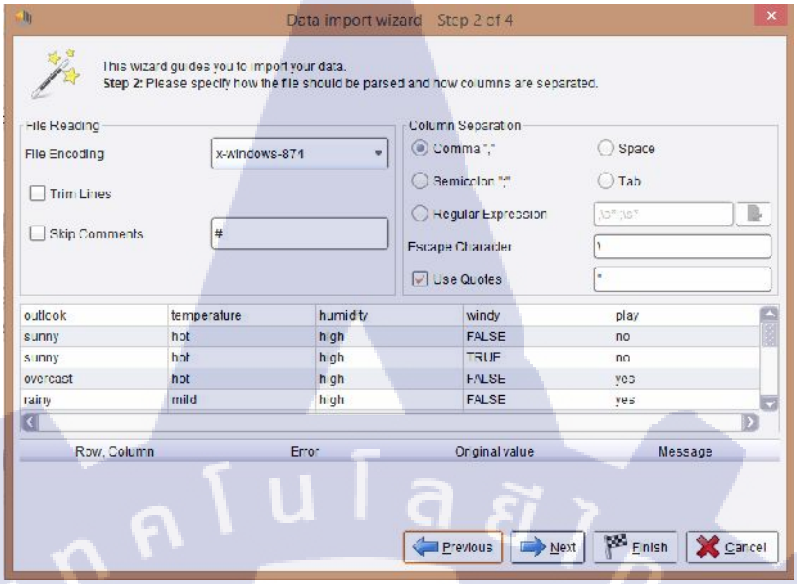
รูปที่ ก.6 โอเปอเรเตอร์ Read CSV ในส่วน Main Process

4.3 คลิกที่ปุ่ม Import Configuration Wizard เพื่อทำการอ่านไฟล์ CSV เข้ามาใช้งานใน RapidMiner Studio 6 หลังจากคลิกที่ปุ่มแล้วจะต้องเลือกไฟล์ CSV ที่จะใช้งานดังในรูปที่ ก.7



รูปที่ ก.7 เลือกไฟล์ข้อมูล .CSV

4.4 เมื่อเลือกไฟล์และกดปุ่ม Next แล้วจะเข้าสู่ขั้นตอนถัดไปซึ่งจะแสดง Option ต่างๆ ให้เราเลือกเช่นถ้าต้องการใช้ไฟล์ที่มีภาษาไทยจะต้องเปลี่ยน File Encoding ให้เป็นประเภท UTF-8 แต่ในการทดลองนี้เราจะใช้ค่าเดิมของ File Encoding แต่จะเปลี่ยนส่วน Column Separator ให้เป็น Comma “,” แทนดังในรูปที่ ก.8 หลังจากนั้นกดปุ่ม Next เพื่อไปขั้นถัดไป



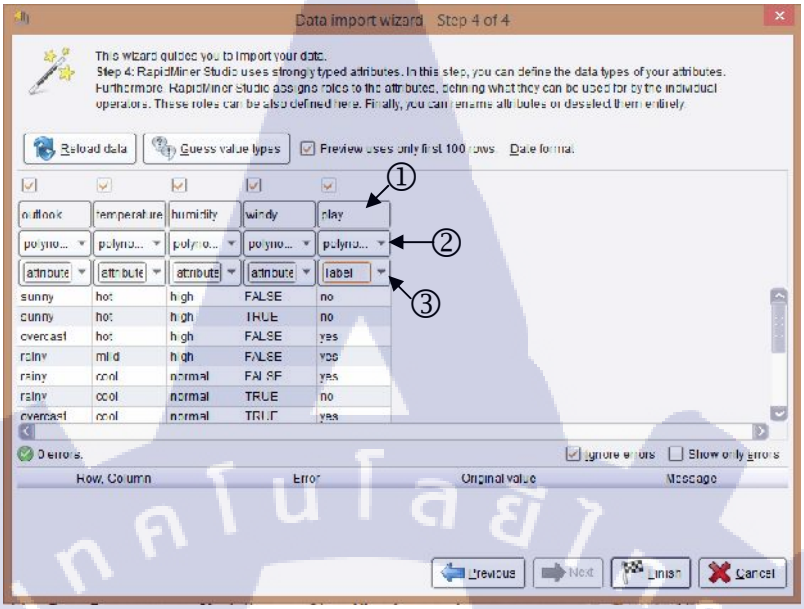
รูปที่ ก.8 การเปลี่ยน File Encoding ส่วน Column Separator ให้เป็น Comma

4.5 ในขั้นตอนนี้เราสามารถเลือกได้ว่าจะให้แถวแรกเป็นชื่อของแอตทริบิวต์หรือไม่โดยการคลิกที่ Name ที่อยู่ในคอลัมน์ Annotation ซึ่งโดยปกติแล้วก็จะถูกเลือกไว้อยู่แล้วดังนั้นในขั้นตอนนี้สามารถกดปุ่ม Next ต่อไปได้เลย

4.6 ถัดมาจะเป็นส่วนของการกำหนดประเภทของข้อมูลที่อยู่ในแต่ละแอตทริบิวต์และประเภทของแต่ละแอตทริบิวต์ดังแสดงในรูปที่ ก.9 โดย 3 แถวแรกจะเป็นตัวกำหนดค่าต่างๆ ดังนี้

1. แถวแรกของตารางจะเป็นชื่อของแอตทริบิวต์ซึ่งเราสามารถพิมพ์เพื่อเปลี่ยนชื่อ
2. แถวที่สองแสดงประเภทของข้อมูลที่เก็บอยู่ในแต่ละแอตทริบิวต์ เช่น Real สำหรับตัวเลขหรือ Polynominal สำหรับค่านอมินอล (Nominal) ที่มีค่าที่แตกต่างกันตั้งแต่ 2 ค่าขึ้นไป
3. แถวที่สามแสดงประเภทของแอตทริบิวต์ซึ่งส่วนใหญ่จะเป็นแอตทริบิวต์และลาเบลซึ่งจะเป็นแอตทริบิวต์ขนิดพิเศษที่มักจะใช้แสดงคำตอบของสิ่งที่เราต้องการจะสร้างโมเดลมาทำนายหรือเรียกว่าคลาส (Class) หรือตัวแปรตาม (Dependent Variable)

ในตัวอย่างนี้เราจะกำหนดให้แอตทริบิวต์ Play เป็นประเภทลาเบลเพื่อใช้เป็นคลาสคำตอบที่โมเดล Decision Tree ของเราต้องการทำนาย หลังจากนั้นกดปุ่ม Finish ก็จะกลับสู่หน้าจอการออกแบบอีกครั้ง

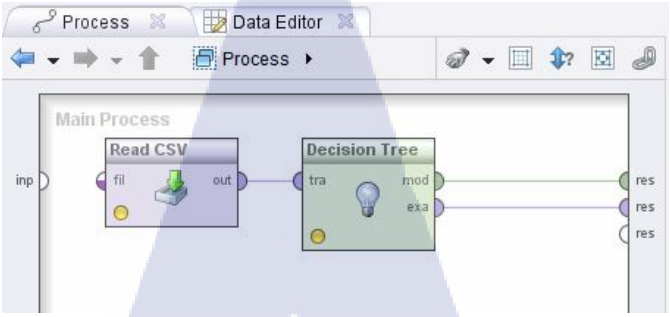


รูปที่ ก.9 ขั้นตอนการเลือกประเภทของแอตทริบิวต์

4.7 สำหรับการสร้างโมเดล Decision Tree ทำได้โดยการเลือกโอเปอเรเตอร์ Decision Tree จากส่วนของ Operators เราสามารถหาโอเปอเรเตอร์นี้ได้โดยการพิมพ์คำว่า Decision Tree ในส่วนของการค้นหา (Search) และกดปุ่ม Enter หลังจากนั้นโอเปอเรเตอร์ Decision Tree

4.8 ลากโอเปอเรเตอร์ Decision Tree มาวางในส่วนของ Main Process และลากเส้นเชื่อมจากพอร์ตที่ชื่อว่า out (ซึ่งย่อมาจากคำว่า Output) ของโอเปอเรเตอร์ Read CSV ไปยังพอร์ตที่ชื่อว่า tra (ย่อมาจากคำว่า Training) ของ Decision Tree

4.9 หลังจากนั้นลากเส้นเชื่อมจากพอร์ต mod (ย่อมาจาก Model) และพอร์ต exa (ย่อมาจาก Example) ของโอเปอเรเตอร์ Decision Tree ไปยังพอร์ต res (ย่อมาจาก Result) ทั้งสองพอร์ตของ Main Process เพื่อไปแสดงในส่วนของหน้าจอผลลัพธ์โดยพอร์ต mod จะทำส่งโมเดล Decision Tree ที่สร้างได้ออกไปแสดงในรูปต้นไม้และพอร์ exa จะส่งข้อมูลจากไฟล์ CSV ไปแสดงในรูปแบบตาราง



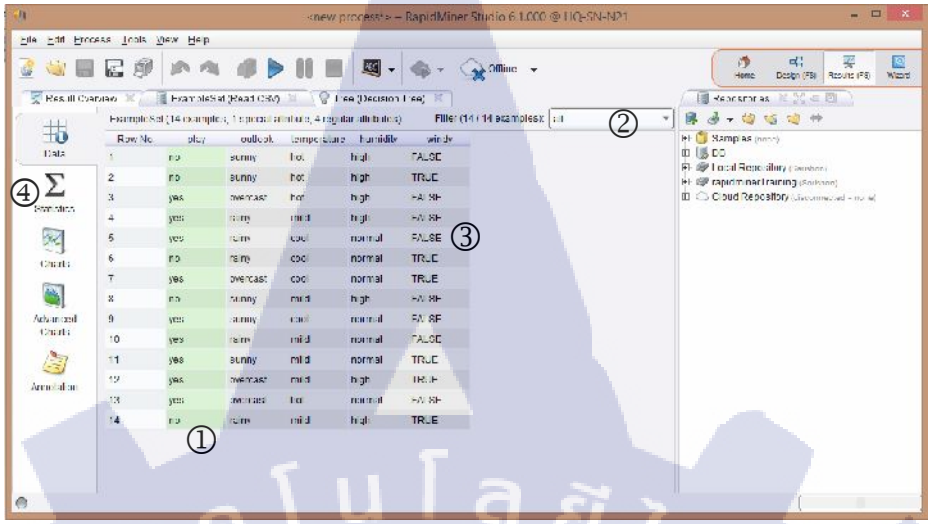
รูปที่ ก.10 การเชื่อมต่อโอเปอเรเตอร์ Read CSV กับ Decision Tree

4.10 กดปุ่ม  เพื่อเริ่มต้นการทำงานของโพรเซส

5. ผลการทดลองการทำงาน Decision Tree ใน RapidMiner Studio 6

เมื่อโพรเซสทำงานเสร็จสิ้นจะเปลี่ยนมายังหน้าจอผลลัพธ์การทำงาน (ดังในรูปที่ ก.10) ซึ่งผลที่ได้จากโพเซสที่เราสร้างจะแสดง 3 แท็บ (Tab) คือ

- Result Overview แสดงรายการผลลัพธ์การทำงานที่ผ่านมา
- Example Set (Read CSV) แสดงตารางข้อมูลที่มาจากไฟล์โดยใช้โอเปอเรเตอร์ Read CSV
- Tree (Decision Tree) แสดงโมเดล Decision Tree ที่สร้างได้จากโอเปอเรเตอร์ Decision Tree ในรูปแบบภาพ



รูปที่ ก.11 ข้อมูลที่นำมาใช้ในหน้าจอแสดงผลลัพธ์การทำงาน

จากรูปที่ ก.11 จะแสดงตารางข้อมูลที่อ่านมาจากไฟล์ CSV และใช้ในการสร้างโมเดล Decision Tree อธิบายส่วนสำคัญหลักๆ ในหน้าจอแสดงผลลัพธ์การทำงาน 4 ส่วนหลักๆ ดังนี้

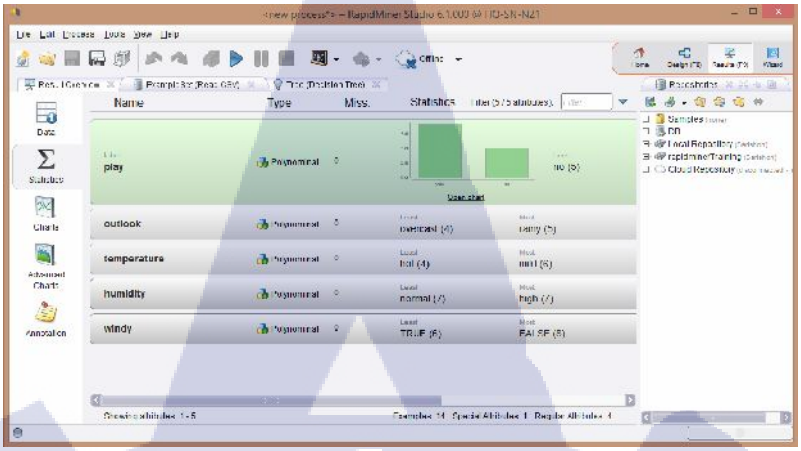
1. แสดงจำนวนตัวอย่างและแอตทริบิวต์ที่ปรากฏในข้อมูลซึ่งในไฟล์ตัวอย่างนี้มีจำนวน 14 ตัวอย่าง 1 แอตทริบิวต์ลาเบลและ 4 แอตทริบิวต์ ทั่วไป

2. ส่วนการกรองข้อมูล (Filter) ซึ่งมีให้เลือกได้ว่าจะดูข้อมูลทั้งหมดหรือข้อมูลที่มีความผิดพลาดอยู่เป็นต้น

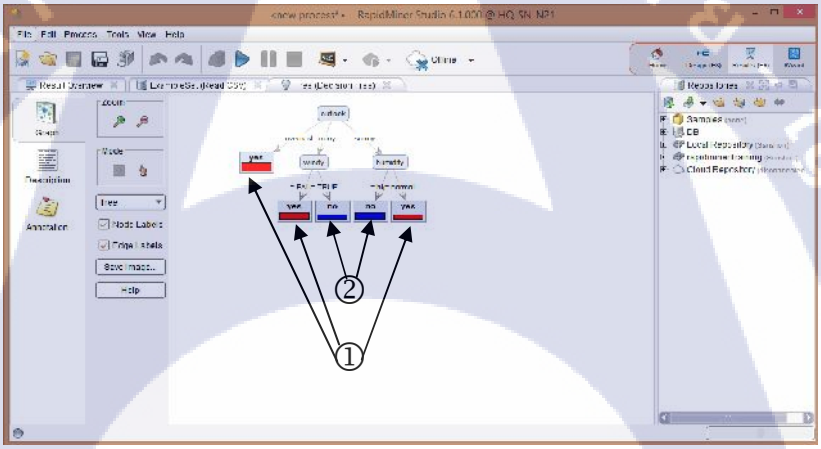
3. ในส่วนของตารางนี้เราสามารถคลิกที่ชื่อแอตทริบิวต์เพื่อทำการเรียงลำดับข้อมูลได้โดยตารางข้อมูลจะแบ่งแอตทริบิวต์ออกเป็น 2 แบบ คือ

- แอตทริบิวต์ที่เป็นลาเบลแสดงด้วยคอลัมน์ที่แสดงตามหมายเลข 1
- แอตทริบิวต์ทั่วไปแสดงด้วยคอลัมน์ที่แสดงตามหมายเลข 3

4. แสดงค่าสรุปทางสถิติของแอตทริบิวต์ต่างๆ เมื่อคลิกที่ไอคอนนี้แล้วหน้าจอจะเปลี่ยนไปซึ่งแสดงค่าทางสถิติของข้อมูลที่อยู่ในแต่ละแอตทริบิวต์โดยจะแสดงชื่อประเภทของข้อมูลที่เก็บอยู่กราฟแสดงค่าความถี่ของค่าข้อมูลในแต่ละแอตทริบิวต์



รูปที่ ก.12 ค่าทางสถิติของแต่ละแอตทริบิวต์ในหน้าจอแสดงผลลัพธ์การทำงาน



รูปที่ ก.13 โมเดล Decision Tree ในหน้าจอแสดงผลลัพธ์การทำงาน

ที่แท็บ Tree จะโมเดลของ Decision Tree ที่สร้างได้จะปรากฏขึ้นมาโดยในแท็บนี้ส่วนสำคัญที่ขออธิบาย 3 ส่วนดังนี้

- ในโมเดล Decision Tree จะมีโหนดต่างๆ อยู่ 2 ประเภท คือ

  โหนดที่เป็นแอตทริบิวต์แสดงด้วยรูปสี่เหลี่ยมที่มีมุมโค้ง

  โหนดลาเบลแสดงด้วยรูปสี่เหลี่ยมที่มีกราฟแท่งแสดงสีต่างๆ อยู่ด้วย ในตัวอย่างนี้มี 2 ลาเบล คือ Yes และ No ถ้าโมเดลตอบว่าเป็น Yes จะมีกราฟสีแสดงตามหมายเลข 1 ปรากฏอยู่และ No จะมีกราฟสีน้ำเงินแสดงตามหมายเลข 2 ปรากฏอยู่ด้วย

- ส่วนของ Mode จะใช้สำหรับปรับโหมดของการใช้งานเมาส์ซึ่งมี 2 โหมด คือ

Transform Mode โดยโหมดนี้เป็นการใช้เมาส์ในการเลื่อนตำแหน่งของ Decision Tree ทั้งต้น

Picking Mode โดยโหมดนี้เป็นการใช้เมาส์เพื่อทำการลากโหนดที่ต้องการเพื่อขยายให้ Decision Tree ดูได้ง่ายขึ้น