การใช้เทคนิคดาต้าไมน์นิ่งในการวิเคราะห์พฤติกรรมลูกค้าในธุรกิจเทปกาว

USING DATA MINING TECHNIQUES IN ORDER TO ANALYZE CUSTOMER BEHAVIOR IN ADHESIVE TAPE

โดย

ศุภกฤต เอี่ยมมีลาภ

SUPAKRIT IAMMEELARP

อาจารย์ที่ปรึกษา

ดร.สุภกิจ นุตยะสกุล

รายงานนี้เป็นส่วนหนึ่งของวิชาการศึกษาอิสระ 2
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2556

USING DATA MINING TECHNIQUES IN ORDER TO ANALYZE CUSTOMER BEHAVIOR IN ADHESIVE TAPE

SUPAKRIT IAMMEELARP

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT OF COURSE
INDEPENDENT STUDY 2
MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATION TECNOLOGY
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2/2013

ใบรับรองโครงการ
การศึกษาอิสระ 2 (Independent Study 2)

เรื่อง

การใช้เทคนิคดาต้าไมน์นิ่งในการวิเคราะห์พฤติกรรมลูกค้าในธุรกิจเทปกาว
Using Data Mining Techniques in Order to Analyze Customer Behavior in Adhesive Tape

ว่าที่ ร.ต.ศุภกฤต เอี่ยมมีลาภ
รหัสประจำตัว 54660704

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ
การศึกษาวิชาโครงการศึกษาอิสระ2 หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)
ภาคเรียนที่ 2 ปีการศึกษา 2556

..... อาจารย์ที่ปรึกษา
(ดร.สุภกิจ นุตยะสกุล)

..... กรรมการสอบ
(ดร.มานพ พันธ์โคกกรวด)

..... กรรมการสอบ
(ดร.สุภวรรณ อันนันหนับ)

ชื่อหัวข้อ	การใช้ดาต้าไมน์นิ่งเพื่อการวิเคราะห์พฤติกรรมลูกค้าในธุรกิจเทปกาว
นักศึกษา	ว่าที่ ร.ต.ศุภกฤต เอี่ยมมีลาภ
Student ID	54660704
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	การจัดการเทคโนโลยีสารสนเทศ
ปีการศึกษา	2556
อาจารย์ที่ปรึกษา	ดร.สุภกิจ นุตยะสกุล

บทคัดย่อ

ในปัจจุบันการดำเนินธุรกิจมีการแข่งขันกันอย่างสูง ส่งผลให้ผู้ประกอบการและองค์กรธุรกิจต่างมุ่งหากลยุทธ์ใหม่ๆเพื่อเพิ่มศักยภาพองค์กรของตนเองให้สามารถแข่งขันกับองค์กรอื่นๆได้ ซึ่งในโครงการศึกษาฉบับนี้ได้นำเสนอเกี่ยวกับการนำเทคนิคของดาต้าไมน์นิ่งมาประยุกต์ใช้เพื่อวิเคราะห์พฤติกรรมลูกค้าในธุรกิจเทปกาวโดยใช้โปรแกรมสำเร็จรูปทางด้านดาต้าไมน์นิ่งที่มีชื่อว่า WEKA ซึ่งเป็นซอฟท์แวร์ไม่มีลิขสิทธิ์ที่เขียนด้วยภาษา JAVA มาเป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูลเกี่ยวกับพฤติกรรมการของลูกค้า

โดยในงานวิจัยนี้ นำโมเดลมาช่วยในการหาคำตอบ ประกอบด้วย 3 โมเดล 6 อัลกอริทึม คือ หาพฤติกรรมของลูกค้าในการซื้อสินค้า โดยใช้เทคนิค apiori เปรียบเทียบกับ FP Growthโมเดลที่ 2 คือ การหาความสัมพันธ์ของข้อมูล คือการซื้อสินค้า กับ ภูมิประเทศ โดยใช้เทคนิค K mean เปรียบเทียบกับ Hierachcal เทคนิคสุดท้ายคือ การทำนายยอดซื้อสินค้าล่วงหน้า โดยใช้เทคนิค Neuron Network เปรียบเทียบกับ Linear Regression

จากผลลัพธ์ที่ได้ในการวิเคราะห์นี้มาช่วยในการตัดสินใจหรือเป็นเครื่องมือในการตัดสินใจซึ่งอยู่ในรูปแบบของโมเดลที่สามารถนำไปใช้ประโยชน์ เพื่อเป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด ซึ่งจะทำให้องค์กรสามารถสร้างรายได้ที่เพิ่มขึ้น ตลอดจนสามารถเข้าใจถึงพฤติกรรมของลูกค้าในธุรกิจเทปกาวได้อย่างดียิ่งขึ้น

Title	Using Data Mining Techniques in Order to Analyze Customer Behavior in Adhesive Tape
Student	Acting Sub. Supakrit Iammeelarp
Student ID	54660704
Degree	Master of Science
Program	Information Technology Management
Major Year	2013
Advisor	Dr.Supakit Nootyaskool

ABSTRACE

Currently, business operation has a high competition result to entrepreneur and business organization seeking new strategies in order to be more self-potential to an advantage of competition. This project presents techniques of data mining to analysis behavior of customer in adhesive business by running on WEKA software. The code of the analysis tools is written by JAVA language to analysis data behavior of customer by finding answer of three models and six algorithms. First, the model finding purchase behavior of customer by Apiori technique compared with FP Growth. Second, the model seeking relation of data in purchase product and geography by K mean technique compare the result with Hierarchical technique. Finally, the model creating prediction summarizes of advance purchase by Neuron Network technique, and the result compared with Linear Regression. The experiment result of the analysis models helps and being a decision tool for customer. Moreover, the analysis models can give information to suggest the customer-related marketing strategy and also giving how to increase income as well as better understanding behavior of customer in business of adhesive.

กิตติกรรมประกาศ

โครงการศึกษากรณีพิเศษครั้งนี้สำเร็จลงได้ ข้าพเจ้าต้องกราบขอบพระคุณ ดร.สุภกิจ นุตยะสกุล ซึ่งเป็นอาจารย์ที่ปรึกษาโครงการเป็นอย่างสูง ที่ได้ให้ความอนุเคราะห์ต่อข้าพเจ้าเกี่ยวกับความรู้และคำนำที่ดีและเป็นประโยชน์ต่อการทำโครงการ จนทำให้โครงการศึกษาในครั้งนี้สำเร็จลุล่วงไปด้วยดี

ขอกราบขอบพระคุณคณาจารย์ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า ทำให้ข้าพเจ้าสามารถนำเอาความรู้ที่ได้จากการศึกษามาประยุกต์ใช้ในโครงการศึกษานี้

ขอขอบพระคุณบริษัท แคบริค (ไทยแลนด์) จำกัด ที่ให้การสนับสนุนในส่วนของข้อมูล ที่ใช้เป็นตัวอย่างในการศึกษาของโครงการนี้

ขอขอบคุณเพื่อนๆพี่ๆ และน้องๆ แขนงวิชาการจัดการเทคโนโลยีสารสนเทศ (ITM) คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง ทุกๆ คนที่คอยให้คำแนะนำและความช่วยเหลือต่างๆ แก่ข้าพเจ้า

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา ของข้าพเจ้า ผู้ซึ่งเป็นกำลังใจที่ดีที่สุดที่ทำให้ข้าพเจ้ามีวันนี้ด้วยความภาคภูมิใจ

คุณค่าและประโยชน์อันพึงมาจากรายงานฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

ศุภกฤต เอี่ยมมีลาภ

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่ 1

บทนำ

ในบทนี้จะกล่าวถึงความเป็นมาของปัญหาของโครงการที่จะทำการศึกษา โดยการนำเอาเทคนิคของดาต้าไมน์นิ่ง เข้ามาช่วยในการค้นหาสารสนเทศที่เป็นประโยชน์ต่อองค์กร ซึ่งซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ ,วัตถุประสงค์ของโครงการการศึกษา, ขอบเขตของการศึกษา และขั้นตอนการดำเนินงานของโครงการ รวมไปถึงประโยชน์ที่จะได้รับจากการพัฒนาโครงการนี้

1.1 ความเป็นมาและความสำคัญของปัญหา

การดำเนินธุรกิจในปัจจุบันไม่ว่าจะเป็นธุรกิจ หรืออุตสาหกรรมใดก็ตาม ต่างก็มีการแข่งขันกันอย่างสูง ทั้งในด้านสินค้าและบริการ ทำให้นักการตลาดและผู้บริหารของแต่ละธุรกิจจำเป็นต้องกำหนดนโยบายและแผนกกลยุทธ์ของตนให้มีประสิทธิภาพมากที่สุด โดยอาศัยทรัพยากรที่สำคัญที่องค์กรมีอยู่ นั่นก็คือ ฐานข้อมูลลูกค้า

ธุรกิจผลิตเทปกาวก็นับเป็นธุรกิจที่มีการแข่งขันกันสูง โดยผู้ผลิตแต่ละรายต่างแข่งขันกันสร้างสรรค์สินค้า และบริการให้มีความแตกต่างจากคู่แข่ง รวมถึงราคาของสินค้าและช่องทางการจัดจำหน่ายสินค้าไปยังลูกค้า เพื่อที่จะสามารถตอบสนองและเข้าถึงความต้องการของลูกค้าให้ได้มากที่สุด ตรงกลุ่มเป้าหมายที่สุด และอยู่ในช่วงเวลาและโอกาสที่เหมาะสมที่สุด จึงจะเป็นผู้ครองส่วนแบ่งทางการตลาดในธุรกิจนั้น ยิ่งไปกว่านั้นการใช้ข้อมูลที่มีอยู่ให้เกิดประโยชน์สูงสุด ยังเท่ากับเป็นการลดต้นทุนทางการแข่งขัน และสามารถสร้างผลกำไรให้กับธุรกิจได้อีกด้วย

โดยในโครงการศึกษาฉบับนี้ ได้นำเสนอแนวทางการวิเคราะห์ข้อมูลในเชิงธุรกิจ โดยใช้วิธีการสืบค้นข้อมูลจากฐานข้อมูล (Knowledge Discovery in Database: KDD) หรือที่เรียกกันว่าการขุดค้นข้อมูล (Data Mining) เพื่อใช้วิเคราะห์พฤติกรรมของลูกค้า โดยนำผลลัพธ์ที่ได้จากการใช้เทคนิคดาต้าไมน์นิ่ง ซึ่งผลลัพธ์ที่ได้จะเป็นสารสนเทศที่มีประโยชน์ ที่จะสามารถนำไปใช้เป็นแนวทางในการกำหนดกลยุทธ์ทางการตลาด เพื่อตอบสนองความต้องการของลูกค้า ทำให้ลูกค้าเกิดความพึงพอใจสูงสุด และยังเป็นเครื่องมือหรือแนวทางที่ช่วยให้ผู้บริหารนำมาในการตัดสินใจทางธุรกิจ ซึ่งอาจให้ผลประโยชน์ในด้านต่างๆ เช่น ทำให้รักษาฐานลูกค้า (Customer Retention) ให้อยู่กับองค์กรอย่างมีประสิทธิภาพมากยิ่งขึ้น ทำให้บริษัทสามารถเพิ่มประสิทธิภาพการขาย โดย

สามารถเข้าใจถึงความต้องการของลูกค้าแต่ละกลุ่มได้อย่างเหมาะสม รวมถึงยังสามารถลดต้นทุนค่าใช้จ่ายในด้านต่างอีกด้วย

1.2 วัตถุประสงค์ของการศึกษา

วิเคราะห์พฤติกรรมลูกค้าในธุรกิจเทปกาวมีวัตถุประสงค์ในการศึกษา ดังต่อไปนี้

1. นำเทคนิคและวิธีการทำเหมืองข้อมูล (Data Mining) ไปใช้กับธุรกิจการเทปกาวในการวิเคราะห์พฤติกรรมลูกค้าได้
2. สามารถเข้าใจเทคนิคในการหากฎความสัมพันธ์ (Association Rule)ของข้อมูลเพื่อตอบโจทย์การหาความสัมพันธ์ของการซื้อสินค้า
3. สามารถเข้าใจเทคนิคในการแบ่งกลุ่มของข้อมูล (Clustering)เพื่อตอบโจทย์การหาความสัมพันธ์ของการกลุ่มข้อมูลการซื้อสินค้าและข้อมูลลูกค้าได้
4. สามารถเข้าใจเทคนิคในการทำนายข้อมูล (Prediction) เพื่อตอบโจทย์จำนวนสินค้าที่มีแนวโน้มการสั่งซื้อล่วงหน้าได้

1.3 ขอบเขตการศึกษา

1. ข้อมูลที่ใช้ในการศึกษาจะเป็นข้อมูลที่ได้จากบริษัทแคบริค (ประเทศไทย) จำกัด ซึ่งเป็นข้อมูลที่อยู่ในช่วงระยะเวลา 12 เดือน ตั้งแต่ 1 พฤษภาคม พ.ศ.2554 ถึง 31 พฤษภาคม พ.ศ.2555 เท่านั้นที่จะใช้ในโครงการศึกษา
2. ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องในกระบวนการทำดาต้าไมน์นิ่งและกระบวนการการตัดสินใจ
3. วิเคราะห์โจทย์ในกรณีศึกษาและหารูปแบบและแนวทางการหาคำตอบจากเทคนิคดาต้าไมน์นิ่งในแบบต่างๆ
4. สามารถเตรียมข้อมูลและจัดการข้อมูลไปสู่กระบวนการทำดาต้าไมน์นิ่งได้อย่างถูกต้องและเหมาะสม
5. วิเคราะห์ข้อมูลต่างๆด้วยซอฟต์แวร์ในการทำดาต้าไมน์นิ่ง
6. นำผลลัพธ์ที่ได้มาแปลงผล และแสดงผลลัพธ์ที่ได้จาการใช้เทคนิคดาต้าไมน์นิ่งในการวิเคราะห์ข้อมูล

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1. กำหนดความมุ่งหมาย อธิบายความเป็นมา กำหนดวัตถุประสงค์ในการศึกษา กำหนด

ขอบเขตการศึกษา และ ขั้นตอนการศึกษาและดำเนินงาน รวมถึง ประโยชน์ที่คาดว่าจะได้รับจากกรณีศึกษาในครั้งนี้ได้อย่างชัดเจน

2. ศึกษาแนวคิดและทฤษฎีพื้นฐานที่ใช้ในการศึกษา และที่เกี่ยวข้อง ทางด้านดาต้าไมน์นิ่ง

3. ทำการรวบรวมข้อมูลที่เกี่ยวข้องในกรณีศึกษา และทำการจัดการข้อมูลในกระบวนการทำดาต้าไมน์นิ่ง รวมถึงกำหนดหัวข้อวัตถุประสงค์ในการวิเคราะห์

4. นำข้อมูลที่ได้ผ่านกระบวนการดาต้าไมน์นิ่งเข้าสู่ซอฟต์แวร์ด้านดาต้าไมน์นิ่ง และทำการวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้าไมน์นิ่ง

5. สรุปผลการศึกษาของโครงการและข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการที่ได้ศึกษาแนวคิดและทฤษฎีของดาต้าไมน์นิ่ง เพื่อวิเคราะห์พฤติกรรมของลูกค้าในธุรกิจเทปกาว คาดว่าจะได้รับประโยชน์ ดังนี้

1. ทำให้มีความรู้ความเข้าใจถึงหลักการ และกระบวนการทำงานของดาต้าไมน์นิ่ง

2. สามารถทำการ จัดเตรียม คัดเลือก และแปลงข้อมูล ให้อยู่ในรูปแบบที่เหมาะสมในการทำดาต้าไมน์นิ่ง

3. สามารถเลือก เทคนิค วิธีการ อัลกอริทึม โมเดล และการแสดงผล รวมถึงการแปลงผลลัพธ์ในการทำดาต้าไมน์นิ่งได้อย่างเหมาะสมและถูกต้อง

4. สามารถนำข้อมูลที่มีอยู่ในองค์กร มาใช้ประโยชน์ได้อย่างมีประสิทธิภาพสูงสุด

5. สามารถนำความรู้ที่ได้ไปประยุกต์ใช้กับปัญหาจริงทางธุรกิจขององค์กรได้ทำให้สามารถทำนายลูกค้าที่มีแนวโน้มในการซื้อเทปกาวในอนาคตได้อย่างถูกต้องแม่นยำมากขึ้น และสามารถนำสารสนเทศที่ได้ไปปรับปรุงกลยุทธ์ทางการตลาด เพื่อตอบสนองความต้องการของลูกค้าทำให้ลูกค้าเกิดความพึงพอใจสูงสุด ทำให้สามารถรักษาฐานลูกค้าให้ยังคงอยู่กับองค์กรต่อไป

บทที่ 2
ทฤษฎีและหลักการของดาต้าไมน์นิ่ง

บทนำ

บริษัท แคบริค (ไทยแลนด์) จำกัด เป็นบริษัทที่ผลิตและจัดจำหน่ายเทปกาวทุกชนิด ได้ก่อตั้งบริษัทมาตั้งแต่ปี พ.ศ.2548 ซึ่งได้เริ่มก่อตั้งเป็น ห้างหุ้นส่วนจำกัด ต่อมาบริษัทมีการเติบโตอย่างรวดเร็ว จึงได้มีการเปลี่ยนเป็นบริษัทแคบริค (ไทยแลนด์)จำกัด โดยบริษัทได้ก่อตั้งเริ่มแรก ได้มีเพียงแต่การจัดจำหน่ายสินค้าเทปกาวเท่านั้น ต่อมาทางผู้บริหารเล็งเห็นช่องทางธุรกิจ รวมถึงธุรกิจมีการขยายตัวอย่างรวดเร็ว จึงได้มีการพัฒนาและเปลี่ยนรูปแบบเป็นทั้งผู้ผลิตและจัดจำหน่ายให้กับลูกค้า เช่น ห้างสรรพสินค้า โรงงานอุตสาหกรรม ร้านเครื่องเขียน รวมถึงลูกค้ารายย่อยต่างๆ ทั่วประเทศ โดยปัจจุบันทางบริษัทมีลูกค้ามากกว่า 3000 บริษัท โดยสินค้าและวัตถุดิบส่วนประกอบได้มีการจัดซื้อจากต่างประเทศ เพื่อนำเข้ามาผลิตเป็นบางส่วน และช่องทางการจัดจำหน่ายที่มีการประชาสัมพันธ์ผ่านสื่อออนไลน์ รวมถึงการให้พนักงานขายของบริษัทเข้าถึงตัวลูกค้า โดยปัญหาหลักที่พบ ได้แก่ การผลิตสินค้าที่ไม่เพียงพอต่อความต้องการของลูกค้า การขนส่งที่ยังไม่สอดคล้องต่อการผลิต และการแข่งขันกับผู้ผลิตและจัดจำหน่ายที่เพิ่มมากขึ้น เป็นต้น ซึ่งจากที่กล่าวมานั้น ทางบริษัทจึงจำเป็นที่ต้องมีเครื่องมือที่ช่วยในการรักษาฐานข้อมูลค้าเพื่อใช้เป็นแนวทางในการได้เปรียบทางธุรกิจนี้

2.1 ความหมายและหลักการของดาต้าไมน์นิ่ง

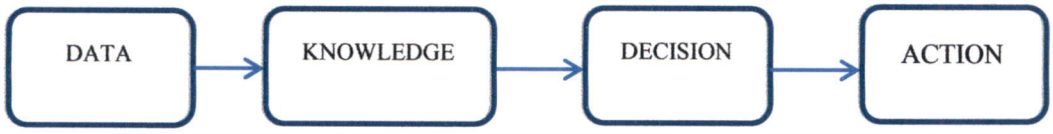
ในอดีตการจะค้นหาข้อมูลที่มีประโยชน์จากฐานข้อมูลนั้นเป็นเรื่องยาก ยิ่งถ้าหากเป็นฐานข้อมูลที่มีขนาดใหญ่มากๆ ก็จะยิ่งต้องใช้เวลาในการค้นหานานมาก จึงทำให้นักพัฒนาระบบต่างคิดค้นวิธีการที่จะทำให้สามารถค้นหาข้อมูลสารสนเทศที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ตลอดจนความสัมพันธ์กันของปัจจัยต่างๆ เพื่อนำมาใช้ประโยชน์ในการวิเคราะห์ การพยากรณ์ที่แม่นยำถูกต้อง ซึ่งสามารถใช้ประโยชน์ในการกำหนดแนวทางหรือแผนในการปฏิบัติงานขององค์กรนั้นให้มีประสิทธิภาพมากที่สุด

ดังนั้น การที่เราจะค้นหาข้อมูลที่เป็นสารสนเทศที่เราต้องการจากแหล่งข้อมูลดิบที่มีมากมายมหาศาลนั้น เราจำเป็นต้องมีเครื่องมือที่จะช่วยในการค้นหาสารสนเทศเหล่านั้น ซึ่งหนึ่งในนั้นก็คือ เทคนิคของดาต้าไมน์นิ่ง

การทำเหมืองข้อมูลหรือ ดาต้าไมน์นิ่ง คือ ชุดซอฟท์แวร์วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนการตัดสินใจของผู้ใช้ ถือว่าเป็นซอฟท์แวร์ที่มีประสิทธิภาพสูงทั้งในเรื่องการค้นหา การทำรายงาน และโปรแกรมรวบรวมข้อมูล และระบบข้อมูลช่วยในการตัดสินใจในการ

บริหารงาน ซึ่งเป็นเครื่องมือที่ช่วยในการค้นหาข้อมูลขนาดใหญ่หรือข้อมูลที่มีประโยชน์ เพื่อเพิ่มคุณค่าให้กับข้อมูลที่มีอยู่และรวมไปถึงเพื่อเป็นตัวช่วยในการตัดสินใจให้กับผู้บริหารได้ด้วย

ระบบการสนับสนุนการตัดสินใจ (Decision Support System) คือการทำอย่างไรให้ข้อมูลที่มีอยู่กลายเป็นข้อมูลที่มีคุณค่า มีความรู้ และสามารถสร้างผลตอบทนให้กับผู้ใช้งานได้



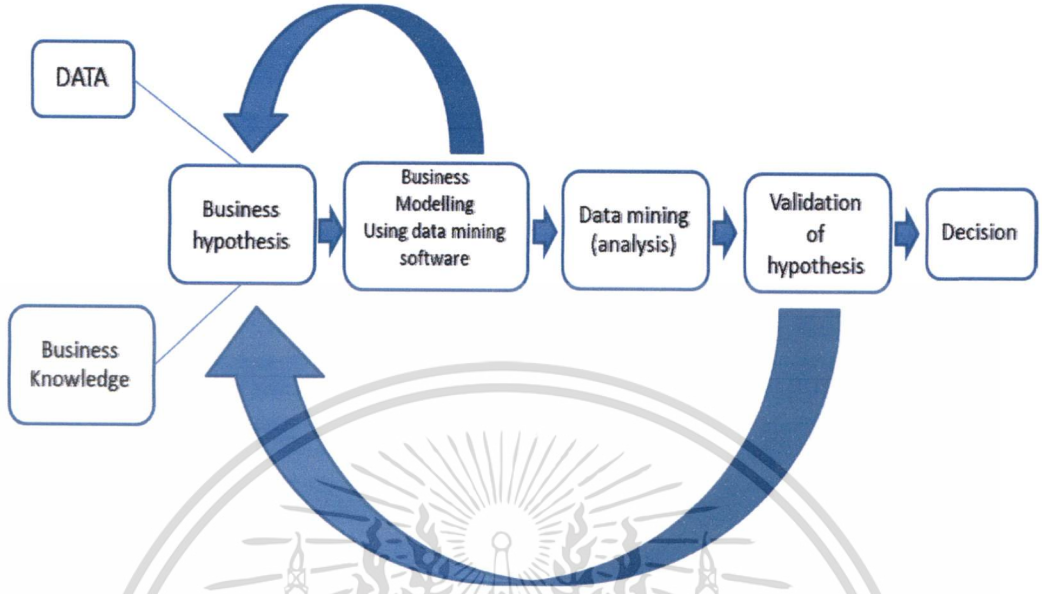
รูปที่ 2.1 แสดงลำดับการประมวลข้อมูลการตัดสินใจและการปฏิบัติ

จากรูปภาพ 2.1 ระบบสนับสนุนการตัดสินใจ (Decision Support System) ประกอบด้วย 4 ขั้นตอน ดังต่อไปนี้ คือ DATA คือข้อมูลที่ผู้ใช้งานมีอยู่ที่จะนำมาประกอบในการตัดสินใจ ลำดับถัดมา คือ KNOWLEDGE คือ ขั้นตอนประมวลความรู้ คือผู้ใช้งานจะต้องมีความรู้ความเข้าใจ ในการที่จะนำข้อมูลอะไรใช้ในการตัดสินใจ หรือ เป็นการรวบรวมความรู้ของผู้ใช้งานและกำหนดวัตถุประสงค์ให้ชัดเจนในการที่จะนำข้อมูลมาสนับสนุนการตัดสินใจ ขั้นตอนถัดมาคือ DECISION เป็นขั้นตอนการตัดสินใจ ขั้นตอนนี้เป็นขั้นตอนที่ผู้ใช้งานจะต้องตัดสินใจว่าข้อมูลที่ได้มา มีความถูกต้อง เหมาะสม และสามารถสนับสนุนการตัดสินใจให้กับผู้ใช้งานได้หรือไม่ ขั้นตอนสุดท้าย คือ ACTION ขั้นตอนการปฏิบัติ หรือ ขั้นตอนการนำไปใช้งานว่า ข้อมูลที่สนับสนุนการตัดสินใจที่ได้มา เรานำมาใช้งานและสามารถตอบโจทย์ที่ตั้งไว้ได้หรือไม่

การนำเหมืองข้อมูลมาใช้สนับสนุนกระบวนเมื่อพิจารณาจากลำดับขั้นตอนนั้น ตามรูปภาพ 2.1แสดงลำดับการประมวลข้อมูลสู่การตัดสินใจและปฏิบัติ จะพบว่า การทำเหมืองข้อมูลนั้นถูกนำไปใช้ในกระบวนการตัดสินใจประกอบขั้น Knowledge ซึ่งขั้นตอนนี้ จะแสดงรายละเอียดดังต่อไปนี้ ตามรูป 2.2

1. ในบริษัทขนาดกลางถึงขนาดเล็ก ขบวนการเหมืองข้อมูล โดยทั่วไปจะเริ่มจากการตั้งสมมุติฐานทางธุรกิจตามความรู้ความเข้าใจของผู้ใช้งาน ที่มีต่อธุรกิจ
2. ใช้ซอฟต์แวร์ดาต้าไมน์นิ่ง และเครื่องมือในการวิเคราะห์ โดยทำการสร้างโมเดล แล้วกลั่นกลองข้อมูลให้อยู่ในรูปแบบที่เหมาะสมตามความต้องการที่ต้องการศึกษาหรือหาความรู้ และตามด้วยการวิเคราะห์ข้อมูล ซึ่งขั้นตอนกระบวนการนี้อาจจะต้องมีการทำซ้ำหลายๆครั้ง หรืออาจจะต้องมีการเริ่มต้นใหม่ตั้งแต่การคัดเลือกและนำเข้าข้อมูล เพื่อให้สามารถตอบโจทย์ในการศึกษาได้อย่างถูกต้องและเหมาะสมที่สุด
3. หลังจากเมื่อทำการวิเคราะห์ข้อมูล แสดงผลลัพธ์ข้อมูลในรูปแบบตามต้องการ รวมถึง

การแปลผลลัพธ์ได้อย่างถูกต้องและตรงความต้องการในการศึกษาแล้ว ถ้าไม่มีสิ่งใดต้องนำมาแก้ไขหรือพิจารณาอีก ก็เป็นขั้นตอนสุดท้ายคือ การตัดสินใจ



รูปที่ 2.2 แสดงรายละเอียดขั้นตอนจากข้อมูลสู่การตัดสินใจ

ในปัจจุบันระบบการตัดสินใจได้เข้ามามีบทบาทในการประกอบการตัดสินใจของผู้บริหารมากขึ้น จึงจำเป็นจะต้องมีซอฟท์แวร์หรือเครื่องมือที่สามารถช่วยให้ผู้ใช้งานหรือผู้บริหารสามารถเข้าถึงข้อมูลที่มีขนาดใหญ่ และสามารถค้นหาวิธีการและรูปแบบที่จะสามารถนำไปปรับปรุงกลยุทธ์ทางธุรกิจได้อย่างมีประสิทธิภาพมากที่สุด ซึ่งนั้นก็คือ การทำดาต้าไมน์นิ่ง

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

1. เป็นข้อมูลที่ไม่เคยรู้ล่วงหน้ามาก่อน (Unknown) หมายถึง ข้อมูลสารสนเทศที่ได้รับนั้น ต้องไม่เคยค้นพบมาก่อนหน้า และไม่สามารถคาดเดาได้ว่าผลที่ได้รับจะออกมาในลักษณะใด
2. ต้องเป็นข้อมูลที่มีความถูกต้อง (Valid) หมายถึง สารสนเทศที่ได้รับต้องเป็นสารสนเทศที่มีความถูกต้อง เนื่องจากต้องนำไปใช้ประกอบกับข้อมูลส่วนอื่นๆ ดังนั้นต้องมีความถูกต้อง น่าเชื่อถือ
3. สามารถนำไปใช้ประโยชน์ได้ (Actionable) คือ ต้องสามารถนำเอาข้อมูลและ

สารสนเทศที่ค้นพบออกมา ไปใช้ประโยชน์ในด้านอื่นๆ ได้ เช่น นำมาช่วยตัดสินใจในการวางแผนการตลาด เพื่อสร้างความได้เปรียบทางการแข่งขันในเชิงธุรกิจ เป็นต้น

ดังนั้นการทำดาต้าไมน์นิ่งจึงเปรียบเสมือนการขุดหาแร่จากเหมืองแร่ที่มีขนาดใหญ่ กว่าที่จะได้แร่ที่มีค่าอย่างที่ต้องการนั้นต้องผ่านกระบวนการมากมายหลายขั้นตอน ในการขุดค้นกลั่นกรองเพื่อที่จะได้แร่ที่มีค่าออกมา นั่นจึงเป็นที่มาของคำว่า ดาต้าไมน์นิ่ง หรือการทำเหมืองข้อมูล

2.1.1 วิวัฒนาการของเทคโนโลยีฐานข้อมูล

วิวัฒนาการของเทคโนโลยีด้านฐานข้อมูลนั้น ได้มีการพัฒนามาทุกยุคทุกสมัย ตั้งแต่ในอดีตจนถึงปัจจุบัน ซึ่งเป็นเทคโนโลยีที่มีความสำคัญมาก เนื่องจากข้อมูลเป็นสิ่งสำคัญในการนำมาใช้ประโยชน์ในด้านต่างๆ อีกทั้งแนวโน้มของการเพิ่มขึ้นของข้อมูล ก็มีแนวโน้มที่เพิ่มขึ้นสูงมาก จึงได้มีการพัฒนาและปรับปรุงวิธีการต่างๆ เพื่อที่จะสามารถเก็บรวบรวมและประมวลผลข้อมูลที่มีอยู่อย่างมหาศาลได้อย่างมีประสิทธิภาพ โดยสามารถสรุปวิวัฒนาการของการพัฒนาเทคโนโลยีด้านฐานข้อมูลได้เป็นช่วงเวลา ดังนี้

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

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

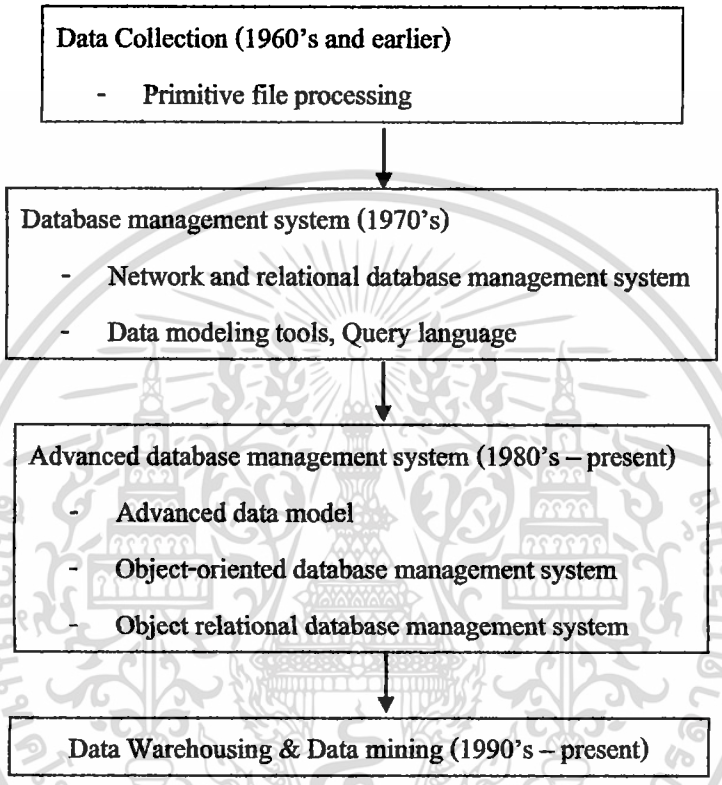
ช่วงปี ค.ศ. 1980 เทคโนโลยีฐานข้อมูลได้เริ่มมีการปรับปรุงและพัฒนาระบบจัดการฐานข้อมูลที่มีศักยภาพมากขึ้น ทำให้สามารถจัดเก็บข้อมูลจำนวนมากที่มีความซับซ้อนได้อย่างมีประสิทธิภาพเพิ่มขึ้น เกิดระบบการจัดการฐานข้อมูลที่มีประสิทธิภาพ เช่น Object-Oriented Database Management System , Object Relational Database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

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

การละเลยข้อมูลควบคู่ไปกับการขาดเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลที่มีศักยภาพนำไปสู่สถานการณ์ที่ว่า "ข้อมูลมากแต่ความรู้น้อย" (data rich but information poor) การเติบโตขึ้น

อย่างรวดเร็วของข้อมูลจำนวนมาก ที่สะสมไว้ในฐานข้อมูลขนาดใหญ่มาก ซึ่งเกินกว่าที่กำลังคนจะสามารถจัดการได้ เป็นผลทำให้มีความจำเป็นที่ต้องมีเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลและหาความเป็นไปได้ของข้อมูลที่เป็นประโยชน์ที่อยู่ในฐานข้อมูลออกมา ซึ่งก็คือ ดาต้าไมน์นิ่ง

จากที่ได้กล่าวถึงประวัติความเป็นมาและวิวัฒนาการของเทคโนโลยีฐานข้อมูลตั้งแต่ในยุคอดีตจนถึงปัจจุบัน จะสามารถแสดงได้ดังรูปที่ 2.3



รูปที่ 2.3 แสดงวิวัฒนาการเทคโนโลยีฐานข้อมูล

2.1.2 ปัจจัยที่ทำให้ดาต้าไมน์นิ่งเป็นที่ได้รับความนิยม

ปัจจัยที่ทำให้ดาต้าไมน์นิ่งเป็นที่ได้รับความนิยม คือ

1. จำนวนและขนาดข้อมูลขนาดใหญ่ถูกผลิต และขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทางอินเทอร์เน็ต , ดาวเทียม และแหล่งผลิตข้อมูล อื่น ๆ เช่น เครื่องอ่านบาร์โค้ด , เครดิตการ์ด , อีคอมเมิร์ซ
2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ (Decision Support System) เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจ

ส่วนมากข้อมูลจะถูกจัดเก็บแยกมาจาก ระบบปฏิบัติการ (Operational System) โดยจัดอยู่ในรูปของคลังหรือเหมืองข้อมูล ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้

3. ระบบคอมพิวเตอร์สมรรถนะสูงมีราคาต่ำลง และเนื่องจากเทคนิคดาต้าไมน์นิ่งประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนและความต้องการการคำนวณสูง จึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์สมรรถนะสูง ปัจจุบันระบบคอมพิวเตอร์สมรรถนะสูงมีราคาต่ำลง พร้อมด้วยเริ่มมีเทคโนโลยีที่นำเครื่องไมโครคอมพิวเตอร์จำนวนมาก มาต่อเชื่อมกันโดยเครือข่ายความเร็วสูง ทำให้ได้ระบบคอมพิวเตอร์ สมรรถนะในราคาต่ำ

4. การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า เนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า มีการผลิตข้อมูลไว้อย่างมากมายแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์จึงเป็นการจำเป็นอย่างยิ่งที่ต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูล ความรู้ที่ได้รับสามารถนำไปวิเคราะห์เพื่อการตัดสินใจ ในการบริหารจัดการในระบบต่าง ๆ ซึ่งเห็นได้ว่าความรู้เหล่านี้ถือว่าเป็นผลิตผลอีกชิ้นหนึ่งเลยทีเดียว

2.1.3 ประเภทข้อมูลที่สามารถนำมาทำดาต้าไมน์นิ่ง

1. ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) เป็นฐานข้อมูลที่จัดเก็บในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย entity relationship (ER model)

2. คลังข้อมูล (Data Warehouses) เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันหรือคนละรูปแบบและรวบรวมไว้ในที่ๆเดียวกัน โดยจัดการโครงสร้างขึ้นมาใหม่เพื่อสะดวกต่อการเรียกใช้งานและการวิเคราะห์ข้อมูลแบบออนไลน์ On-Line Analytical Processing (OLAP)

3. ฐานข้อมูลแบบทรานเซกชัน (Transactional Database) ประกอบด้วยข้อมูลที่แต่ละทรานเซกชันแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

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

2.1.4 ลักษณะเฉพาะของข้อมูลที่สามารถทำดาต้าไมน์นิ่ง

1. ข้อมูลขนาดใหญ่ เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่าหรือโดยการใช้ Database Management System (DBMS) ในการจัดการฐานข้อมูล

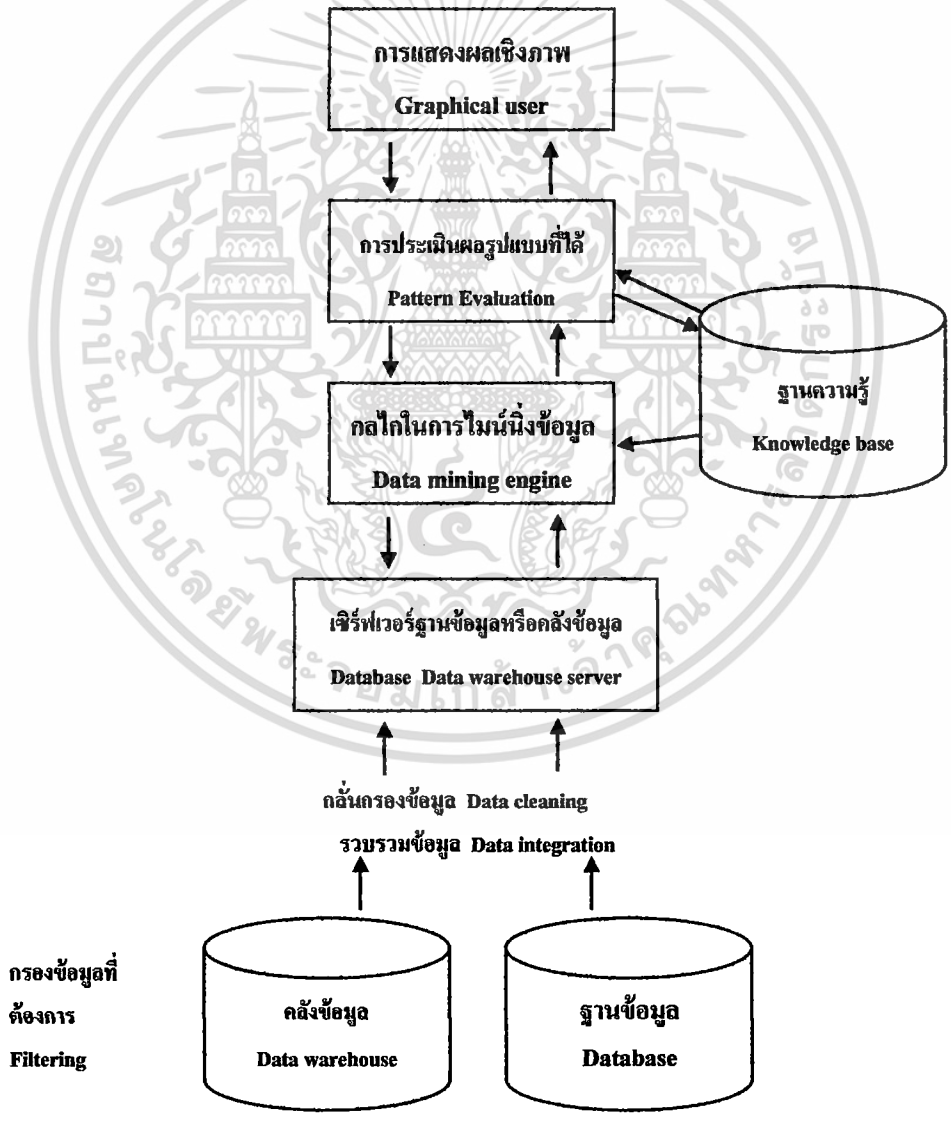
2. ข้อมูลที่มาจากหลายแหล่ง โดยอาจรวบรวมมาจากหลายระบบปฏิบัติการหรือหลาย DBMS เช่น Oracle , DB2 , MS SQL , MS Access เป็นต้น

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลง ตลอดช่วงเวลาที่ทำการไมน์นิ่งข้อมูล หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำดาต้าไมน์นิ่ง แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่ง สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำดาต้าไมน์นิ่งใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำไมน์นิ่งได้เช่นกัน แต่ต้องใช้เทคนิคการทำดาต้าไมน์นิ่งขั้นสูง

2.1.5 สถาปัตยกรรมพื้นฐานของดาต้าไมน์นิ่ง

สถาปัตยกรรมพื้นฐานของดาต้าไมน์นิ่งมีส่วนประกอบหลักดังรูปที่ 2.4 (กฤษณะ ไวยมัย. 2549)



รูปที่ 2.4 สถาปัตยกรรมพื้นฐานของดาต้าไมน์นิ่ง

1. ฐานข้อมูล คลังข้อมูล และแหล่งเก็บสะสมข้อมูลอื่นๆ หมายถึงฐานข้อมูลตั้งแต่หนึ่งฐานข้อมูลขึ้นไปหรอกลุ่มของฐานข้อมูล คลังข้อมูล สเปรดชีตหรือแหล่งที่เก็บข้อมูลในรูปแบบอื่นๆ การได้มาซึ่งข้อมูลนั้นต้องผ่านกระบวนการเตรียมข้อมูลโดยใช้เทคนิคกลั่นกรองข้อมูล และการรวบรวมข้อมูลก่อน

2. เซิร์ฟเวอร์ฐานข้อมูลหรือคลังข้อมูล จะทำหน้าที่ในการให้บริการข้อมูลที่ต้องการตามที่ได้ร้องขอในกระบวนการทำดาต้าไมน์นิ่งนั่นเอง

3. ฐานความรู้ จะเป็นที่เก็บรวบรวมความรู้ หรือรูปแบบกฎเกณฑ์ที่ได้ค้นพบ

4. กลไกในการทำดาต้าไมน์นิ่ง เป็นกระบวนการในการนำเทคนิคดาต้าไมน์นิ่งมาใช้กับข้อมูลอย่างเหมาะสม

5. การประเมินผลรูปแบบที่ได้ เป็นขั้นตอนการแปลความหมายและประเมินผลลัพธ์ที่ได้นั้นมีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งต้องใช้ทักษะการวิเคราะห์เชิงธุรกิจเข้าช่วย

6. การแสดงผลเชิงภาพ เป็นขั้นตอนการนำผลลัพธ์ที่ได้มาแสดงผลโดยทั่วไปควรมีการแสดงผลในรูปแบบที่สามารถเข้าใจได้โดยง่าย เช่น กราฟ แผนภูมิเชิงมิติ เป็นต้น

2.2 ขั้นตอนการทำดาต้าไมน์นิ่ง (Process of Data Mining)

ขั้นตอนในการทำดาต้าไมน์นิ่งหรือเรียกอีกอย่างหนึ่งว่า Knowledge Discovery in Database (KDD) มีขั้นตอนในการทำงานที่สำคัญ ดังต่อไปนี้

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

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ

เป็นตัวจักรที่สำคัญในการทำดาต้าไมน์นิ่ง เนื่องจากเป็นการกำหนดขอบเขตเป้าหมาย ของการทำดาต้าไมน์นิ่ง ซึ่งจะมีผลต่อทุกๆขั้นตอนของการทำดาต้าไมน์นิ่ง โดยนักวิเคราะห์ธุรกิจ

(Business Analyst) จะต้องระบุ ปัญหาที่เกิดขึ้นในการทำธุรกิจให้ครอบคลุมและชัดเจนรวมทั้งวัตถุประสงค์ ซึ่งขั้นตอนนี้จะสามารถมองถึงอัลกอริทึมและฐานข้อมูลที่จะใช้งานเบื้องต้น เป็นการนำไปสู่การสร้างแบบจำลองที่เหมาะสม ขึ้นอยู่กับเป้าหมายทางธุรกิจ

2.2.2 การเตรียมข้อมูล

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้ระยะเวลาประมาณ 60% ของการทำดาต้าไมน์นิ่ง นับเป็นขั้นตอนที่ใช้เวลานานที่สุด ประกอบด้วย

2.2.2.1 การคัดเลือกข้อมูล (Data Selection)

จุดมุ่งหมายของการคัดเลือกข้อมูล คือ การระบุถึงแหล่งข้อมูลที่นำมาใช้ที่จำเป็นต่อการนำมาวิเคราะห์ข้อมูลเบื้องต้น รวมถึงจะต้องมีความเข้าใจเกี่ยวกับลักษณะและตัวแปรของข้อมูลที่จะนำมาทำดาต้าไมน์นิ่งด้วย ซึ่งตัวแปรของข้อมูลแบ่งออกเป็น 2 ประเภท ดังนี้

1. ข้อมูลที่แบ่งเป็นกลุ่ม (Categorical data) มี 2 ประเภท คือ

1) Nominal คือ ข้อมูลแบบที่ไม่คำนึงถึงลำดับ หรือ ลำดับไม่มีความสำคัญ เช่น สถานภาพ (โสด,แต่งงาน,หย่าร้าง) เพศ (ชาย,หญิง) ระดับการศึกษา (มัธยมศึกษา,ปริญญาตรี, ปริญญาโท,ปริญญาเอก)

2) Ordinal คือ ข้อมูลแบบที่คำนึงถึงลำดับ หรือลำดับมีความสำคัญ เช่น การจัดระดับเครดิตของลูกค้า (ดี,ปานกลาง,แย่)

2. ข้อมูลแบบที่เป็นตัวเลข (Quantitative data) มี 2 ประเภท คือ

1) Continuous ข้อมูลที่เป็นจำนวนจริง เช่น

2) รายได้, รายจ่าย, ผลกำไร, ค่าเฉลี่ย เป็นต้น

3) Discrete ข้อมูลที่มีค่าไม่ต่อเนื่อง เป็นจำนวนเต็ม เช่น จำนวนพนักงาน , จำนวนลูกค้า เป็นต้น

ข้อพิจารณาเบื้องต้นในการคัดเลือกข้อมูล ได้แก่

- เลือกเฉพาะข้อมูลที่สนใจ
- ไม่นำคอลัมน์ที่มีค่าสำหรับทุกแถวเป็นค่าเดียวกันมาใช้
- ตัดคอลัมน์ที่มีค่าไม่ซ้ำกันเลยออก เนื่องจากข้อมูลเหล่านี้ไม่สามารถหาแถวที่มี
- ข้อมูลสัมพันธ์กันได้เลย ยกเว้นบางคอลัมน์ที่ต้องการนำไปใช้จริง แต่จะมีการจัดการกับข้อมูลนี้ในลำดับต่อไป
- คอลัมน์ที่มีค่าส่วนมากเป็นอย่างเดียวกัน จะต้องพิจารณาว่าค่าที่แตกต่างกันนั้นสำคัญหรือไม่ หากข้อมูลส่วนน้อยนั้นไม่สำคัญก็สามารถตัดคอลัมน์นั้นทิ้งไปได้

- คุณสมบัติหรือแอตทริบิวต์บางอย่างที่อยู่ในตารางข้อมูลสามารถใช้สร้างเป็นคุณสมบัติใหม่ได้ โดยอาศัยเงื่อนไขต่างๆ มากำหนด

2.2.2.2 การประมวลข้อมูลก่อน (Data Preprocessing)

จุดมุ่งหมายของขั้นตอนนี้เป็นการนำเอาข้อมูลที่จะใช้ในการทำดาต้าไมน์นิ่งมาทำให้เป็นข้อมูลที่มีคุณภาพดีก่อนที่จะนำไปใช้งานต่อไป โดยเป็นการตรวจสอบว่าข้อมูลที่ได้เลือกไว้ในขั้นตอนการคัดเลือกข้อมูลนั้น มีความเหมาะสมหรือไม่ เช่น ข้อมูลแบบ Categorical ใช้วัดการกระจายของข้อมูล เพื่อทำความเข้าใจข้อมูลได้ดียิ่งขึ้นโดยอาศัยเครื่องมือทางด้านการสร้างภาพนามธรรม (Visualize)แสดงข้อมูล เช่น กราฟแท่ง ส่วนข้อมูลแบบ Quantitative ที่เป็นตัวเลข วัดโดยการหาค่าสูงสุด ต่ำสุด ค่าเฉลี่ย ค่ามัธยฐาน และตัววัดทางสถิติอื่นๆ ซึ่งการประมวลข้อมูลก่อนนี้ ประกอบด้วย

1. การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนที่ทำให้ ข้อมูลมีความสมบูรณ์ถูกต้อง และสอดคล้องกัน เป็นการเพิ่มค่าที่ขาดหายไป (Missing Values) การระบุ Noisy Data ค่าความผิดพลาดหรือความแปรปรวน ที่เกิดขึ้นจากการเก็บรวบรวมข้อมูล การป้อนข้อมูลเข้าสู่ระบบและการรับส่งข้อมูล ความไม่สอดคล้องกันจากการตั้งชื่อแล้วจึงทำการปรับปรุงค่าข้อมูลให้มีความสอดคล้องกัน เช่น ข้อมูลในฟิลด์ที่ขาดหายไป อาจจะแทนค่าข้อมูลที่ขาดหายไปด้วย Unknown หรือถ้าหากข้อมูลขาดหายไปเป็นจำนวนมากและข้อมูลนั้นไม่สำคัญมากนักอาจจะทำการตัดฟิลด์นั้นทิ้งไป

2. การรวบรวมข้อมูล (Data Integration) เป็นขั้นตอนที่รวบรวมข้อมูลมาจากหลายๆ แหล่งแล้วทำการตรวจหา และขจัดความขัดแย้งและความซ้ำซ้อนของข้อมูล

2.2.3 การแปลงรูปแบบข้อมูล (Data Transformation)

เป็นขั้นตอนที่ทำการรวบรวมข้อมูลหรือเปลี่ยนแปลงข้อมูล เพื่อให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมที่ใช้ในการทำดาต้าไมน์นิ่งของงาน ซึ่งความเหมาะสมของข้อมูลก็ขึ้นอยู่กับโมเดลที่เราจะใช้งาน ตัวอย่างของโมเดลที่จะใช้งานไม่สามารถทำการคำนวณข้อมูลที่เป็นตัวอักษรได้ ก็จะต้องแปลงตัวอักษรไปเป็นตัวเลขก่อน เช่น ระดับการศึกษา ปริญญาตรี ปริญญาโท และปริญญาเอก ไปเป็นตัวเลข 1,2,3 เพื่อให้สอดคล้องกับโมเดลที่จะใช้งาน

2.2.4 การทำดาต้าไมน์นิ่ง

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

2.2.5 การวิเคราะห์ผลลัพธ์

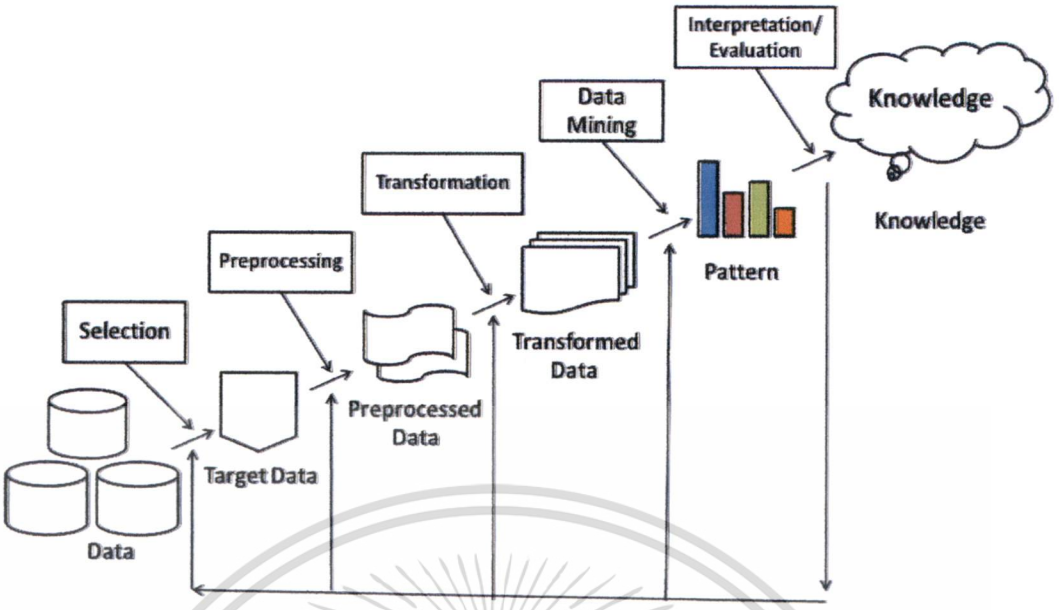
การวิเคราะห์ผลลัพธ์เป็นขั้นตอนที่ทำการวิเคราะห์ผลลัพธ์ และแปลความหมายที่ได้จากการทำดาต้าไมน์นิ่งในขั้นตอนที่ผ่านมา โดยต้องอาศัยทักษะจากประสบการณ์ ความรู้ความชำนาญในเรื่องที่เกี่ยวข้อง รวมถึงต้องเป็นไปตามดำเนินการและเทคนิคที่เลือก

2.2.6 การนำความรู้ที่ได้ไปใช้งาน

การนำความรู้ที่ได้ไปใช้งาน เป็นขั้นตอนในการเลือกและรวบรวมความรู้ที่ได้จากการวิเคราะห์ผลลัพธ์ นำไปประยุกต์ใช้กับองค์กรจริงๆ เนื่องจากผลลัพธ์ที่ได้อาจมีได้หลายรูปแบบ ซึ่งพบว่าบางผลลัพธ์อาจไม่เป็นประโยชน์กับองค์กร ทำให้ต้องมีการวัดความน่าสนใจของผลลัพธ์โดยสามารถวัดได้จาก

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

ดังนั้น จากการศึกษาถึงขั้นตอนต่างๆ ในการทำดาต้าไมน์นิ่ง ทำให้เราทราบถึงลักษณะการทำงานในแต่ละขั้นตอน ซึ่งขั้นตอนต่างๆ ของการทำดาต้าไมน์นิ่ง สามารถแสดงเป็นรูปภาพได้ดังรูปที่ 2.5



รูปที่ 2.5 แสดงขั้นตอนต่างๆของการทำดาต้าไมน์นิ่ง

2.3 โมเดลในการทำดาต้าไมน์นิ่ง

ในการทำดาต้าไมน์นิ่ง มีโมเดลหลักๆ 2 ประเภท (กรกต เกรียงพันธุ์อมร,2543) คือ

1. Predictive Model หรือ Supevised Learning

โมเดลที่สร้างขึ้นจากข้อมูลที่รู้ผลลัพธ์อยู่แล้ว เพื่อใช้เป็นโมเดลในการทำนายผลลัพธ์ของข้อมูลชุดใหม่โดยเน้นที่ความถูกต้องของโมเดลมากกว่าการค้นหาโมเดลที่น่าสนใจโมเดลประเภทนี้ได้แก่ โมเดล Classification โมเดล Regression (prediction)

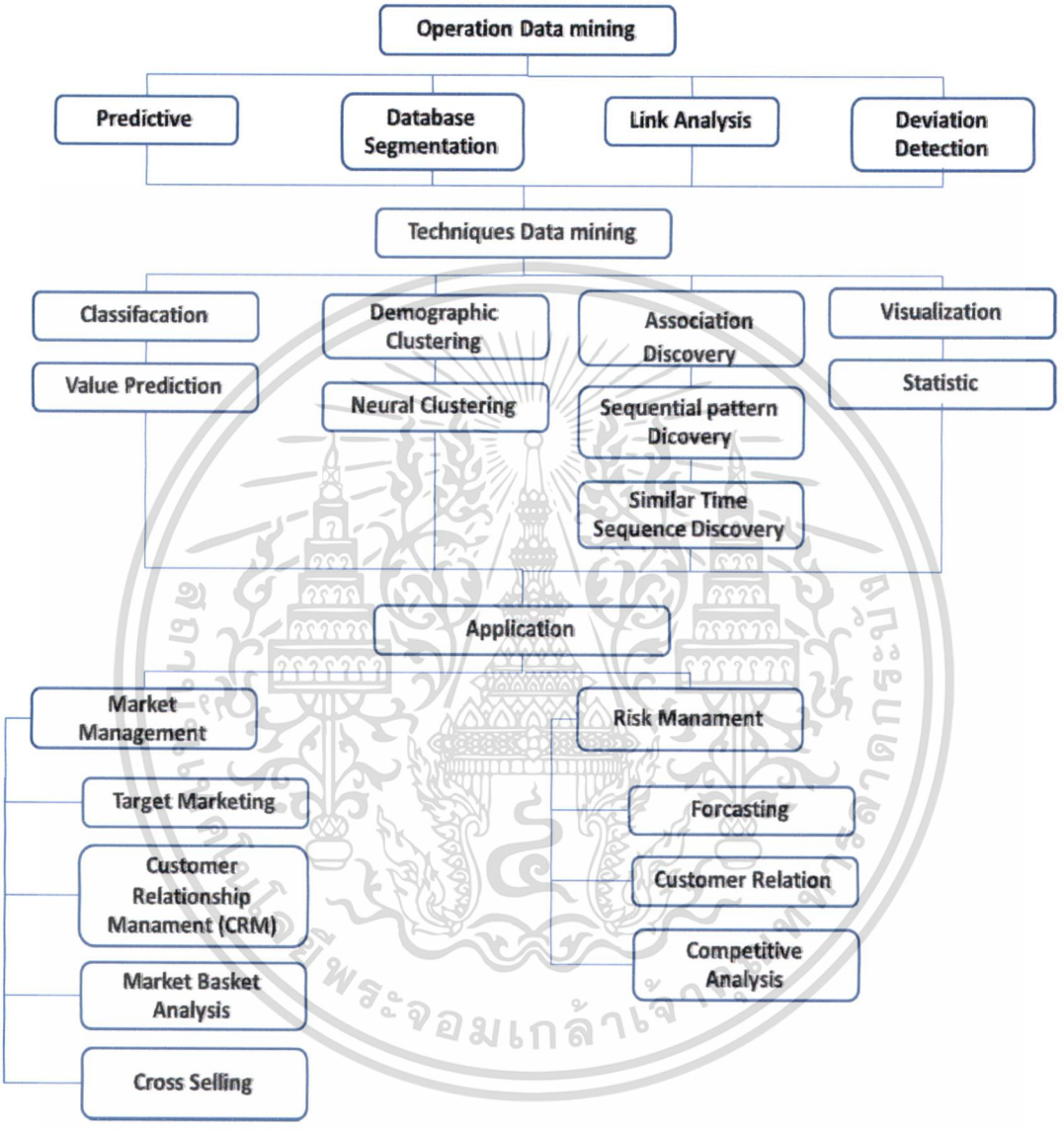
2. Descriptive Model หรือ Unsupervised Learning

โมเดลนี้เป็นโมเดลที่สร้างขึ้นจากข้อมูลที่ผู้ใช้ไม่รู้ผลลัพธ์มาก่อน ซึ่งใช้อธิบายรูปแบบความสัมพันธ์ภายในข้อมูล เพื่อช่วยในการตัดสินใจส่วนใหญ่จะมีเป้าหมายเพื่อทำความเข้าใจ หรือค้นหาความรู้ใหม่ๆที่แฝงอยู่ในข้อมูล โมเดลประเภทนี้ได้แก่ โมเดล Clustering และ โมเดล Association

2.4 เทคนิคของดาต้าไมน์นิ่ง

เทคนิคของการทำดาต้าไมน์นิ่ง เป็นขั้นตอนในการเลือกรูปแบบการประมวลผล วิธีการ

โมเดล อัลกอริทึม ให้เหมาะสมกับการใช้งาน หรือหาความรู้จากการใช้เทคนิคนั้นๆ ซึ่งบางผลลัพธ์อาจจะต้องประกอบด้วยหลายเทคนิคเข้าด้วยกัน เพื่อหาผลลัพธ์ตามที่ต้องการ โดยทั่วไปแล้ว มีเทคนิควิธีการ และผลลัพธ์ที่นำไปใช้ประโยชน์ได้หลากหลายรูปแบบ ดังนี้



รูปที่ 2.6 ประเภทของเทคนิคดาต้าไมน์นิ่งและตัวอย่างการทำงานและวัตถุประสงค์ของผลลัพธ์

2.4.1 Link Analysis

เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูลที่นิยมอีกรูปแบบหนึ่ง และสามารถนำเอาไปประยุกต์ใช้จริงกับงานต่างๆ หลักการทำงานของวิธีนี้คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลที่มีอยู่เป็นจำนวนมาก จุดมุ่งหมายของ Link Analysis คือ การสร้าง Link ที่เรียกว่า

"Association" ระหว่าง เรคอร์ดหรือกลุ่มของข้อมูล ในฐานข้อมูล เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆ หรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้า ซึ่งประเมินได้จากข้อมูลที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ "กฎความสัมพันธ์" (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

Link Analysis แบ่งออกเป็น 3 ชนิด

1.Association Discovery

2.Sequential Pattern Discovery

3.Similar time Sequential Discovery

ตัวอย่างที่นำเทคนิคนี้เอาไปใช้ ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆมักจะซื้อหนังสือใดด้วยพร้อมกันเสมอ เป็นต้น

กฎความสัมพันธ์ (Link Association) เป็นการค้นหาความสัมพันธ์ของข้อมูล ซึ่งนิยมใช้ในการหาความสัมพันธ์ของสินค้าที่เกิดขึ้นในรายการเดียวกัน ที่มีแนวโน้มว่าจะเกิดขึ้นพร้อมๆ กัน เช่น พิจารณาสินค้าที่มักจะถูกซื้อควบคู่กันไปในคราวเดียวกัน การวิเคราะห์ในลักษณะนี้เรียกว่า "Market Basket Analysis" ซึ่งจะนำไปใช้วิเคราะห์การซื้อสินค้าจากลูกค้าทำให้ผู้ประกอบการธุรกิจสามารถนำไปช่วยในการวางแผนทางการตลาด หรือกำหนดกลยุทธ์ทางด้านการจำหน่ายสินค้าและบริการได้เช่น การจัดโปรโมชั่น การวางตำแหน่งของสินค้า เป็นต้น

ผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่งด้วยเทคนิคนี้จะได้ออกมาเป็นการหากฎความสัมพันธ์ (Association rules) ซึ่งรูปแบบของกฎความสัมพันธ์ สามารถแสดงได้ดังนี้

If X Then Y

When Condition 1 then Condition 2

X =>Y

เมื่อ X หรือ Condition 1 คือกฎหรือเหตุการณ์ที่เกิดขึ้นก่อน (Antecedent)

Y หรือ Condition 2 คือ ผลที่ตามมา (Consequent)

ปัจจัยที่เกี่ยวข้องในการสร้างกฎนี้ได้แก่

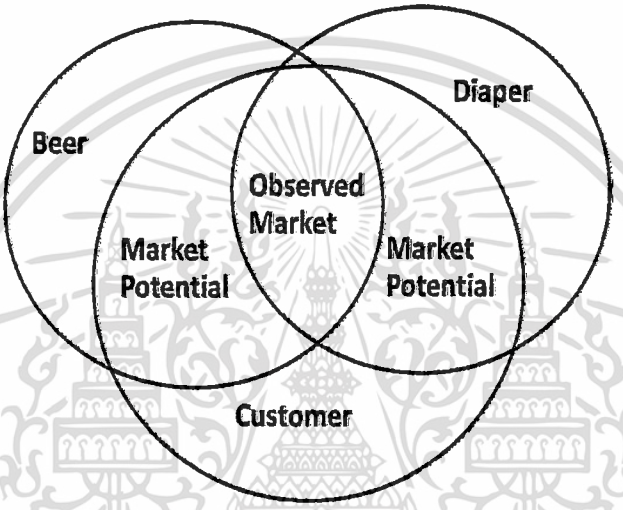
- ค่าความถี่ของเหตุการณ์ X และ Y ที่เกิดขึ้นจากจำนวนเหตุการณ์ทั้งหมด (Support Factor)

- ค่าความน่าเชื่อถือของเหตุการณ์ Y ที่จะน่าเกิดขึ้น เทียบจากเหตุการณ์ X และ Y พร้อมกัน (Confidence Factor)

ตัวอย่างกฎความสัมพันธ์

Buys (X,"diapers") => buys (X,"beers") [2%,50%]

แปลความหมายได้ว่า มีลูกค้าที่นิยมซื้อผ้าอ้อมและเบียร์ไปพร้อมกันด้วยความถี่ 2% ของรายการที่ซื้อสินค้าทั้งหมดและมีรายการที่ซื้อผ้าอ้อมแล้วซื้อเบียร์ไปด้วย ค่าความน่าเชื่อถือ 50% ด้วยกัน การขายสินค้าประเภทผ้าอ้อมและเบียร์ไปพร้อมๆกัน แสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 ผลการขายสินค้าประเภทผ้าอ้อมและเบียร์

โดยส่วนใหญ่ กฎความสัมพันธ์ที่น่าสนใจ คือกฎที่มีค่าความน่าเชื่อถือที่สูง เนื่องจากมีโอกาสที่จะเกิดขึ้นสูงตามด้วยและนอกจากความถี่และค่าความน่าเชื่อถือ ยังมีตัววัดค่าความน่าเชื่อถือของกฎที่สร้างขึ้นเรียกว่า ลิฟต์

ค่าลิฟต์ จะแสดงถึงความสำคัญของความสัมพันธ์หรือเหตุการณ์ว่ามีมากน้อยเพียงใด โดยหาได้จากผลหารของค่าความน่าเชื่อถือกับจำนวนข้อมูลที่มีรายการซื้อสินค้าอย่างที่ 2 ซึ่งหากค่าลิฟต์มีค่ามากกว่า 1 หมายถึงเหตุการณ์นั้นค่าสนใจ และในกรณีที่ค่าลิฟต์มากหรือน้อยเกินไปอาจพิจารณาได้ว่า กฎนั้นไม่เป็นจริง

ผลที่ได้จากการหากฎความสัมพันธ์จะมีอยู่มากมาย หลากหลายจงต้องมีการกำจัดหรือตัดกฎบางกฎที่ไม่น่าสนใจออก หรือมีค่าน่าสนใจน้อยออกเพื่อเป็นกรลดจำนวนกฎที่มีโอกาสที่จะเกิดขึ้นน้อยออกไป โดยการกำหนดค่าความถี่น้อยที่สุด (Minimum Support) และค่าความน่าเชื่อถือน้อยที่สุด (Minimum Confidence) ซึ่งหากค่าความถี่และค่าความน่าเชื่อถือที่ได้จากกฎ

ความสัมพันธ์มีค่าน้อยกว่าที่กำหนดจะไม่นำมาพิจารณา ทำให้เวลาที่ใช้การหากฎความสัมพันธ์สั้นลงและในการกำหนดค่าความถี่น้อยที่สุดและค่าความน่าเชื่อถือน้อยที่สุดนี้ ขึ้นอยู่กับสถานการณ์หรือลักษณะของธุรกิจ

ข้อดี

- ผู้ใช้สามารถควบคุมจำนวนผลลัพธ์ได้โดยการระบุค่าความถี่น้อยที่สุดและค่าความน่าเชื่อถือน้อยที่สุด
- สามารถทำงานได้ดีกบข้อมูลขนาดใหญ่
- ในกรณีที่ข้อมูลไม่สมบูรณ์ ก็สามารถทำการไมน์นิ่งกับข้อมูลบางส่วนได้
- ไม่จำเป็นต้องระบุขอบเขตของกลุ่มข้อมูล
- สามารถจัดเก็บข้อมูลที่อยู่ในรูปแบบที่แตกต่างกันได้
- มีการแสดงผลด้วยสัญลักษณ์ ทำให้ง่ายต่อการทำความเข้าใจ

ข้อเสีย

- ในการกำหนดค่าความถี่น้อยที่สุดและค่าความน่าเชื่อถือน้อยที่สุด นั้นขึ้นอยู่ลักษณะของธุรกิจและนักวิเคราะห์ซึ่งหากกำหนดไม่ดีอาจทำให้เกิดความผิดพลาดในการวิเคราะห์ข้อมูลได้
- กฎที่ได้มานั้นอาจเป็นกฎที่เกิดขึ้นบ่อยๆ และรู้อยู่แล้วทำให้ไม่เกิดการนำไปใช้งานได้จริงในทางปฏิบัติ
- กฎที่ได้มาสามาถบอกได้เพียงแนวโน้มที่จะเกิดขึ้นด้วยกัน ไม่ได้บอกเรื่องของความเป็นเหตุเป็นผลของกฎซึ่งต้องอาศัยประสบการณ์ของนักวิเคราะห์ในการวิเคราะห์กฎต่างๆ

การพยากรณ์ข้อมูลอนุกรมเชิงเวลา

การวิเคราะห์อนุกรมเวลาเป็นระเบียบทางสถิติที่สามรถแปลงประสบการณ์ในอดีตไปพยากรณ์เหตุการณ์ในอนาคต

1. **อนุกรมเวลา**

อนุกรมเวลา (Time Series) หมายถึงข้อมูลหรือค่าสังเกตที่เปลี่ยนแปลงไปตามลำดับเวลาที่เกิดขึ้น โดยข้อมูลที่สังเกตได้ จะเก็บรวบรวมในช่วงเวลาใดก็ได้ ซึ่งจะทำให้เห็นรูปแบบของการเปลี่ยนแปลงค่าสังเกตในช่วงเวลาที่ผ่านมาเพื่อใช้ในการพยากรณ์ค่าสังเกตดังกล่าวในอนาคตได้

วิธีการพยากรณ์อนุกรมเวลา (Time Series Forecasting Method) เป็นการวิเคราะห์ลักษณะพฤติกรรมในอดีตของตัวแปรอนุกรมเวลา เพื่อพยากรณ์พฤติกรรมในอนาคตโดยถ้าค้นพบพฤติกรรมที่เป็นระบบบางอย่างในตัวแปรอนุกรมเวลาผู้ตัดสินใจก็จะสามารถสร้างแบบจำลองของพฤติกรรมของตัวแปรตามแล้วนำมาใช้ในการพยากรณ์พฤติกรรมของตัวแปรเหล่านั้นในอนาคตได้

2. องค์ประกอบของอนุกรมเวลามีอะไรบ้าง

ในการวิเคราะห์อนุกรมเวลาผู้วิเคราะห์จะแยกองค์ประกอบต่างๆ ที่ประกอบกันขึ้นเป็นอนุกรมเวลาโดยจะมีการเปลี่ยนแปลงไปตามอิทธิพลต่างๆ เช่นการเปลี่ยนแปลงการผลิต เทคโนโลยีสภาวะอากาศ เป็นต้น ในการหาคุณลักษณะของอนุกรมเวลานั้นเราสามารถใช้แบบจำลองได้หลายแบบ ซึ่งแบบจำลองที่ใช้โดยนักเศรษฐศาสตร์แบบหนึ่ง คือแบบจำลองแบบคลาสสิก (Classical Model) เป็นแบบจำลองที่อธิบายถึงองค์ประกอบของการแปรผันของอนุกรมเวลา 4 ส่วน ดังนี้

- ค่าแนวโน้ม (Secular trend) แทนด้วย T_t
- การเปลี่ยนแปลงหรือความแปรผันตามฤดูกาล (Seasonal Variation) แทนด้วย S_t
- การเปลี่ยนแปลงหรือความผันแปรตามวัฏจักร (Cyclical Variation) แทนด้วย C_t
- การเปลี่ยนแปลงหรือความผันแปรเนื่องจากเหตุการณ์ผิดปกติ (Irregular Variation) แทนด้วย I

3. รูปแบบของอนุกรมเวลา

จากปัจจัยทั้ง 4 ข้างต้น ถ้า Y แทนข้อมูลอนุกรมเวลาชุดหนึ่งๆ เราสามารถกำหนดแบบจำลองได้ 2 แบบ ดังนี้

- แบบจำลองผลบวก (Additive model) ถือว่าข้อมูลในแต่ละอนุกรมเวลาประกอบด้วยผลบวกขององค์ประกอบทั้ง 4 อย่าง

$$Y_t = T_t + S_t + C_t + I_t$$

- แบบจำลองผลคูณ (Multiplicative model) ถือว่าข้อมูลในแต่ละอนุกรมเวลาประกอบด้วยผลคูณขององค์ประกอบทั้ง 4 อย่าง

$$Y_t = T_t * S_t * C_t * I_t$$

4. การพยากรณ์ (Introduction to Forecasting)

การพยากรณ์ความตองการทางการตลาด ทำให้ผู้ผลิตสามารถวางแผนการผลิตให้สอดคล้องกับความต้องการนั้นในแต่ละช่วงเวลา อีกทั้งจะช่วยให้วางแผนการจัดซื้อวัตถุดิบต่างๆ วางแผนกำลังคน และวางแผนปริมาณสินค้าคงคลัง และวางแผนการซ่อมบำรุงรักษาได้อีกด้วย

ลักษณะของการพยากรณ์ที่ดี

- มีการกำหนดช่วงเวลาที่เหมาะสม เช่นการพยากรณ์ความต้องการในเดือนถัดไป หรือไตรมาสถัดไป
- เลือกเทคนิคที่มีความเหมาะสม และเกิดความคลาดเคลื่อนน้อย เช่น การพยากรณ์ความต้องการสินค้าบางอย่างต้องพยากรณ์ตามฤดูกาล เป็นต้น
- มีการกำหนดหน่วยพยากรณ์ เช่น จำนวนชิ้น จำนวนเงิน เป็นต้น

5. ขั้นตอนในการพยากรณ์

- กำหนดจุดประสงค์ในการพยากรณ์ เพื่อเลือกเทคนิคที่เหมาะสม
- กำหนดช่วงเวลาที่เหมาะสม ถ้าช่วงเวลาไกลหรือใกล้เกินไป จะเกิดความคลาดเคลื่อนมากขึ้น
- เลือกเทคนิคในการพยากรณ์ โดยพิจารณาข้อมูลที่หาได้ ความง่าย งบประมาณ บุคลากร และทรัพยากรที่มีอยู่
- รวบรวมและวิเคราะห์ข้อมูลในอดีต หารูปแบบของข้อมูล เช่น
- ทำการพยากรณ์
- แสดงผลการพยากรณ์ คำนวณค่าความคลาดเคลื่อน (Error, e_t) เมื่อเปรียบเทียบข้อมูลจริงในอดีต

ความคลาดเคลื่อน(e_t) = ค่าจริง (A_t) – ค่าพยากรณ์ (F_t)

- ปรับปรุงการพยากรณ์โดยอาจจะเก็บข้อมูลเพิ่มเติม หรือ ใช้เทคนิคการพยากรณ์ใหม่

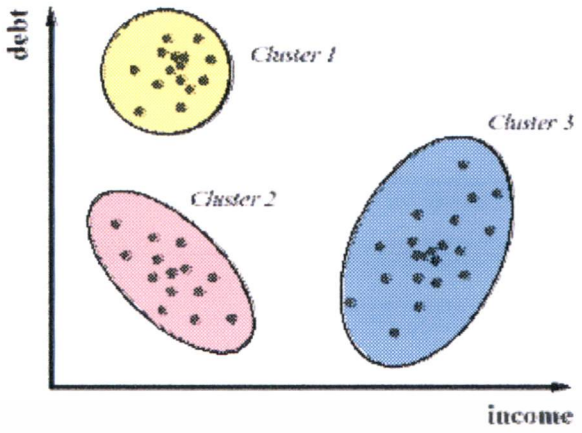
ประเภทของการพยากรณ์ สามารถแบ่งได้เป็น 3 ประเภทดังนี้

- การพยากรณ์ระยะสั้น (Short range forecasting) เช่น การพยากรณ์ช่วงเดือน, หรือช่วงไตรมาส เป็นต้น
- การพยากรณ์ระยะปานกลาง (Intermediate range forecasting) เช่น การพยากรณ์ในช่วง 1 ปี ใช้ในการวางแผนการผลิตหลัก เป็นต้น
- การพยากรณ์ระยะยาว (Long range forecasting) เช่น ถ้ามาการขยายกำลังการผลิตเพิ่มเติม อาจจะมีการพยากรณ์ ช่วงยาวกว่า 1 ปี เป็นต้น

2.4.2 Database clustering หรือ Segmentation

เมื่อฐานข้อมูลมีขนาดใหญ่ขึ้นและเก็บข้อมูลที่มีความหลากหลายมากขึ้น จึงมีความจำเป็นที่จะต้องแบ่งฐานข้อมูลออกเป็นส่วนๆ โดยข้อมูลในแต่ละส่วนจะมีความสัมพันธ์กันอย่างเหมาะสม เพื่อสะดวกในการทำเหมืองข้อมูล (Data Mining) การแบ่งฐานข้อมูลนี้อาจใช้การจัดกลุ่ม (Clustering) ซึ่งเป็นวิธีกำหนดกลุ่มหรือประเภทที่มีความแตกต่างกัน การแบ่งฐานข้อมูลนี้มักมีประโยชน์ในด้านการส่งเสริมการขายอีกทางหนึ่งด้วย ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1. กลุ่มผู้มีรายได้สูง
2. กลุ่มผู้มีรายได้ปานกลาง
3. กลุ่มผู้มีรายได้ต่ำ



รูปที่ 2.8 แสดงตัวอย่างรูป clustering

จากข้อมูลข้างต้นทำให้รู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

เทคนิคในการทำเหมืองข้อมูลเพื่อแก้ปัญหาแบบ clustering มีดังนี้

1. Demographic Clustering
2. Neural Clustering

1. Demographic Clustering

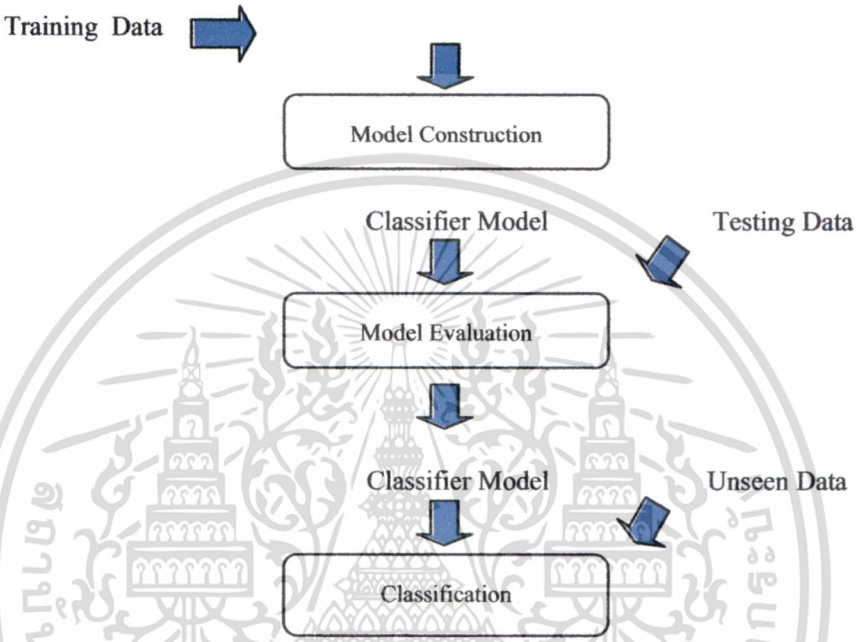
Demographic Clustering แนวคิดพื้นฐานของ Demographic Clustering คือการสร้าง segment โดยการเปรียบเทียบข้อมูล แต่ละตัวกับทุก ๆ segment ที่สร้างขึ้นในขณะที่กำลังทำ Data Mining โดยการสร้างความแตกต่างระหว่างคะแนน ให้มากที่สุด algorithm จะใส่ข้อมูลลงในแต่ละ segment ซึ่ง segment ใหม่สามารถถูกสร้างขึ้นได้ตลอดเวลาที่ทำ Data Mining ข้อดีของเทคนิคนี้คือ มันสามารถกำหนดจำนวนของ segment ที่ต้องสร้างขึ้นได้โดยอัตโนมัติและ ผลลัพธ์ของชุดข้อมูลขนาดใหญ่ที่ถูกแบ่งอย่างชัดเจน Demographic Clustering เหมาะกับข้อมูลที่มีลักษณะเป็นกลุ่ม โดยเฉพาะจำนวนของกลุ่มน้อยๆ

2. Nueral Clustering

Neural Clustering เทคนิคนี้นำ Kohonen feature neural network มาใช้ Kohonenfeuture map ใช้กระบวนการที่เรียกว่า self-organization ในการตั้งค่าหน่วยของผลลัพธ์เข้าสู่ topological map ประกอบด้วยชั้นของหน่วยประมวลผล 2 ชั้น โดยชั้นของข้อมูลเข้า (Input Layer)จะเชื่อมต่อเข้ากับข้อมูลของผลลัพธ์ (Output Layer)อย่างสมบูรณ์ เมื่อรูปแบบของข้อมูลเข้าถูกแสดงสู่ feature map หน่วยต่างๆในชั้นของผลลัพธ์ จะแข่งขันกันเพื่อสิทธิ์ที่จะได้เป็นผู้ชนะ หน่วยผลลัพธ์ที่ชนะคือ หน่วยที่น้ำหนักการเชื่อมต่อใกล้เคียงกับรูปแบบของข้อมูลเข้ามากที่สุด ในความหมายของ Euclidean distance นั้น Kohonenfeuture map สร้าง topological map โดยปรับแต่งไม่เพียงแต่น้ำหนักของผู้ชนะเท่านั้นแต่ยังปรับแต่งของหน่วยผลลัพธ์ที่อยู่ประชิดกับผู้ชนะอีกด้วย

2.4.3 Classification

เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดให้ โดยมีการแบ่งกลุ่มชัดเจนและแน่นอน ตัวอย่างเช่น ลูกค้าชั้นดี,ลูกค้าชั้นปานกลาง,และลูกค้าทั่วไป รวมถึงลูกค้าที่ไม่น่าเชื่อถือ เป็นต้น โดยพิจารณาจากประวัติลูกค้า และข้อมูลการชำระเงินของลูกค้า หรือแบ่งประเภทของการซื้อสินค้าของลูกค้า โดยการพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ Classification นี้แบ่งเป็นออก 3 ขั้นตอน ดังรูปที่ 2.9



รูปที่ 2.9 แสดงกระบวนการของ Classification

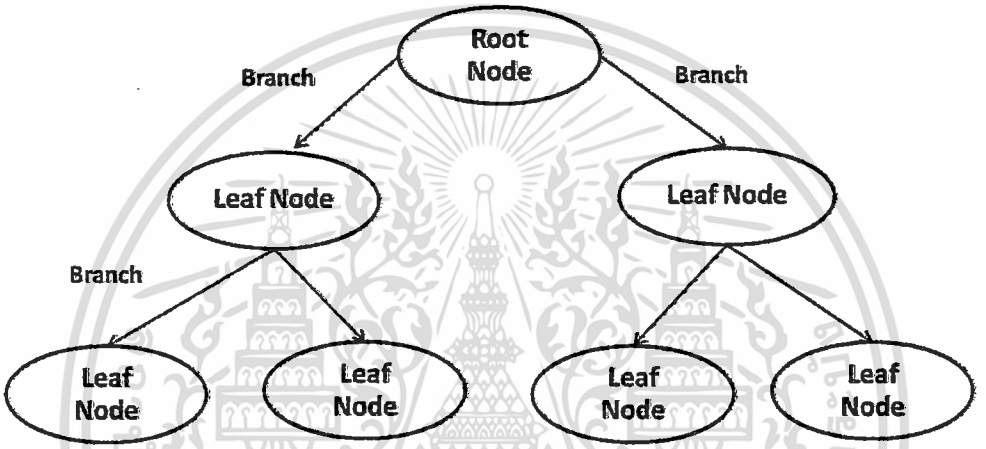
2.4.4 โครงสร้างแบบต้นไม้ (Decision Tree)

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

ซึ่งเทคนิคของ Tree Decision นั้น จะสามารถรองรับข้อมูลที่มีลักษณะของข้อมูลได้หลายลักษณะ ดังนี้

1. Nominal เป็นลักษณะข้อมูลที่เป็นข้อมูลตัวเลข เช่น 1, 2, 3, 4, 5 เป็นต้น
2. Ordinal เป็นข้อมูลที่มีลักษณะของข้อมูลที่สามารถแยกออกเป็นประเภทต่างๆ ได้ เช่น ปริญญาตรี, ปริญญาโท, ปริญญาเอก เป็นต้น
3. Interval เป็นข้อมูลที่มีลักษณะเป็นค่าต่อเนื่อง หรือค่าเฉลี่ย เช่น อุณหภูมิ, รายได้ เป็นต้น

โดยจะแสดงผลในรูปแผนภูมิที่จะแสดงการแตกผลลัพธ์ของตัวแปรแต่ละตัวออกมาเรื่อยๆ จนสุดท้ายจะได้คำตอบที่ต้องการ เหมือนลักษณะของกิ่งของต้นไม้ที่แตกออกมา ดังรูปที่ 2.10 เป็นรูปตัวอย่างของ Decision Tree



รูปที่ 2.10 แสดงตัวอย่างการแสดงผลของ Decision Tree

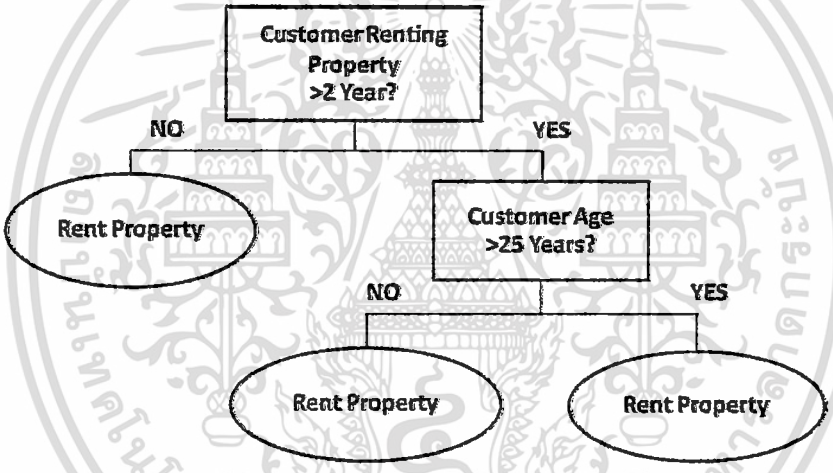
จากรูปจะประกอบไปด้วย Node ต่างๆ จุดที่เริ่มต้นของแผนภูมิจะเรียกว่า Root Node หลังจากนั้นจะแตกข้อมูลออกเป็นกิ่งต่างๆ (Branch) ตามทางเลือกของ Node ต่างๆ ซึ่งกระบวนการนี้จะดำเนินการต่อไปเรื่อยๆ จนกระทั่งได้ผลลัพธ์สุดท้ายของตัวแปรที่เป็นเป้าหมาย (Target Attribute) เรียกว่า Leaf Node ซึ่งจะเก็บค่าของคำตอบไว้ที่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

ยกตัวอย่าง สมมติว่าบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะ ดูแลอาคารต่าง ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบ

ฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา (Branch) ข้อมูลพนักงาน (Staff) ข้อมูลทรัพย์สิน (Property) และข้อมูลลูกค้า (Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer rental) รายการให้เช่า (Rentals) รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า

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

ดังรูปที่ 2.11 แสดงให้เห็นถึง Decision Tree สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า



รูปที่ 2.11 ตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน

ข้อดีของแผนภาพต้นไม้

1. วิธีการและหลักการเข้าใจได้ง่าย
2. ในการแบ่งกลุ่มข้อมูลกระทำได้ง่ายและสามารถเลือกแอตทริบิวต์ที่ดีที่สุดในการแบ่งได้

ข้อเสียของแผนภาพต้นไม้

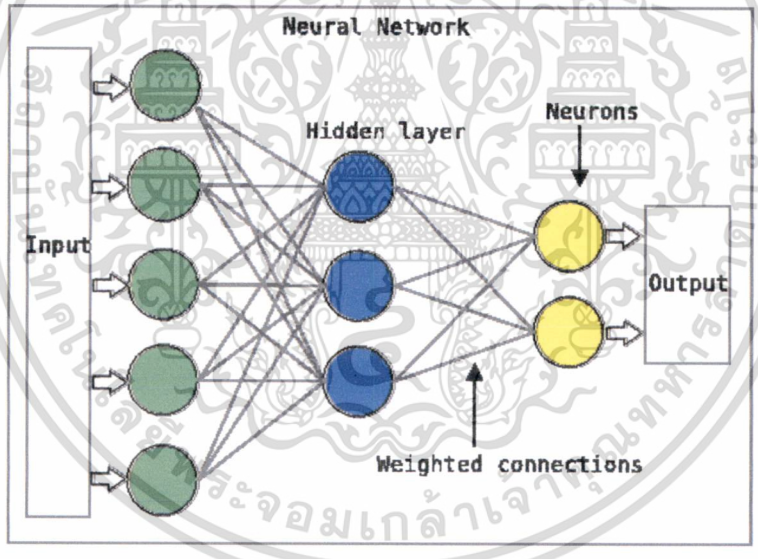
1. ความน่าเชื่อถือจะลดลงเมื่อระดับของแผนต้นไม้มีความซับซ้อนและมีจำนวนโหนดมากขึ้น ซึ่งจะก่อให้เกิดปัญหา Overfilling หรือ Overstraining
2. ข้อมูลที่จัดกลุ่มจะน้อยลงทำให้เกิดปัญหา Fragmentation ตามมา
3. กรณีข้อมูลเป็นค่าต่อเนื่อง เช่น รายได้ อายุ ซึ่งการจัดเก็บค่าของข้อมูลอาจเป็นช่วง เมื่อ

จัดกลุ่มอาจไม่สามารถทำได้

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

2.4.5 นิวรอลเน็ต หรือ นิวรอลเน็ตเวิร์ก (Neural Net)

เพื่อใช้ในการคำนวณหาค่าฟังก์ชั่นจากกลุ่มข้อมูล วิธีการของนิวรอลเน็ต หรือเรียกเต็มว่า Artificial Neural Networks : ANN เป็นวิธีการที่ให้คอมพิวเตอร์เรียนรู้จากตัวอย่างต้นแบบ แล้วฝึกให้ระบบได้รู้จักคิดแก้ปัญหาที่เกิดขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะประกอบด้วย Node สำหรับ Input Output และการประมวลผลของข้อมูล กระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ Input Layer,Output Layer ,Hidden Layer การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่างๆ ในเลเยอร์เหล่านี้ ดังเช่นตัวอย่างในรูปที่ 2.12 นิวรอลเน็ต



รูปที่ 2.12 แสดงตัวอย่างรูปนิวรอลเน็ตเวิร์ก

ข้อเสียของ นิวรอลเน็ตเวิร์ก มีดังนี้

1. นิวรอลเน็ตเวิร์กเป็นวิธีที่ยากต่อการทำความเข้าใจในโมเดลที่ถูกผลิตออกมา
2. นิวรอลเน็ตเวิร์กมีคุณสมบัติที่ไวต่อรูปแบบของอินพุท ถ้าแทนข้อมูลด้วยรูปแบบอินพุทที่แตกต่างกันก็จะสามารถผลิตผลลัพธ์ที่แตกต่างกันออกมาได้ ดังนั้นการกำหนดค่าเริ่มต้นจึงเป็นส่วนที่สำคัญ

2.5 งานของดาต้าไมน์นิ่ง (Data Mining Task)

ในทางการทำงานจริง การทำดาต้าไมน์นิ่งจะประสบความสำเร็จกับงานบางกลุ่มเท่านั้น และต้องอยู่ใต้ภาวะที่มีการแก้ปัญหาอย่างเหมาะสมกับการใช้เทคนิค การทำดาต้าไมน์นิ่งจะต้องใช้ผู้ที่มีความรู้ ความเข้าใจ ใช้เหตุผลในการแก้ปัญหา รวมถึงสามารถประยุกต์ใช้กับงานในด้านต่างๆ ได้อย่างเหมาะสม ซึ่งสามารถจัดรูปแบบของธุรกิจให้อยู่ในรูปแบบของงานทั้ง 6 งาน (ธนพงษ์ นิติการุณ และปราการ อัศวชัวันธรกุล,2545) ได้ดังต่อไปนี้

1. การจัดหมวดหมู่ (Classification)
2. การประเมินค่า (Estimation)
3. การทำนายล่วงหน้า (Prediction)
4. การจัดกลุ่มโดยอาศัยความใกล้ชิด (Affinity Group)
5. การรวมตัว (Clustering)
6. การบรรยาย (Description)

ไม่มีเทคนิคหรือเครื่องมือเพียงชนิดเดียวของการทำดาต้าไมน์นิ่งที่เหมาะสมกับงานทุกชนิด ดังนั้นผู้ใช้จึง จำเป็นต้องทราบวัตถุประสงค์ และผลลัพธ์ที่ต้องการจะได้จากงานนั้น ซึ่งงานในแต่ละชนิดของการทำดาต้าไมน์นิ่งจะมีวิธีการและเทคนิคแตกต่างกันไป ขึ้นอยู่กับชนิดของงาน

2.5.1 การจัดหมวดหมู่ (Classification)

การจัดหมวดหมู่ประกอบไปด้วยการสำรวจจุดเด่นของวัตถุที่ปรากฏออกมา และทำการกำหนดจุดเด่นนั้นๆ เป็นตัวที่ใช้แบ่งหมวดหมู่ งานในการแบ่งหมวดหมู่ คือ การบ่งบอกลักษณะโดยการอธิบายจุดเด่นที่เป็นที่รู้จักดี หรือเป็นคุณลักษณะเด่นในหมวดหมู่นั้น และเทรนนิ่งเซต (Training Set) ของตัวอย่างในแต่ละหมวดหมู่ ซึ่งมีภาระหน้าที่ในการสร้างโมเดลของบางชนิดที่ไม่สามารถจะจัดหมวดหมู่ของข้อมูลได้ ทำให้สามารถจัดหมวดหมู่ได้ ตัวอย่างของการจัดหมวดหมู่ เช่น กลุ่มข้อมูลลูกค้า เป็น กลุ่มลูกค้าที่ดี กลุ่มลูกค้าปานกลาง กลุ่มลูกค้าทั่วไป และกลุ่มลูกค้าที่ไม่น่าเชื่อถือ โดยวัดจากระดับความเสี่ยงที่จะได้รับการชำระเงิน เป็นต้น

2.5.2 การประเมินค่า (Estimation)

การประเมินค่าทางธุรกิจอย่างต่อเนื่องจะก่อให้เกิดผลลัพธ์ที่มีประโยชน์กับธุรกิจ เช่น การป้อนข้อมูลที่มีอยู่ลงไป เพื่อใช้ในการประเมินความค่าต่างๆ ที่จะก่อให้เกิดผลลัพธ์หรือสำหรับตัวแปรที่เราไม่รู้ค่าแน่นอน เช่น รายได้จากการขาย จุดสูงสุดจากการขาย หรือสภาพการเงินรายรับรายจ่าย ในทางปฏิบัติการประเมินค่าจะถูกใช้ในการทำงานประเภทการจัดหมวดหมู่ ตัวอย่างการ

ประเมินค่า เช่น การประเมินรายได้ยอดขายของปีนั้นๆ หรือการประเมินค่าใช้จ่ายในรอบเดือน เป็นต้น

2.5.3 การทำนายล่วงหน้า (Prediction)

การทำนายล่วงหน้าก้อเป็นงานที่คล้ายกับการจัดหมวดหมู่ หรือ การประเมินค่า ยกเว้นเพียงแต่จะใช้สถิติการบันทึกของการจัดหมวดหมู่เข้ามาช่วยในการทำนายอนาคตของพฤติกรรมหรือการประเมินค่าที่จะเกอดขึ้นในอนาคต ตัวอย่างการทำนายล่วงหน้า เช่น การทำนายการยกเลิกการใช้บริการโทรศัพท์มือถือ หรือ การทำนายลูกค้าของบริษัทในอีก 6 เดือนข้างหน้า เป็นต้น

2.5.4 การจัดกลุ่มโดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ของตลาด (Affinity Group)

งานในการจัดกลุ่มหรือการวิเคราะห์ตลาด เป็นการตัดสินใจรวมสิ่งที่มีความคล้ายคลึงกันเข้าด้วยกันไว้ในกลุ่มเดียวกัน ตัวอย่างของการจัดกลุ่ม โดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ของตลาด เช่น การตัดสินใจว่าสิ่งใดบ้างที่ควรจัดวางหรือจัดเรียงใกล้กันในห้างสรรพสินค้า เป็นต้น

2.5.5 การรวมตัว (Clustering)

การรวมตัว คือ การรวมงานที่ทำ หรือรวมสิ่งต่างๆในชนิดที่ต่างกันให้รวมอยู่ด้วยกันในกลุ่มย่อยๆ หรือ การรวมตัวก็คือการจัดหมวดหมู่นั้นเอง โดยการรวมตัวจะไม่พึ่งพาอาศัยการกำหนดหมวดหมู่ล่วงหน้าและไม่ใช้ตัวอย่าง ข้อมูลจะรวมตัวกันบนพื้นฐานความคล้ายคลึงของข้อมูลหรือสินค้า

2.5.6 การบรรยาย (Description)

ในบางครั้งวัตถุประสงค์ของการทำ ดาต้าไมน์นิ่ง คือการต้องการอธิบายความซับซ้อนของข้อมูล โดยการที่จะต้องการเพิ่มความเข้าใจให้กับผู้ใช้งาน หรือเพิ่มความเข้าใจในกระบวนการมากขึ้น ดังนั้น การบรรยายจะสามารถทำให้ผู้ใช้งานสามารถเข้าใจถึงการทำดาต้าไมน์นิ่ง ได้อย่างเหมาะสมมากขึ้น

2.6 การประยุกต์ใช้งานดาต้าไมน์นิ่ง

- ธุรกิจร้านหนังสือ สามารถนำเอาเทคนิคของดาต้าไมน์นิ่ง ไปใช้ในทางธุรกิจได้อย่างมีประสิทธิภาพ เช่น ธุรกิจร้านหนังสือ ออนไลน์ ของ Amazon ที่นำเอาเทคนิคดาต้าไมน์นิ่งไปใช้โดยใช้โมเดลในรูปแบบของความสัมพันธ์ (Association) ในการจำหน่ายหนังสือให้กับลูกค้าบนอินเตอร์เน็ต โดนมีเทคนิคการแนะนำหนังสือเล่มต่อไปที่มีเนื้อหาที่เกี่ยวข้อง หรือ มีความเชื่อมโยงสัมพันธ์กัน เพื่อที่จะให้ลูกค้าซื้อสินค้าเพิ่มขึ้น รวมถึงมีการลดราคา ในกรณีที่ลูกค้าทำการซื้อหนังสือที่แนะนำควบคู่กันไปด้วย

- กิจการโทรคมนาคม เช่น บริษัท ทศท คอร์ปอเรชั่น ได้มีการนำเอาเทคนิคดาต้าไมน์นิ่ง ไปใช้ในการในการหาพฤติกรรมของลูกค้าที่จะมีแนวโน้มการยกเลิกโทรศัพท์พื้นฐาน (ดลใจ กระแสเสน,2547) โดยการสร้างจำลองข้อมูลเพื่อนำมาวิเคราะห์ในโมเดล และหาโมเดลที่เหมาะสมที่ให้ค่าได้ถูกต้องมากที่สุดเพื่อนำเอาโมเดลต้นแบบนี้ไปใช้ในการหาค่าข้อมูลที่แท้จริง ซึ่งเป็นการทำนายล่วงหน้าโดยใช้ K- Means Algorithm และใช้ Database Segmentation ในการแบ่งกลุ่มข้อมูลของลูกค้า และใช้ต้นไม้ตัดสินใจ (Decision Tree)ในการแสดงผลลัพธ์
- งานด้านการศึกษา (ชลนิศา สาระ,2550) เสนอวิธีการจำแนกกลุ่มสถานภาพการสำเร็จการศึกษาโดยใช้แบบจำลองต้นไม้ตัดสินใจ ซึ่งใช้อัลกอริทึม C4.5 ซึ่งทำให้เข้าใจถึงกระบวนการ พารามิเตอร์ที่ใช้ และการวัดประสิทธิภาพของผลลัพธ์ที่ได้ ในการสร้างแบบจำลองต้นไม้สำหรับอัลกอริทึม C4.5
- การใช้เทคนิคการทำเหมืองข้อมูลเพื่อช่วยในการวิเคราะห์การขาย (บวร น้อยแสง, 2549) โดยใช้รูปแบบ ต้นไม้ตัดสินใจ (Decision Tree) และ Neural Network แบบ Back Propagation ใช้ในการวิเคราะห์ผลลัพธ์และตัวแปรที่จะใช้ในการประมวลผล โดยสรุปผลใช้ต้นไม้ตัดสินใจในการแสดงผลลัพธ์

บทที่ 3
การคัดเลือกและเตรียมข้อมูล

ในหัวข้อนี้จะกล่าวถึงการวิเคราะห์ปัญหาหรือความต้องการทางธุรกิจเพื่อที่จะนำไปสู่ขั้นตอนการคัดเลือกและการเตรียมข้อมูลตามหลักการและทฤษฎีของดาต้าไมน์นิ่งเพื่อให้สามารถนำข้อมูลไปใช้วิเคราะห์ได้อย่างถูกต้อง รวดเร็ว และเหมาะสมกับเทคนิคดาต้าไมน์นิ่งแบบต่างๆ

3.1 แหล่งข้อมูล

แหล่งข้อมูลที่นำมาวิเคราะห์มาจากระบบปฏิบัติงานหลัก นำข้อมูลที่ได้รับการอนุญาตจากทางบริษัท มาใช้ในการทำงานดาต้าไมน์นิ่ง โดยปริมาณข้อมูลที่ได้รับมานั้น จะเป็นข้อมูลดิบที่ทางบริษัทเรียกข้อมูลรายงานออกมาให้ โดยยังไม่มีการดัดแปลง แก้ไข ข้อมูลที่ได้มานี้ จะอยู่ในช่วงระยะเวลา 1 ปี โดยมีระยะเวลาตั้งแต่ วันที่ 1 มกราคม พ.ศ.2555 ถึงวันที่ 30 ธันวาคม พ.ศ.2555 เท่านั้น โดยข้อมูลที่ได้รับมาจะแบ่งจำแนกได้ ออกเป็น 3 ส่วนใหญ่ๆ คือ

1. ข้อมูลลูกค้า ประกอบด้วยข้อมูลทั้งหมด 5,221 เรคคอร์ด
2. ข้อมูลสินค้า ประกอบด้วยข้อมูลทั้งหมด 1,950 เรคคอร์ด
3. ข้อมูลการขายสินค้า ประกอบด้วยข้อมูลทั้งหมด 65,500เรคคอร์ด

โดยในข้อมูลที่ได้รับมานั้น จะมีรายละเอียดข้อมูลภายในเป็นจำนวนมาก ซึ่งเป็นทั้งข้อมูลที่มีความจำเป็นเกี่ยวข้องกับการทำดาต้าไมน์นิ่ง และข้อมูลบางส่วนที่ไม่เกี่ยวข้องกับการทำงานเลย จึงจำเป็นจะต้องมีการคัดเลือกข้อมูล แยกประเภทข้อมูล และการทำความสะอาดข้อมูล โดยผ่านกระบวนการขั้นตอนต่างๆในการทำดาต้าไมน์นิ่งก่อนถึงจะสามารถนำข้อมูลที่ได้รับมา ไปทำการประมวลผล เพื่อที่จะต้องการหาคำตอบนั้นๆ และนำความรู้ที่ได้จากการประมวลผลที่ตีความแล้วไปประยุกต์ใช้ต่อไป

3.2 ระบบปัจจุบันและปัญหาของระบบ

จากเทคนิคดาต้าไมน์นิ่งดังที่กล่าวมาในบทข้างต้นนั้น สามารถนำมาประยุกต์ใช้ได้กับองค์กรที่ต้องการวิเคราะห์ข้อมูลที่ซ่อนอยู่ในฐานข้อมูลโดยองค์กรที่นำมาเป็นกรณีศึกษา ได้แก่ บริษัท แคบริค (ประเทศไทย) จำกัด ซึ่งดำเนินธุรกิจเกี่ยวกับการผลิต และจัดจำหน่ายสินค้าประเภทเทปกาว ให้กับกลุ่มลูกค้าที่หลายประเภท เช่น ห้างสรรพสินค้า ร้านเครื่องเขียน โรงงานอุตสาหกรรม รวมถึง ตัวแทนจำหน่ายสินค้าต่อไปยังลูกค้ารายย่อย เป็นต้น ซึ่งทำให้มีความหลากหลายของกลุ่มข้อมูลลูกค้า โดยบริษัทมีช่องทางการจำหน่ายผ่านทางพนักงานขายของท่านั้น ไม่มีหน้าร้านเพื่อจัดจำหน่ายสินค้าโดยตรงและมีคลังเก็บสินค้า ซึ่งตั้งอยู่เขตมีนบุรีจังกรุงเทพมหานครฯ

3.2.1 ระบบปัจจุบันและปัญหาของระบบ

ระบบปฏิบัติการหลักของบริษัทฯ มาจากการขายสินค้าผ่านทางพนักงานขายของบริษัท รวมถึงผ่านทางการสั่งซื้อสินค้าทางโทรศัพท์ โดยมีเจ้าหน้าที่ของบริษัท ทำการเก็บข้อมูลและบันทึกข้อมูลในการสั่งซื้อสินค้าแต่ละวันโดยใช้ฐานข้อมูลเชิงสัมพันธ์ ซึ่งปัจจุบันยังไม่มีระบบสารสนเทศที่ช่วยในการสนับสนุนการตัดสินใจในการวางแผนกลยุทธ์หรือ มาช่วยในการวิเคราะห์ข้อมูล ทำให้ผู้บริหารไม่สามารถวิเคราะห์ข้อมูลได้อย่างสะดวก รวดเร็วและถูกต้องเท่าที่ควร ปัญหาที่ผู้บริหารพบหรือต้องการมีดังนี้

- ไม่สามารถทราบถึงพฤติกรรมการซื้อสินค้าของลูกค้าหรือความนิยมของผู้ซื้อสินค้าแต่ละประเภท
- ต้องการทราบถึงกลุ่มของลูกค้าที่ซื้อสินค้าเพื่อช่วยในการจัดโปรโมชั่นและทำ Direct Mail ถึงลูกค้าแต่ละกลุ่ม
- ต้องการคาดคะเนล่วงหน้าถึงยอดขายในเดือนหรือในปีถัดไปเพื่อช่วยในการวางแผนกลยุทธ์ของบริษัท
- แนวทางการขายสินค้าว่าในช่วงเวลาใดที่สินค้าประเภทใดขายดี หรือเป็นที่นิยมต่อลูกค้า และสินค้าประเภทใดที่ไม่ได้รับความนิยมต่อลูกค้าเท่าที่ควร

3.2.2 การกำหนดวัตถุประสงค์ทางธุรกิจ

จากปัญหาที่เกิดขึ้นของบริษัททำให้บริษัทจำเป็นต้องหาหนทางที่จะแก้ไขปัญหาดังกล่าว ซึ่งสามารถแก้ไขปัญหาได้โดยวิธีการของดาต้าไมน์นิ่ง ดังนั้น ขั้นตอนแรกของการทำดาต้าไมน์นิ่งคือการกำหนดวัตถุประสงค์ หรือ กำหนดปัญหาทางธุรกิจของ ผู้บริหาร ซึ่งจากการสอบถามผู้บริหารของบริษัท สามารถวิเคราะห์ความต้องการ ได้ดังนี้

- วิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้า เช่น ลูกค้าส่วนใหญ่เมื่อซื้อสินค้าอย่างหนึ่งแล้ว มักจะซื้อ สินค้าคู่กับสินค้าอะไร
- วิเคราะห์กลุ่มลูกค้าในการสั่งซื้อสินค้า โดยจำแนกความสัมพันธ์ด้านพื้นที่ และธุรกิจขององค์กรลูกค้า
- ทำนายยอดปริมาณสินค้าในการสั่งซื้อล่วงหน้า

3.3 การเตรียมข้อมูล

การเลือกข้อมูลจากฐานข้อมูลสำหรับการนำข้อมูลไปใช้ในการทำดาต้าไมน์นิ่ง โดยหลักการในการเลือกข้อมูลจากฐานข้อมูลนั้นเพื่อให้ได้ข้อมูลที่เหมาะสมและครอบคลุมการวิเคราะห์ข้อมูลจะพิจารณาจากการวิเคราะห์ปัญหาในหัวข้อ 3.2.2 การกำหนดวัตถุประสงค์ทางธุรกิจ ซึ่งจะต้องคัดเลือกข้อมูลที่สำคัญ และเกี่ยวข้องกับการวิเคราะห์ปัญหาในแต่ละหัวข้อดังนี้

3.3.1 ต้องการวิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้า ได้แก่ "ลูกค้าส่วนใหญ่เมื่อซื้อสินค้าอย่างหนึ่งแล้วมักจะซื้อสินค้าคู่กับสินค้าอะไร

- การคัดเลือกข้อมูล ในกระบวนการนี้จะคัดเลือกข้อมูล โดยการนำเอาข้อมูลการขายสินค้าของบริษัทมา ใช้ โดยภายในข้อมูลการขายสินค้า ข้อมูลที่ได้รับมาจะมีการแบ่งรายชื่อลูกค้าและวันที่ลูกค้าที่มีการซื้อสินค้า ซึ่งเรียกว่าตารางการขายสินค้า จะประกอบด้วยข้อมูล ดังนี้ รหัสลูกค้า,ชื่อลูกค้า,รหัสสินค้า,รายละเอียดสินค้า,ส่วนลด,ปริมาณการซื้อ,ราคาสินค้า รวมถึง ยอดรวมการซื้อสินค้า

- การทำความสะอาดข้อมูล เนื่องจากข้อมูลที่เก็บอยู่ในฐานข้อมูลนั้นอาจไม่ถูกต้องหรือไม่เหมาะสมสำหรับการทำดาต้าไมน์นิ่ง ดังนั้น จำเป็นต้องตรวจสอบข้อมูลให้ถูกต้องก่อน โดยเริ่มจากการตรวจสอบข้อมูลที่ขาดหายไป (Missing data) ในแต่ละคอลัมน์ ถ้าพบว่าข้อมูลขาดหายไป ก็จะมีการเติมข้อมูลลงไปให้ครบ โดยใช้ข้อมูลที่ได้กำหนดไว้แล้วตามความเหมาะสมของธุรกิจ จากการตรวจสอบข้อมูลพบว่าเลขที่ใบสั่งซื้อสินค้าและชื่อสินค้า นั้นไม่มีข้อมูลขาดหายไป เนื่องจากเป็นข้อมูลที่ระบบสร้างขึ้นและไม่อนุญาตให้มีค่าว่างเกิดขึ้น แต่สิ่งที่ทำการตรวจพบต่อมา คือ ช่องเครดิต ช่องอ้างอิง และช่องส่วนลด ซึ่งมีค่าว่างเกิดขึ้นอยู่เป็นจำนวนมาก โดยจะทำการตัดทิ้งทั้งแถวไม่ได้ เนื่องจากจะทำให้มีข้อมูลที่ขาดหาย ซึ่งอาจจะทำให้ข้อมูลที่จะไปทำดาต้าไมน์นิ่งไม่สมบูรณ์ จึงทำการตัดคอลัมน์ ดังกล่าวทิ้งไป เพื่อให้ข้อมูลที่เหลือนั้นมีความถูกต้องมากที่สุด จากนั้นตรวจสอบข้อมูลที่มีสิ่งรบกวน (Noisy data) ซึ่งเป็นความผิดพลาดที่เกิดขึ้นแบบไม่ตั้งใจ โดยเกิดจากการกรอกข้อมูลที่ผิดพลาด ทำให้ค่าไม่เป็นดังที่ควรจะเป็น จึงจำเป็นต้องมีการตรวจสอบความถูกต้อง ให้ครบถ้วน ดังรูปที่3.1 ที่ยังไม่ทำความสะอาดข้อมูลและกำจัดข้อมูลที่ขาดหาย เป็นต้น

รหัส	ชื่อลูกค้า วันที่	เลขที่	คลัง	เครดิต	รหัสสินค้า	ชื่อสินค้า	หน่วย	อ้างอิง	ปริมาณ	ราคา/หน่วย	ลด	เงินบาท
132041	บริษัท กลชาญวิทย์ เซ็นเตอร์ จำกัด											
	8/16/2554	IV 110808		4 Credit: 30 [	DBG1A104	เทปOPP ก	ม้วน	SO-2011-0	360	10.13		3,646.80
	8/16/2554	IV 110808		4 Credit: 30 [	DDE2E310	เทปกาวสอ	ม้วน	SO-2011-0	480	5.25		2,520.00
	8/20/2554	IV 110811		4 Credit: 30 [	DBG1A104	เทปOPP ก	ม้วน	SO-2011-0	360	10.13		3,646.80
132042	ห้างหุ้นส่วนจำกัด เอ็นเอสที											
	6/7/2554	IV 110603		4	DBA1B325	เทปใส้กาว	ม้วน		32	16.50	10%	475.20
	6/7/2554	IV 110603		4	DBA1B325	เทปใส้กาว	ม้วน		32	16.50	10%	475.20
	6/7/2554	IV 110603		4	DDE2E310	เทปกาวสอ	ม้วน		96	5.40	10%	466.56
132043	ห้างหุ้นส่วนจำกัด ชายไพรมูเอ้ดีไก้											
	6/1/2554	IV 110600		4	DEM5C127	สติกเกอร์เล	ม้วน		96	16.20	3%	1,508.54
	6/1/2554	IV 110600		4	DEM5C127	สติกเกอร์เล	ม้วน		96	16.20	3%	1,508.54
132044	บริษัท บี.พ้อยท์ มาร์เก็ตติ้ง จำกัด											
	5/23/2554	IV 110509		4	DBA1B325	เทปใส้กาว	ม้วน		60	28.00		1,680.00
	10/6/2554	IV 111002		4	DBG1A104	เทปOPP ก	ม้วน	SO-2011-1	24	14.00		336.00
	10/6/2554	IV 111002		4	DBG1A104	เทปOPP ก	ม้วน	SO-2011-1	18	14.00		252.00
132050	บริษัท สมไจ (อ้อมน้อย) จำกัด											
	7/12/2554	IV 110705		4 Credit: 30 [	DEU1E210	Filament t	ม้วน	SO-2011-0	144	25.50		3,672.00
	7/12/2554	IV 110705		4 Credit: 30 [	DEU1E210	Filament t	ม้วน	SO-2011-0	72	51.00		3,672.00
	7/12/2554	IV 110705		4 Credit: 30 [	DEK1C115	PVC สีแสด	ม้วน	SO-2011-0	12	130.00		1,560.00
	7/12/2554	IV 110705		4 Credit: 30 [	DBG1A104	เทปOPP ก	ม้วน	SO-2011-0	360	9.45		3,402.00
	8/26/2554	IV 110814		4 Credit: 30 [	DBB2B113	เทปกาวสน	ม้วน	SO-2011-0	12	36.00	10%	388.80

รูปที่ 3.1 รายละเอียดของรูปที่ใช้ในการวิเคราะห์พฤติกรรมการซื้อสินค้าที่ยังไม่ได้ทำการแปลงข้อมูล

ตารางที่ 3.1 รายละเอียดของตารางข้อมูลการขายสินค้า

ชื่อ	ความหมาย
รหัส	หมายเลขรหัสลูกค้า มีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อลูกค้า	ชื่อลูกค้าที่มีการซื้อสินค้ากับบริษัท โดยมี บุคคล ,ห้างหุ้นส่วน,บริษัทและนิติบุคคล
วันที่	วันที่ลูกค้าทำการซื้อสินค้า
เลขที่	เลขที่ใบสั่งซื้อสินค้า
คลัง	คลังจัดเก็บสินค้าเพื่อจำหน่าย
เครดิต	ระยะเวลาในการชำระเงิน
รหัสสินค้า	เลขรหัสสินค้ามีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อสินค้า	รายละเอียดสินค้า
หน่วย	หน่วยนับของสินค้า
อ้างอิง	เลขที่การสั่งซื้อสินค้า
ปริมาณ	จำนวนสินค้า
ราคา/หน่วย	ราคาสินค้า
ลด	ส่วนลด
เงินบาท	จำนวนเงินรวมการซื้อสินค้า

- แปลงข้อมูล เป็นการทำให้ข้อมูลอยู่ในรูปแบบที่เหมาะสม โดยอาศัยเงื่อนไข เพื่อให้ข้อมูลสอดคล้องกับการวิเคราะห์ โดยพิจารณาข้อมูลทั้งหมดว่าส่วนไหนบ้างที่ไม่จำเป็นในการนำมาพิจารณา รวมถึงนำข้อมูลที่ได้ไปแปลงให้อยู่ในรูปแบบที่เหมาะสมในการทำการวิเคราะห์ โดยทำการแปลงข้อมูลการขายสินค้าจากเดิมที่มีการบันทึกการขายสินค้าเป็นจำรวนรวมต่อหนึ่งรายการสินค้าจะทำการแปลงเป็นกลุ่มข้อมูลสินค้า เพื่อง่ายต่อการนำไปวิเคราะห์ โดยทำการแบ่งกลุ่มข้อมูลสินค้า ออกเป็นกลุ่ม แล้วทำการกรอกรายละเอียดการซื้อของลูกค้าแต่ละราย ฌปืนต้น ดังรูปภาพต่อไปนี้

- การรวบรวมข้อมูลจากที่ได้เห็นข้อมูลตัวอย่างแล้ว เมื่อผ่านขั้นตอนการแปลงข้อมูล รวมถึงการทำความสะอาดข้อมูลแล้ว เราก็จะทำการรวมข้อมูลให้อยู่ในรูปแบบเดียวกัน โดยมีข้อมูลที่จำเป็นในการวิเคราะห์เท่านั้น โดยข้อมูลที่ผ่านกระบวนการแปลงและการทำความสะอาดข้อมูลแล้ว จะเห็นได้ดังรูปภาพ3.2

ชื่อลูกค้า	ชื่อสินค้า													จังหวัด	ภาค	ประเภทลูกค้า
	เทปผ้ากาว	เทปกระดาษกาวย่น	เทปกาวย่น	เทป opp	ตราช้างเทป	เทปใสกาวยาง	เทปใสกาวน้ำ	Filament tape	สติ๊กเกอร์	เทปกาวสองหน้า	เทปพันสายไฟ	เทปอลูมิเนียม	เทปกาวโฟมสองหน้า			
ศูนย์หนังสือเชียงคำ	L	L	L	0	0	0	0	0	L	L	0	0	0	พะเยา	1	3
ร้านเพื่อนนักเรียน	L	L	L	0	0	L	0	0	0	L	0	0	0	สมุทรปราการ	2	4
จิปาถะ	L	L	L	L	0	0	0	0	L	L	0	0	L	สุโขทัย	1	4
ศูนย์หนังสือเสมาทอง	M	L	0	L	0	0	0	0	L	M	0	0	L	สุพรรณบุรี	2	3
ไม้น้ำ	0	0	0	0	0	0	0	0	M	M	0	0	L	พะเยา	1	4
เหรียญทอง	M	0	0	L	L	0	0	0	0	M	L	0	0	สุพรรณบุรี	2	4
จันทรภัณฑ์	L	0	0	L	0	L	0	L	M	L	0	0	0	สุพรรณบุรี	2	4
ทรัพย์ถาวร เซ็นเตอร์														กรุงเทพ	2	1
โฮมโปรดักส์ เซ็นเตอร์ (มหาชน)	H	0	0	0	0	0	0	0	0	0	0	0	0	อยุธยา	2	1
ธนินท์เรศ	H	M	M	L	L	L	0	L	L	L	L	0	L	สุโขทัย	1	1
ก.นวดกรรม	H	M	M	M	L	L	0	L	M	M	0	L	0	สุโขทัย	1	1
ที.เอส.อาร์ พรีเมี่ยม โปรดักส์	0	0	0	M	0	0	0	0	0	0	0	0	0	กรุงเทพ	2	1
ดีนเลอร์	0	0	0	0	0	0	0	0	M	M	0	0	0	กำแพงเพชร	1	4
อิมพ์ซีน เบิ้ล ซัพพลาย	L	0	0	H	0	0	H	0	0	0	0	0	0	ชลบุรี	4	1
วิแอนท์ อินเตอร์โฮลดิ้ง	0	L	0	H	0	M	0	0	0	L	0	0	0	กรุงเทพ	2	1

< 1000 LOW = L

> 1000 - 5000 Mim = M　　> 5000ขึ้นไป High = H

รูปที่ 3.2 รูปแปลงข้อมูลที่นำไปใช้ในการวิเคราะห์

จากรูป 3.2 เป็นตัวอย่างข้อมูลที่ผ่านกระบวนการเตรียมข้อมูล ทำความสะอาดข้อมูล และแปลงข้อมูลแล้ว สามารถนำข้อมูลนี้ไปใช้ในการวิเคราะห์ได้อย่างถูกต้องและมีประสิทธิภาพ

เทคนิคและอัลกอริทึม ที่ใช้ในการวิเคราะห์

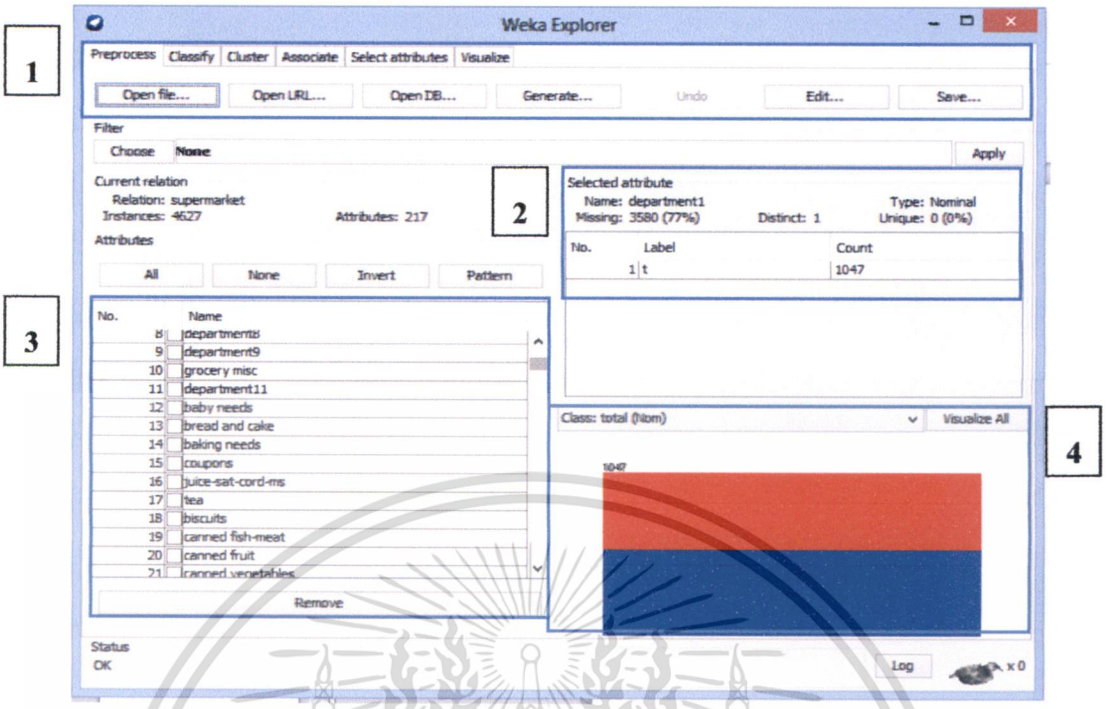
กรอบส่วนสีน้ำเงิน เป็นส่วนของการนำข้อมูลลูกค้ามาจัดเรียงใหม่ โดยนำข้อมูลลูกค้าที่เป็นรายชื่อลูกค้าเท่านั้นมาจัดเรียง

กรอบส่วนสีแดง เป็นส่วนของการนำข้อมูลสินค้ามาจัดเรียง โดยนำข้อมูลสินค้าที่แบ่งออกเป็นกลุ่ม แต่ละประเภทมากำหนดตามรูปแบบการซื้อสินค้าของลูกค้า

กรอบส่วนสีเขียว เป็นส่วนของการนำข้อมูลลูกค้า เช่น จังหวัด ภาค และรูปแบบลูกค้าโดยกำหนดเป็นภาค

ในการวิเคราะห์เพื่อหาคำตอบนี้ ได้ใช้เทคนิค กฎความสัมพันธ์ (Link Association) เป็นการค้นหาความสัมพันธ์ของข้อมูล ซึ่งนิยมใช้ในการหาความสัมพันธ์ของสินค้าที่เกิดขึ้นในรายการเดียวกัน ที่มีแนวโน้มว่าจะเกิดขึ้นพร้อมๆ กัน เช่น พิจารณาสินค้าที่มักจะถูกซื้อควบคู่กันไปในตราเดียวกัน การวิเคราะห์ในลักษณะนี้เรียกว่า “Market Basket Analysis” ซึ่งจะนำไปใช้วิเคราะห์การซื้อสินค้าจากลูกค้าทำให้ผู้ประกอการธุรกิจสามารถนำไปช่วยในการวางแผนทางการตลาดหรือกำหนดกลยุทธ์ทางด้านการจำหน่ายสินค้าและบริการได้เช่น การจัดโปรโมชั่น การวางตำแหน่งของสินค้า เป็นต้น

หลักการทำงานของวิธีนี้คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลที่มีอยู่เป็นจำนวนมาก จุดมุ่งหมายของ Link Analysis คือ การสร้าง Link ที่เรียกว่า “Association” ระหว่าง เรคอร์ดหรือกลุ่มของข้อมูล ในฐานข้อมูล เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆ หรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้า ซึ่งประเมินได้จากข้อมูลที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล ตามรูปที่ 3.3 รูปแบบตัวอย่างการใช้เทคนิค Apiori



รูปที่ 3.3 รูปตัวอย่างข้อมูลในเทคนิค Ascociation Rule

รูปที่ 3.3 รูปภาพแสดงตัวอย่างการนำเข้าข้อมูลด้วยโปรแกรม Weka ซึ่งแสดงข้อมูลที่จะใช้ในการทำคาต้าไมน์นิ่งด้วยเทคนิค Ascociation Rule โดยในรูปภาพหน้าต่างการใช้งานโปรแกรม WEKA นี้ประกอบด้วย 4 ส่วนหลักๆ ดังนี้

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล
2. ส่วนแสดงผลข้อมูล ในแต่ละแอตทริบิวต์ ว่ามีรูปแบบข้อมูลเป็นแบบใดและข้อมูลประกอบด้วยจำนวนเท่าใด
3. ส่วนแสดงข้อมูลทั้งหมด ว่ามีข้อมูลใดประกอบบ้าง ซึ่งเราสามารถเลือกข้อมูลที่ต้องการ และสามารถเลือกข้อมูลที่ไม่ต้องการ สามารถนำข้อมูลที่ไม่ต้องการออกได้
4. ส่วนแสดงผลเป็นรูปภาพ จะสามารถเห็นข้อมูลโดยรวมทั้งหมดให้เป็นรูปแบบรูปภาพ

1

2

elation: supermarket

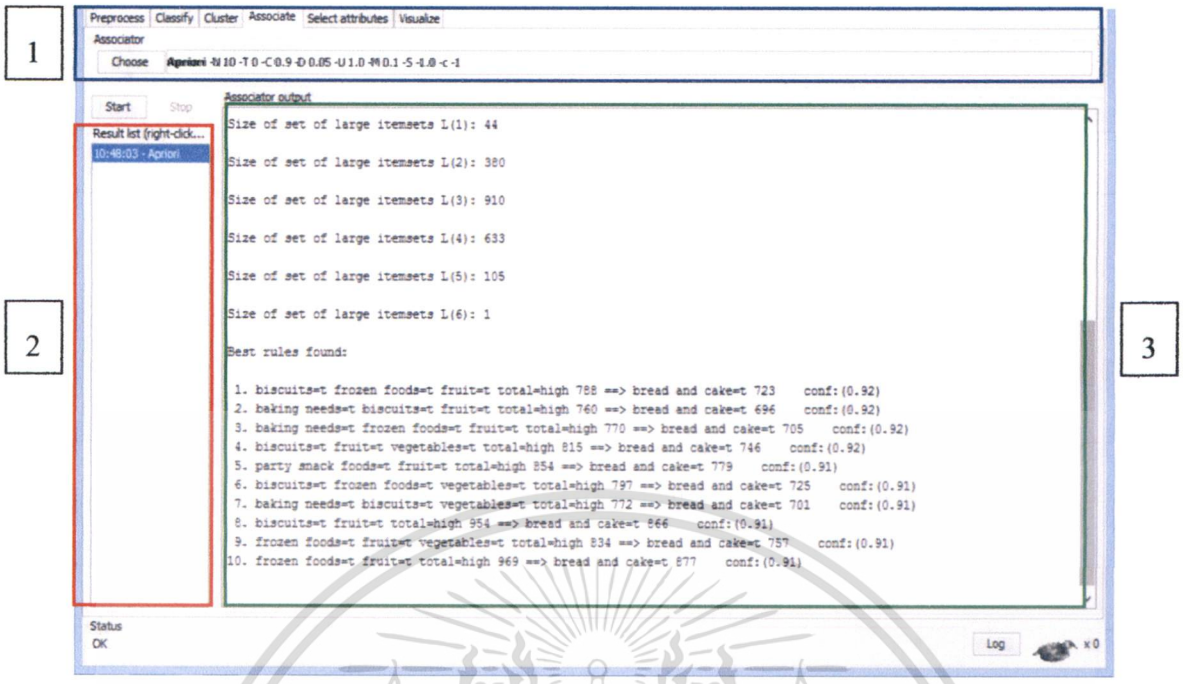
bread and cake Nominal	baking needs Nominal	coupons Nominal	juice-sat-cord-ms Nominal	tea Nominal	biscuits Nominal	canned fish-meat Nominal	canned fruit Nominal	canned vegetables Nominal	breakfast food Nominal	cigs-tobacco pkts Nominal
	t		t		t			t		
						t	t	t		
	t		t		t		t			
	t		t		t			t	t	
	t		t	t						
	t		t	t	t			t	t	
	t		t	t	t		t	t	t	
	t		t		t			t		
	t		t		t		t	t		
	t		t		t		t	t	t	
	t		t			t	t	t		
	t		t		t				t	t
	t									
			t		t				t	
	t		t	t	t		t	t	t	
			t	t	t		t			
					t					
	t								t	t
					t	t				
					t					t
	t		t	t	t			t	t	
	t				t					
			t		t	t				
	t			t						
	t						t	t	t	t
	t		t		t			t	t	t
	t		t		t			t	t	
	t			t				t		
	t		t		t			t		
			t	t		t				
					t		t	t	t	
					t		t	t		

รูปที่ 3.4 รูปตัวอย่างข้อมูลในการขายสินค้าของห้างสรรพสินค้า

รูปที่ 3.4 รูปภาพตัวอย่างของข้อมูลในการขายสินค้าของห้างสรรพสินค้า เป็นรูปแบบข้อมูลที่จะนำมาใช้ใช้ในการทำดาต้าไมน์นิ่ง โดยในส่วนรูปภาพนี้จะเป็นการแสดงข้อมูลในโปรแกรม WEKA ซึ่งจะประกอบด้วยข้อมูล 2 ส่วน

1.ส่วนข้อมูลรายการสินค้า จะประกอบด้วยรายการข้อมูลสินค้าแต่ละชนิดที่มีอยู่ในข้อมูล

2.ส่วนรูปแบบข้อมูล การจำหน่ายสินค้าชนิดนั้นๆ โดยข้อมูลที่มีการขายสินค้านั้น จะมีข้อมูลบันทึกเป็นค่า t ส่วนรายการสินค้าที่ไม่มีการซื้อ จะไม่มีข้อมูลภายใน



รูปที่ 3.5 รูปตัวอย่างการประมวลผลข้อมูลในเทคนิค Ascociation Rule

รูปที่ 3.5 รูปภาพแสดงตัวอย่างการประมวลผลด้วยโปรแกรม Weka ในเทคนิค Ascociation Rule ซึ่งแสดงวิธีการทำด้วย Apiori โดยในรูปภาพนี้ประกอบด้วย 3 ส่วน ได้แก่

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล
2. ส่วนแสดงผลในการเลือกโมเดลและอัลกอริทึมที่ใช้
3. ส่วนแสดงผลลัพธ์ที่ได้จากการคำนวณ โดยในรูปภาพจะแสดงข้อมูลผลลัพธ์ที่ได้จากการคำนวนโดยใช้อัลกอริทึมเลือกและแสดงผลลัพธ์ตามที่ต้องการ

สำหรับ วิธีการ ในการ ทำ Association rule นั้น มีหลายวิธีอยู่เหมือนกัน แต่ก็มีที่เข้าใจง่ายๆ เช่น

- Apriori
- Fequence FP Growth

ข้อดี

- ช่วย ให้ทราบพฤติกรรมของเป้าหมาย ได้

ข้อเสีย

- บางข้อมูลใช้วิธี Apriori จะให้ความเที่ยงตรงสูงกว่าใช้ FP Growth บางข้อมูลใช้ วิธี FP Growth ก็จะให้ความ เที่ยงตรงสูงกว่า Apriori

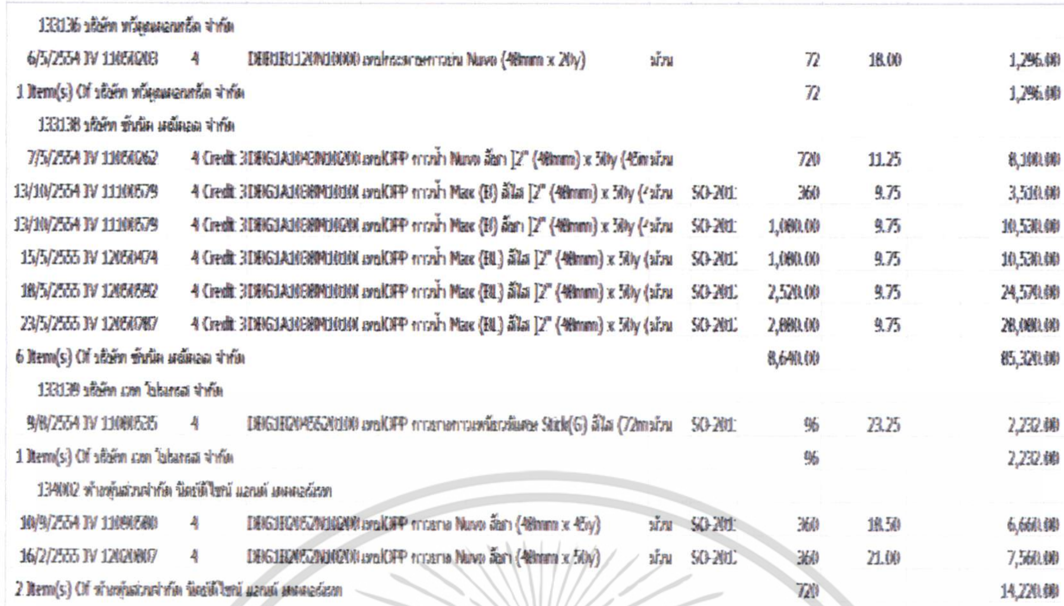
ดังนั้น ในทางปฏิบัติจะต้องทำการทดสอบหลายๆโมเดล และต้องเอา Model ที่ได้มา Evaluated เพื่อ เลือก วิธีที่เหมาะสม และทำการทดลองหรือนำไปทดสอบเพื่อวัดประสิทธิภาพของโมเดล

3.3.2 วิเคราะห์พฤติกรรมลูกค้ากับการซื้อสินค้า (แยกตามภูมิศาสตร์ และรูปแบบองค์กรลูกค้า)

- **การคัดเลือกข้อมูล** ในกระบวนการนี้จะคัดเลือก ข้อมูลที่เกี่ยวข้อง โดยข้อมูลที่จะนำมาทำการวิเคราะห์ประกอบด้วย 2 กลุ่มข้อมูล คือ ตารางข้อมูลลูกค้า และ ตารางการขายสินค้า โดยจะนำข้อมูลทั้งสองตารางและมีความเกี่ยวข้องกัน มารวมกัน เพื่อที่จะได้สามารถทำการวิเคราะห์จากโจทย์ที่ตั้งคำถามได้ถูกต้องมากยิ่งขึ้น ดังตารางต่อไปนี้

ตารางที่ 3.2 รายละเอียดของตารางข้อมูลลูกค้าที่สั่งซื้อสินค้า

ชื่อ	ความหมาย
รหัส	หมายเลขรหัสลูกค้า มีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อลูกค้า	ชื่อลูกค้าที่มีการซื้อสินค้ากับบริษัท โดยมี บุคคล ,ห้างหุ้นส่วน,บริษัท,และนิติบุคคล
วันที่	วันที่ลูกค้าทำการซื้อสินค้า
เลขที่	เลขที่ใบสั่งซื้อสินค้า
คลัง	คลังจัดเก็บสินค้าเพื่อจำหน่าย
เครดิต	ระยะเวลาในการชำระเงิน
รหัสสินค้า	เลขรหัสสินค้ามีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อสินค้า	รายละเอียดสินค้า
หน่วย	หน่วยนับของสินค้า
อ้างอิง	เลขที่การสั่งซื้อสินค้า
ปริมาณ	จำนวนสินค้า
ราคา/หน่วย	ราคาสินค้า
ลด	ส่วนลด
เงินบาท	จำนวนเงินรวมการซื้อสินค้า



133136 บริษัท ทวีสุขแพลกกรีต จำกัด							
6/5/2554 IV 11050203	4	DEB01E0112WN10000 เทปกระดาษกาวย่น Nuvo (48mm x 20y)	ม้วน		72	18.00	1,296.00
1 Item(s) Of บริษัท ทวีสุขแพลกกรีต จำกัด					72		1,296.00
133138 บริษัท ซันกัน [illegible] จำกัด							
7/5/2554 IV 11050262	4 Credit	3DRG1A1043N10200 เทปOPP กาวน้ำ Nuvo สีชา [2" (48mm) x 50y (45ไมครอน	ม้วน		720	11.25	8,100.00
13/10/2554 IV 11100579	4 Credit	3DRG1A1038N10100 เทปOPP กาวน้ำ Max (B) สีใส [2" (48mm) x 50y (4	ม้วน	SO-201	360	9.75	3,510.00
13/10/2554 IV 11100579	4 Credit	3DRG1A1038N10200 เทปOPP กาวน้ำ Max (B) สีชา [2" (48mm) x 50y (4	ม้วน	SO-201	1,080.00	9.75	10,530.00
15/5/2555 IV 12050474	4 Credit	3DRG1A1038N10100 เทปOPP กาวน้ำ Max (BL) สีใส [2" (48mm) x 50y (	ม้วน	SO-201	1,080.00	9.75	10,530.00
18/5/2555 IV 12050592	4 Credit	3DRG1A1038N10100 เทปOPP กาวน้ำ Max (BL) สีใส [2" (48mm) x 50y (	ม้วน	SO-201	2,520.00	9.75	24,570.00
23/5/2555 IV 12050787	4 Credit	3DRG1A1038N10100 เทปOPP กาวน้ำ Max (BL) สีใส [2" (48mm) x 50y (	ม้วน	SO-201	2,880.00	9.75	28,080.00
6 Item(s) Of บริษัท ซันกัน [illegible] จำกัด					8,640.00		85,320.00
133139 บริษัท เวด โปรเกรส จำกัด							
9/8/2554 IV 11080535	4	DRG1EQ0455201100 เทปOPP กาวยางกาวเหนียว[illegible] Stick(G) สีใส (72mm	ม้วน	SO-201	96	23.25	2,232.00
1 Item(s) Of บริษัท เวด โปรเกรส จำกัด					96		2,232.00
134002 ห้างหุ้นส่วนจำกัด นิคซ์ซี่ไซน์ แอนด์ [illegible]							
10/9/2554 IV 11090580	4	DRG1EQ052N10200 เทปOPP กาวยาง Nuvo สีชา (48mm x 45y)	ม้วน	SO-201	360	18.50	6,660.00
16/2/2555 IV 12020007	4	DRG1EQ052N10200 เทปOPP กาวยาง Nuvo สีชา (48mm x 50y)	ม้วน	SO-201	360	21.00	7,560.00
2 Item(s) Of ห้างหุ้นส่วนจำกัด นิคซ์ซี่ไซน์ แอนด์ [illegible]					720		14,220.00

รูปที่ 3.6 รูปตารางข้อมูลลูกค้าที่สั่งซื้อสินค้า

โดยในรูปที่ 3.6 รูปตารางข้อมูลลูกค้าที่สั่งซื้อสินค้านั้น เป็นรูปตารางข้อมูลแบบเก่าซึ่งยังไม่ได้มีการเปลี่ยนแปลงข้อมูลเพื่อความเหมาะสม และข้อมูลนี้ไม่เหมาะสมในการนำไปใช้ซึ่งจะต้องทำการการแปลงข้อมูลให้เหมาะสมกับรูปแบบโมเดลและอัลกอริทึมที่จะทำการคำนวน

ตารางที่ 3.3 รายละเอียดของตารางข้อมูลลูกค้า

ชื่อ	**ความหมาย**
รหัสลูกค้า	หมายเลขรหัสลูกค้า มีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อลูกค้า	ชื่อลูกค้าที่มีการซื้อสินค้ากับบริษัท เช่น บุคคล,ห้างหุ้นส่วน,บริษัท,และนิติบุคคล
ที่อยู่	วันที่ลูกค้าทำการซื้อสินค้า
โทรศัพท์	เลขที่ใบสั่งซื้อสินค้า
โทรสาร	คลังจัดเก็บสินค้าเพื่อจำหน่าย
อีเมล	ระยะเวลาในการชำระเงิน
วงเงินเครดิต	เลขรหัสสินค้ามีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อสินค้า	รายละเอียดสินค้า
ยอดลูกหนี้	หน่วยนับของสินค้า
จำนวนวันเครดิต	เลขที่การสั่งซื้อสินค้า

รหัส รหัสลูกค้า	ชื่อกลุ่ม ชื่อลูกค้า	ที่อยู่	โทรศัพท์	โทรสาร	[illegible]
0	เงินสด				
1	ห้างหุ้นส่วนจำกัด[illegible] (ประเทศไทย)	1/817 หมู่ 17 ต.คูคต อ.ลำลูกกา จ.ปทุมธานี			
2	บริษัท[illegible] จำกัด	1/817 หมู่ 17 ต.คูคต อ.ลำลูกกา จ.ปทุมธานี 12130	02-9938848	02-99388[illegible]	
3	บริษัท [illegible] ([illegible]) จำกัด	1/817 หมู่ 17 ต.คูคต อ.ลำลูกกา จ.ปทุมธานี	02-917-4611		
101001	บริษัท[illegible] จำกัด	44/14 หมู่ที่ 9 ตำบล[illegible] อำเภอปากเกร็ด จังหวัดนนทบุรี 11120	02-5847085-7	02-5847[illegible]	
101002	ร้าน[illegible]	111/181 หมู่บ้าน[illegible] ซ.บางบอน 12 ถ.[illegible] 1 แขวง[illegible] เขต[illegible]	02-8023226	02-80[illegible]	
1010023	หจก.[illegible]	44/63-65 ถ.[illegible] แขวง[illegible] เขต[illegible] กทม.	02-8640578-80		
101003	[illegible]	บางแค กทม.			
101004	บริษัท[illegible] จำกัด	21/122 อาคาร[illegible] ซอย[illegible] ถนน[illegible] 9 แขวง[illegible] เขตห้วย[illegible]	02-2060373	081-[illegible]	
101005	บริษัท[illegible] เทรดดิ้ง จำกัด	365/248 ม.8 แขวง[illegible] เขต[illegible] กทม. 10170	02-8898167	02-88[illegible]	
101006	บริษัท [illegible] จำกัด	[illegible] อาคาร2 เลขที่ 133 ม.1 ต.[illegible] อ.สามพราน จ.นครปฐม	02-889-4646-51	02-889-[illegible]	
101008	บริษัท[illegible] จำกัด	802/280 ถ.[illegible] ต.คูคต อ.ลำลูกกา จ.ปทุมธานี 12130	02-7338866[illegible]	02-73[illegible]	
101010	บริษัท[illegible] จำกัด	23/4 หมู่ที่ 10 ตำบล[illegible] อำเภอสามพราน จังหวัดนครปฐม 73110	034-322151,081-8315235	034-327519	
101012	[illegible]	[illegible] 22 กทม.			
101014	[illegible]	58/100 [illegible] 3 ต.บางพูด อ.ปากเกร็ด จ.นนทบุรี	02-9829338		
101015	บริษัท [illegible] จำกัด	371/9-10 หมู่ 8 ถ.[illegible] ต.[illegible] อ.เมืองสมุทรปราการ จ.สมุทรปราการ	02-7019271-3	02-70[illegible]	
101016	[illegible]	21/602 ม.12 ซ.[illegible] ถ.[illegible] กม.2 บางนา กทม.	02-7490931-2		
101017A	[illegible]	102 ซ.[illegible] 4 ถ.[illegible] แขวง[illegible] เขต[illegible] กทม. 10700	02-4353652	02-4[illegible]	
101017B	บริษัท [illegible] จำกัด	342 ซ.[illegible] 27 แขวง[illegible] เขต[illegible] กทม. 10150	02-417-7515-7/089-171-3835	02-417-75[illegible]	
101018A	บริษัท[illegible] จำกัด	401/1 หมู่ที่ 8 ถนน[illegible] ตำบลคูคต อำเภอลำลูกกา จังหวัดปทุมธานี 12130	02-9925821#2051	02-992[illegible]	
101018B	บริษัท[illegible] จำกัด	401 หมู่ที่ 8 ถนน[illegible] ตำบลคูคต อำเภอลำลูกกา จังหวัดปทุมธานี 12130	02-9925800		

รูปที่ 3.7 ข้อมูลลูกค้าที่ยังไม่ได้ทำการแปลงข้อมูล

- การทำความสะอาดข้อมูล โดยเริ่มจากการตรวจสอบข้อมูลที่ขาดหายไปในแต่ละคอลัมน์ จากข้อมูลที่ได้รับมาพบว่าในคอลัมน์บางคอลัมน์มีการขาดหายของข้อมูล เช่น คอลัมน์อีเมลล์ เบอร์โทรศัพท์ ซึ่งคอลัมน์นี้ไม่มีความหมายในการจะนำมาวิเคราะห์ จึงสามารถตัดทิ้งทั้งคอลัมน์ได้ ส่วนคอลัมน์ที่จะใช้ในการวิเคราะห์ของตารางข้อมูลนั้นไม่พบข้อมูลที่ขาดหายไป เนื่องจากระบบไม่อนุญาตให้มีค่าว่างเกิดขึ้น เช่น ชื่อลูกค้า ที่อยู่ลูกค้า เป็นต้น และตรวจสอบความผิดปกติของข้อมูลว่ามีค่าที่ไม่เป็นไปได้หรือไม่ ซึ่งพบว่าค่าทุกค่าของข้อมูลนั้นถูกต้อง

- การแปลงข้อมูล เป็นการทำให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมโดยอาศัยเงื่อนไข เพื่อให้ข้อมูลสอดคล้องกับการวิเคราะห์ ได้แก่ ที่อยู่ลูกค้า เปลี่ยนคุณสมบัติโดยแบ่งออกเป็น 6 กลุ่ม ได้แก่ ภาคกลาง, ภาคเหนือ, ภาคอีสาน, ภาคใต้, ภาคตะวันออก และกรุงเทพฯ (กรุงเทพฯจะรวมจังหวัดปริมณฑลด้วย) และ ข้อมูลการซื้อสินค้าเป็นจำนวนเงิน โดยทำการรวมจำนวนเงินที่ลูกค้าซื้อสินค้าต่อเดือน โดยใส่ข้อมูลให้อยู่ในแต่ละช่องของลูกค้าอย่างเหมาะสม และทำการแยกกลุ่มสินค้า หรือประเภทสินค้า ออกเป็นกลุ่มใหญ่ๆ เช่น กลุ่มเทปผ้ากาว กลุ่มเทปกาวน้ำ กลุ่มเทปกาวใส กลุ่มเทปกระดาษ เป็นต้น ตามรูปภาพ 3.5 ตารางข้อมูลที่ผ่านการทำความสะอาดและแปลงข้อมูลอยู่ในรูปแบบที่เหมาะสม

- การรวบรวมข้อมูล แต่ละตารางข้อมูลนั้นมีการเชื่อมโยงความสัมพันธ์กันแต่มีการถูกจัดแบ่งมาออกเป็นสองตาราง ดังนั้น จึงจำเป็นที่จะต้องนำข้อมูลทั้งสองตารางมารวมกัน และทำการตัดข้อมูลบางส่วนที่ไม่มีความจำเป็นในการวิเคราะห์ออก และทำการรวมข้อมูลใหม่ให้อยู่ในตารางเดียว เพื่อที่จะสามารถนำเข้าไปทำการวิเคราะห์ได้อย่างมีประสิทธิภาพ

ชื่อลูกค้า	ชื่อสินค้า													จังหวัด	ภาค	ประเภทลูกค้า
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	สติ๊กเกอร์	[illegible]	[illegible]	[illegible]	[illegible]			
[illegible]	L	L	L	0	0	0	0	0	L	L	0	0	0	พะเยา	1	3
[illegible]	L	L	L	0	0	L	0	0	0	L	0	0	0	[illegible]	2	4
[illegible]	L	L	L	L	0	0	0	0	L	L	0	0	L	สุโขทัย	1	4
[illegible]	M	L	0	L	0	0	0	0	L	M	0	0	L	สุพรรณบุรี	2	3
[illegible]	0	0	0	0	0	0	0	0	M	M	0	0	L	พะเยา	1	4
[illegible]	M	0	0	L	L	0	0	0	0	M	L	0	0	สุพรรณบุรี	2	4
[illegible]	L	0	0	L	0	L	0	L	M	L	0	0	0	สุพรรณบุรี	2	4
[illegible]														กรุงเทพ	2	1
[illegible]	H	0	0	0	0	0	0	0	0	0	0	0	0	อยุธยา	2	1
[illegible]	H	M	M	L	L	L	0	L	L	L	L	0	L	สุโขทัย	1	1
[illegible]	H	M	M	M	L	L	0	L	M	M	0	L	0	สุโขทัย	1	1
[illegible]	0	0	0	M	0	0	0	0	0	0	0	0	0	กรุงเทพ	2	1
[illegible]	0	0	0	0	0	0	0	0	M	M	0	0	0	กำแพงเพชร	1	4
[illegible]	L	0	0	H	0	0	H	0	0	0	0	0	0	ชลบุรี	4	1
[illegible]	0	L	0	H	0	M	0	0	0	L	0	0	0	กรุงเทพ	2	1

< 1000 LOW = L

> 1000 - 5000 Mim = M > 5000 ขึ้นไป High = H

รูปที่ 3.8 ตารางข้อมูลที่ผ่านการทำความสะอาดและแปลงข้อมูลอยู่ในรูปแบบที่เหมาะสม

ในส่วนรูปที่ 3.8 ตารางข้อมูลที่ผ่านการทำความสะอาดและแปลงข้อมูลอยู่ในรูปแบบที่เหมาะสม โดยเราจะทำแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมที่พร้อมจะนำไปใช้ในการวิเคราะห์ในเทคนิคการแบ่งกลุ่ม เป็นต้น

กรอบส่วนสีน้ำเงิน เป็นส่วนของการนำข้อมูลลูกค้ามาจัดเรียงใหม่ โดยนำข้อมูลลูกค้าที่เป็นรายชื่อลูกค้าเท่านั้นมาจัดเรียง

กรอบส่วนสีแดง เป็นส่วนของการนำข้อมูลสินค้ามาจัดเรียง โดยนำข้อมูลสินค้าที่แบ่งออกเป็นกลุ่ม แต่ละประเภทมากำหนดตามรูปแบบการซื้อสินค้าของลูกค้า

กรอบส่วนสีเขียว เป็นส่วนของการนำข้อมูลลูกค้า เช่น จังหวัด ภาค และรูปแบบลูกค้าโดยกำหนดเป็นภาค

เทคนิคและอัลกอริทึม ที่ใช้ในการวิเคราะห์

ในการวิเคราะห์เพื่อหาคำตอบนี้ ได้ใช้เทคนิค การแบ่งกลุ่มฐานข้อมูล อาจใช้การจัดกลุ่ม (Clustering) ซึ่งเป็นวิธีกำหนดกลุ่มหรือแยกกลุ่มประเภทที่มีความเหมือนหรือแตกต่างกัน การแบ่งฐานข้อมูลนี้มักมีประโยชน์ในด้านการส่งเสริมการขายอีกทางหนึ่งด้วย เป็นลักษณะ การ ทำ

unsupervise learning คือ เรา จะยังไม่ทราบเป้าหมายชัดเจนของผลลัพธ์ แต่เรา ต้องการที่แบ่งกลุ่มข้อมูลเหล่านั้นออกมา เพื่อ ที่จะนำมาอนุมาน ให้เกิด ประโยชน์ ต่อไป หรือ ถ้ามอง ในเชิงธุกิจง่ายๆ อาจจะนำมาแบ่งกลุ่มลูกค้า เพื่อ การตัดสินใจทำ อะไรบางอย่างเช่น ส่ง จดหมาย ให้ลูกค้าว่าควรจะส่งให้กลุ่มใด

รูปที่ 3.9 รูปตัวอย่างข้อมูลในเทคนิค K Mean

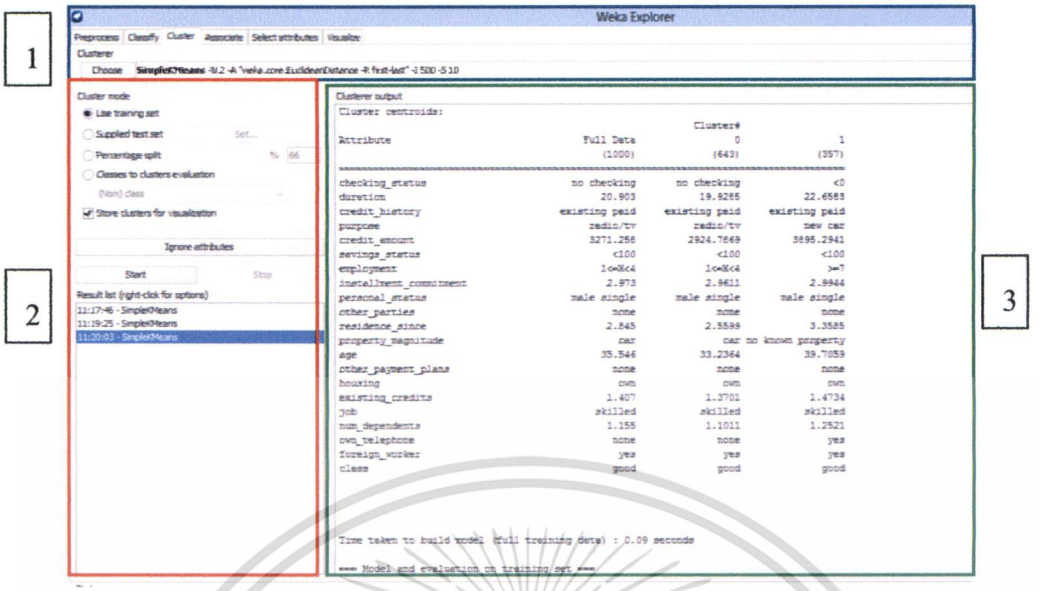
รูปภาพที่ 3.9 รูปภาพตัวอย่างการนำเข้าข้อมูลที่จะนำใช้ในการทำดาต้าไมน์นิ่งด้วยเทคนิค K Mean ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆตามรูปภาพ โดยในรูปภาพหน้าต่างการใช้งานโปรแกรม WEKA นี้ประกอบด้วย 4 ส่วนหลักๆ ดังนี้

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล

2. ส่วนแสดงผลข้อมูล ในแต่ละแอตทริบิวต์ ว่ามีรูปแบบข้อมูลเป็นแบบใดและข้อมูลประกอบด้วยจำนวนเท่าใด

3. ส่วนแสดงข้อมูลทั้งหมด ว่ามีข้อมูลใดประกอบบ้าง ซึ่งเราสามารถเลือกข้อมูลที่ต้องการ และสามารถเลือกข้อมูลที่ไม่ต้องการและสามารถนำข้อมูลที่ไม่ต้องการออกได้

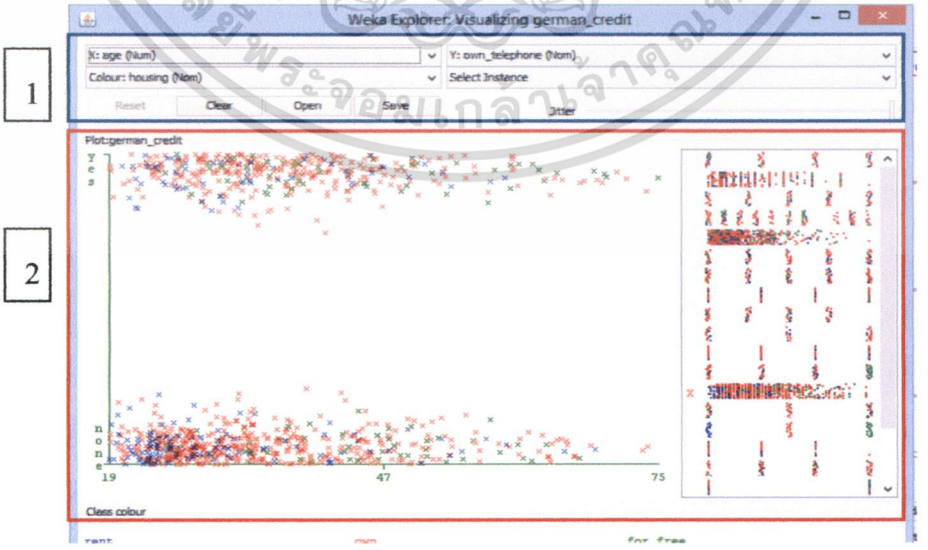
4. ส่วนแสดงผลเป็นรูปภาพ จะสามารถเห็นข้อมูลโดยรวมทั้งหมดให้เป็นรูปแบบรูปภาพ



รูปที่ 3.10 รูปตัวอย่างการประมวลผลข้อมูลในเทคนิค K Mean

รูปภาพที่ 3.10 รูปภาพตัวอย่างการประมวลผลลัพธ์ข้อมูลที่ได้ด้วยเทคนิค K Mean ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆตามรูปภาพ โดยในรูปภาพนี้ประกอบด้วย 3 ส่วน ได้แก่

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล
2. ส่วนแสดงผลในการเลือกโมเดลและอัลกอริทึมที่ใช้
3. ส่วนแสดงผลลัพธ์ที่ได้จากการคำนวณ โดยในรูปภาพจะแสดงข้อมูลผลลัพธ์ที่ได้จากการคำนวนโดยใช้อัลกอริทึมเลือกและแสดงผลลัพธ์ตามที่ต้องการ



รูปที่ 3.11 รูปแสดงผลข้อมูลแบบในเทคนิค K Mean

รูปภาพที่ 3.11 รูปภาพตัวอย่างที่เป็นแบบ Visualizing ของข้อมูลที่ได้ด้วยเทคนิค K Mean ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆตามรูปภาพ โดยแบ่งออกเป็น 2 ส่วน คือ

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบสี และรูปแบบผลลัพธ์ของข้อมูล
2. ส่วนแสดงผลข้อมูลในการเลือกโมเดลและอัลกอริทึมที่ใช้ ซึ่งจะเห็นว่าข้อมูลมีการแบ่งออกเป็น 2 กลุ่มใหญ่ ตามรูปภาพ

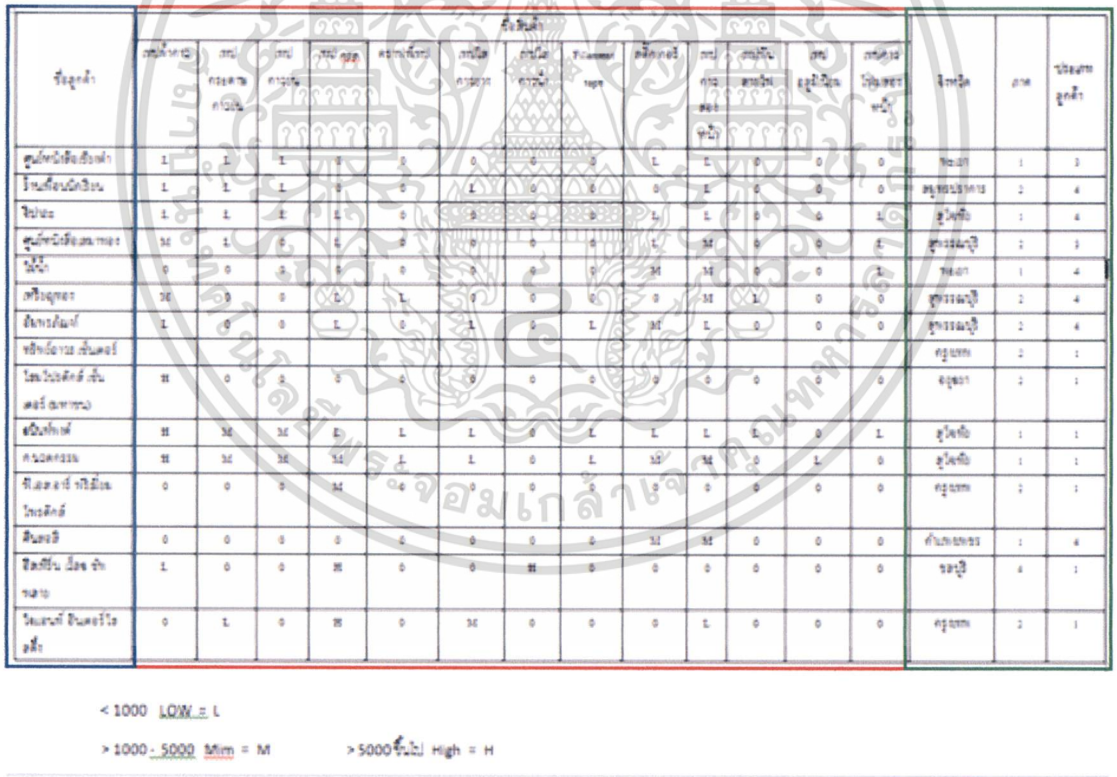
3.3.3 ทำนายยอดปริมาณสินค้าในการสั่งซื้อล่วงหน้า

- การคัดเลือกข้อมูล ในกระบวนการนี้จะคัดเลือกโดยการนำตารางการขายสินค้าในแต่ละเดือนโดยทำการแยกการซื้อสินค้าของลูกค้าแต่ละรายออก โดยมีผลรวมจำนวนเงินที่ขายได้ในแต่ละวัน/เดือน และข้อมูลรายละเอียดการขายสินค้า เพื่อทำการดูแนวโน้มการขายสินค้าสินค้าดังตารางที่ 3.3

ตารางที่ 3.4 รายละเอียดของตารางการขายสินค้าที่ใช้สำหรับทำนายยอดขายสินค้า

ชื่อ	ความหมาย
รหัส	หมายเลขรหัสลูกค้า มีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อลูกค้า	ชื่อลูกค้าที่มีการซื้อสินค้ากับบริษัท โดยมี บุคคล ,ห้างหุ้นส่วน,บริษัท,และนิติบุคคล
วันที่	วันที่ลูกค้าทำการซื้อสินค้า
เลขที่	เลขที่ใบสั่งซื้อสินค้า
คลัง	คลังจัดเก็บสินค้าเพื่อจำหน่าย
เครดิต	ระยะเวลาในการชำระเงิน
รหัสสินค้า	เลขรหัสสินค้ามีทั้งตัวเลขและตัวหนังสือภาษาอังกฤษ
ชื่อสินค้า	รายละเอียดสินค้า
หน่วย	หน่วยนับของสินค้า
อ้างอิง	เลขที่การสั่งซื้อสินค้า
ปริมาณ	จำนวนสินค้า
ราคา/หน่วย	ราคาสินค้า
ลด	ส่วนลด
เงินบาท	จำนวนเงินรวมการซื้อสินค้า

- การทำความสะอาดข้อมูล โดยเริ่มจากการตรวจสอบข้อมูลที่ขาดหายไปในแต่ละคอลัมน์ จากระบบพบว่าในคอลัมน์นั้นไม่พบข้อมูลที่ขาดหายไปในส่วนที่จำเป็นต้องนำมาวิเคราะห์ข้อมูล ส่วนบางคอลัมน์นั้นมีบางส่วนที่ขาดหายแต่ไม่ได้นำมาวิเคราะห์ จึงไม่จำเป็นต้องทำการแก้ไขหรือเพิ่มเติมข้อมูล ซึ่งสามารถตัดคอลัมน์นั้นทิ้งไปได้ และตรวจสอบความผิดปกติของข้อมูลว่ามีค่าที่ไม่เป็นไปได้หรือไม่ ซึ่งพบว่าคาทุกค่าของข้อมูลนั้นถูกต้อง เนื่องจากเป็นการป้อนข้อมูลโดยเลือกข้อมูลที่มีอยู่ในระบบเท่านั้น ไม่ได้มีการป้อนข้อมูลเอง
- การแปลงข้อมูล เป็นการทำให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมโดยอาศัยเงื่อนไขเพื่อให้ข้อมูลสอดคล้องกับการวิเคราะห์ สำหรับข้อมูลบางส่วนนั้นรูปแบบของข้อมูลสามารถนำไปใช้งานได้ทันที และส่วนสำหรับข้อมูลจำนวนยอดเงินรวมในแต่ละวัน/เดือน จะทำการแปลงข้อมูลหรือเพิ่มข้อมูลนั้นขึ้นมาเพื่อให้สามารถทำการวิเคราะห์ข้อมูลได้อย่างถูกต้อง
- การรวบรวมข้อมูล จากข้อมูลข้างต้นเป็นข้อมูลที่มาจากตารางการขายสินค้า ซึ่งมีบางส่วนของข้อมูลที่ไม่มีความจำเป็นต้องนำมาใช้ จึงสามารถทำการตัดทิ้ง เหลือแต่ส่วนที่จำเป็นในการวิเคราะห์ได้ ดังรูปภาพ3.6 ตัวอย่างข้อมูลสำหรับทำนายยอดการขายสินค้า



< 1000 LOW = L

> 1000 - 5000 Mim = M > 5000ขึ้นไป High = H

รูปที่ 3.12 ตัวอย่างข้อมูลสำหรับทำนายยอดการขายสินค้า

รูปภาพที่ 3.12 รูปภาพตัวอย่างข้อมูลที่จะใช้ในการเข้าประมวลผลด้วยเทคนิค Classify เพื่อที่เอาไว้ใช้ในการทำนายผลที่ผ่านการทำความสะอาดและแปลงข้อมูลอยู่ในรูปแบบที่เหมาะสมโดยเราจะทำแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมที่พร้อมจะนำไปใช้ในการวิเคราะห์ในเทคนิค Classify

กรอบส่วนสีน้ำเงิน เป็นส่วนของการนำข้อมูลลูกค้ามาจัดเรียงใหม่ โดยนำข้อมูลลูกค้าที่เป็นรายชื่อลูกค้าเท่านั้นมาจัดเรียง

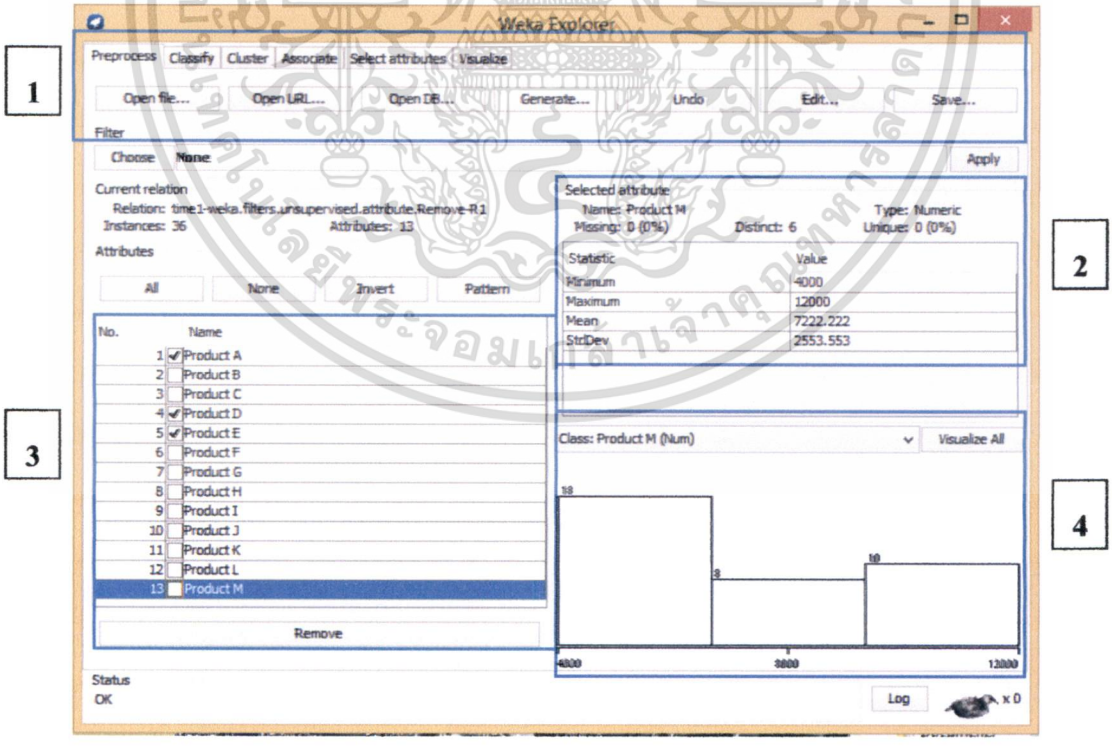
กรอบส่วนสีแดง เป็นส่วนของการนำข้อมูลสินค้ามาจัดเรียง โดยนำข้อมูลสินค้าที่แบ่งออกเป็นกลุ่ม แต่ละประเภทมากำหนดตามรูปแบบการซื้อสินค้าของลูกค้า

กรอบส่วนสีเขียว เป็นส่วนของการนำข้อมูลลูกค้า เช่น จังหวัด ภาค และรูปแบบลูกค้าโดยกำหนดเป็นภาค

เทคนิคและอัลกอริทึม ที่ใช้ในการวิเคราะห์

ในการทำการวิเคราะห์ เนื่องจากผลลัพธ์ที่ได้จากการใช้วิธีนี้ จะสามารถตีความหมายของผลการพยากรณ์ได้ง่ายและยังสามารถทำความเข้าใจกระบวนการใช้งานได้ง่าย อีกทั้งยังสามารถหาสาเหตุที่มาที่ไปของผลลัพธ์ได้

รูปแบบตัวอย่างการใช้เทคนิค Neuron Network



รูปที่ 3.13 รูปตัวอย่างข้อมูลในเทคนิค Neuron Network

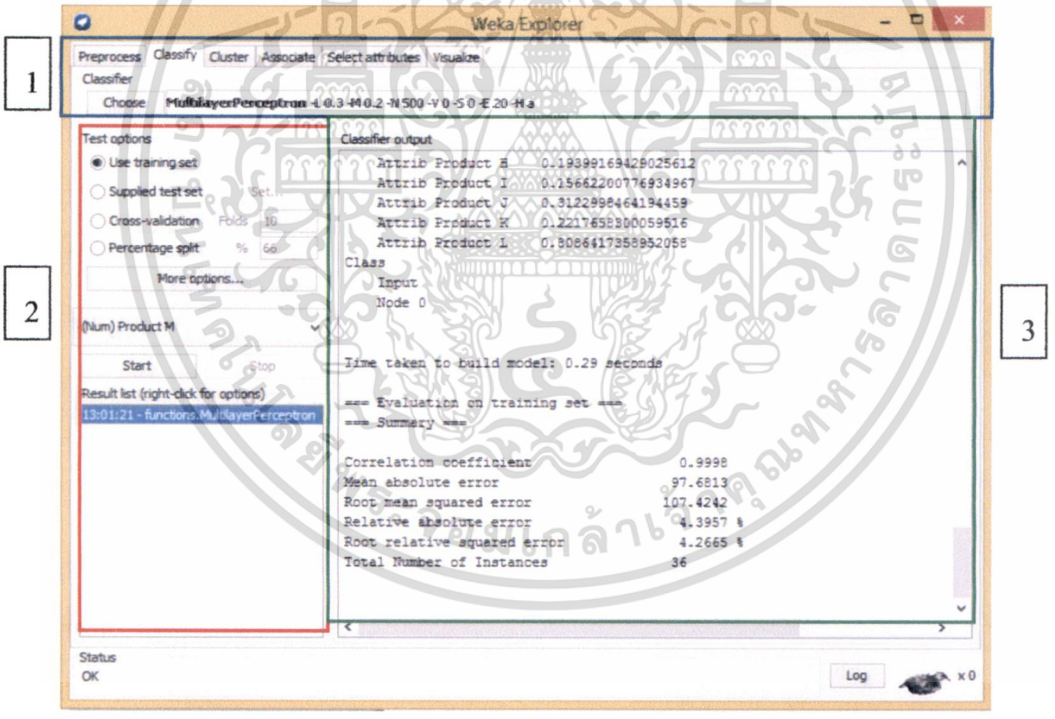
รูปภาพที่ 3.13 รูปภาพตัวอย่างการนำเข้าข้อมูลที่จะนำใช้ในการทำดาต้าไมน์นิ่งด้วยเทคนิค Neuron Network ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆตามรูปภาพ โดยในรูปภาพหน้าต่างการใช้งานโปรแกรม WEKA นี้ประกอบด้วย 4 ส่วนหลักๆ ดังนี้

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล

2. ส่วนแสดงผลข้อมูล ในแต่ละแอตทริบิวต์ ว่ามีรูปแบบข้อมูลเป็นแบบใดและข้อมูลประกอบด้วยจำนวนเท่าใด

3. ส่วนแสดงข้อมูลทั้งหมด ว่ามีข้อมูลใดประกอบบ้าง ซึ่งเราสามารถเลือกข้อมูลที่ต้องการ และสามารถเลือกข้อมูลที่ไม่ต้องการและสามารถนำข้อมูลที่ไม่ต้องการออกได้

4. ส่วนแสดงผลเป็นรูปภาพ จะสามารถเห็นข้อมูลโดยรวมทั้งหมดให้เป็นรูปแบบรูปภาพ

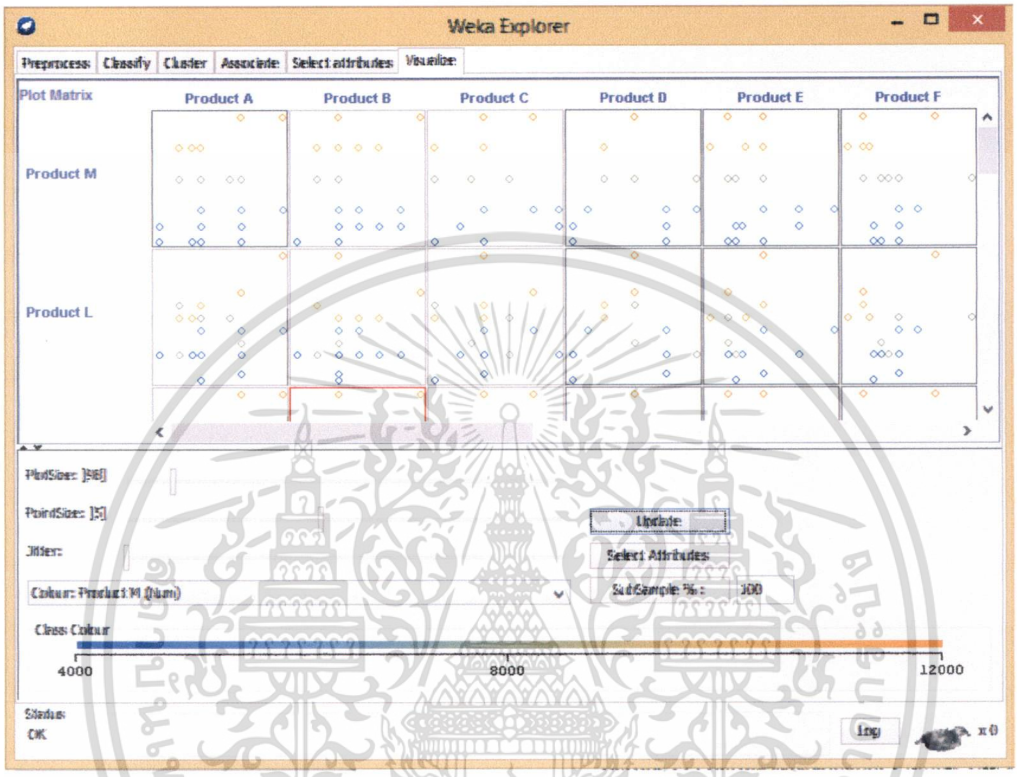


รูปที่ 3.14 รูปตัวอย่างการประมวลผลข้อมูลในเทคนิค Neuron Network

รูปภาพที่ 3.14 รูปภาพตัวอย่างข้อมูลผลลัพธ์ ด้วยเทคนิค Neuron Network ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆตามรูปภาพ โดยในรูปภาพนี้ประกอบด้วย 3 ส่วน ได้แก่

1. ส่วนของเมนูและฟังก์ชั่นการใช้งาน โดยเป็นส่วนการเลือกข้อมูลการนำเข้า และเลือกรูปแบบโมเดล และอัลกอริทึม ที่จะนำเข้ามาในการประมวลผล

2. ส่วนแสดงผลในการเลือกโมเดลและอัลกอริทึมที่ใช้
3. ส่วนแสดงผลลัพธ์ที่ได้จากการคำนวณ โดยในรูปภาพจะแสดงข้อมูลผลลัพธ์ที่ได้จากการคำนวนโดยใช้อัลกอริทึมเลือกและแสดงผลลัพธ์ตามที่ต้องการ



รูปที่ 3.15 รูปแสดงผลข้อมูลแบบในเทคนิค Neuron Network

รูปภาพที่ 3.15 รูปภาพตัวอย่างที่เป็นแบบ Classifier Tree Visualizer ด้วยเทคนิค Neuron Network ซึ่งข้อมูลจะแสดงผลแต่ละผลิตภัณฑ์ โดยแบ่งออกเป็นโหนดๆ

3.4 สรุปท้ายบท

จากบทที่ 3 นี้เราสามารถสรุปได้ว่า การเตรียมข้อมูล และการแปลงข้อมูลนั้น จะพบว่าจะต้องต้องหาแหล่งข้อมูลที่มีความพร้อมและถูกต้องของข้อมูล รวมถึงจะต้องแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม และเหมาะกับเทคนิค และอัลกอริทึมที่จะใช้ โดยที่เทคนิคและอัลกอริทึมที่ใช้ในการคำนวณนั้น จะมีความแตกต่างในการเก็บข้อมูลที่แตกต่างกันไป โดยข้อมูลที่จัดเก็บนั้น จะแบ่งออกเป็น 2 ประเภทใหญ่ๆ ได้แก่

1. ข้อมูลที่เป็นตัวหนังสือ โดยข้อมูลที่จัดเก็บในรูปแบบนี้ ส่วนมากจะนำมาใช้กับข้อมูลเพื่อหาความสัมพันธ์ เกี่ยวข้องกัน เช่น การทำ Association Rule เป็นต้น

2. ข้อมูลที่เป็นตัวเลข โดยข้อมูลที่จัดเก็บในรูปแบบนี้ เทคนิคและอัลกอริทึมที่เราใช้จะสนใจข้อมูลที่เป็นตัวเลข หรือข้อมูลเกี่ยวเนื่องกัน โดยเทคนิคและอัลกอริทึมที่ใช้ข้อมูลตัวเลขนั้น จะเป็น การทำ Classify เป็นต้น

โดยในบทนี้สามารถสรุปใจความสำคัญได้ว่า ขั้นตอนการทำ ดาต้าไมน์นิ่งนั้นเราจะต้องคำนึงหลักในการทำงาน 3 ส่วน ด้วยกัน ได้แก่

1. การรวบรวมข้อมูล
2. การเลือกเทคนิค และอัลกอริทึมที่จะใช้ในการหาผลลัพธ์
3. การแปลงข้อมูล และการเติมข้อมูลในส่วนที่ขาดหาย โดยคำนึงถึงความสอดคล้องของข้อมูล หรือการตัดข้อมูลที่ไม่มีความจำเป็น หรือเกี่ยวข้องในการทำงาน

บทที่ 4
ผลการดำเนินงาน

ผลการดำเนินงานของการวิเคราะห์ความสัมพันธ์ของเทปกาวและพฤติกรรมของผู้บริโภค นี้ได้แบ่งผลการดำเนินงานออกเป็น3ส่วนแยกตามเทคนิคการทำเหมืองข้อมูลคือขั้นตอนการนำเข้าระบบผลการดำเนินงานและผลการทดสอบเพื่อวัดประสิทธิภาพของระบบ ซึ่งมีรายละเอียดดังต่อไปนี้

4.1 กฎความสัมพันธ์ (Association Rule)

4.1.1 ขั้นตอนการนำเข้าระบบ

4.1.2 ผลการดำเนินงาน

4.1.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

4.2 การแบ่งกลุ่ม (Clustering)

4.2.1ขั้นตอนการนำเข้าระบบ

4.2.2 ผลการดำเนินงาน

4.2.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

4.3 การทำนายผล

4.3.1 ขั้นตอนการนำเข้าระบบ

4.3.2 ผลการดำเนินงาน

4.3.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

4.1 กฎความสัมพันธ์ (Association Rule)

ในรูปแบบนี้ต้องการวิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้าได้แก่“ลูกค้าส่วนใหญ่เมื่อซื้อสินค้าอย่างหนึ่งแล้วมักจะซื้อสินค้าคู่กับสินค้าอะไร จึงจำเป็นต้องใช้กฎความสัมพันธ์ในการทำดาต้าไมน์นิ่ง เพื่อหาคำตอบที่เหมาะสม โดยกฎความสัมพันธ์นี้ เราจะใช้ อัลกอริทึม 2 แบบ ได้แก่

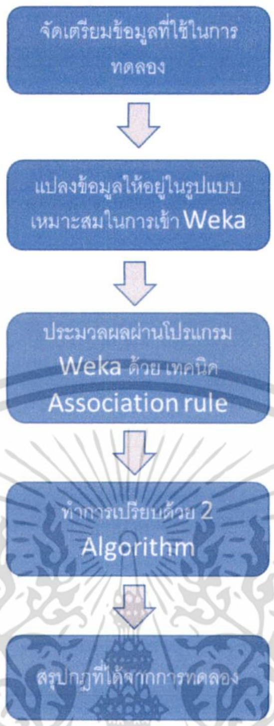
1. APIORI
2. FP GROWTH

4.1.1 ขั้นตอนการนำเข้าระบบ

การแปลงข้อมูล

จากบทที่แล้ว ได้แสดงข้อมูลที่จะนำมาใช้ในการทดสอบ ซึ่งข้อมูลที่ได้จะประกอบด้วยข้อมูลหลากหลายแบบ ซึ่งบางข้อมูลสามารถนำมาใช้ได้ แต่บางข้อมูลไม่มีความจำเป็นที่นำเข้ามาใช้ในการทดสอบ เราจึงจำเป็นต้องมีการแปลงข้อมูล หรือการเลือกนำข้อมูลที่มีความเหมาะ

ถูกต้องที่จะใช้ในการทดสอบเพื่อในการหาคำตอบ และให้สอดคล้องกับรูปแบบในแต่ละเทคนิคของการทำดาต้าไมน์นิ่ง



รูปภาพที่ 4.1 ลำดับขั้นตอนการทำงานด้วยเทคนิค Association Rule

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Name	Product A	Product B	Product C	Product D	Product E	Product F	Product G	Product H	Product I	Product J	Product K	Product L	Product M
2	บี.แอล.สเตชั่นเนอรี่	Y	Y	Y	Y	?	?	?	?	?	Y	?	?	Y
3	[illegible] เทรดดิ้ง	Y	?	?	Y	?	Y	?	?	Y	Y	?	?	Y
4	บริษัท [illegible]สเตชั่นเนอรี่	?	?	?	Y	?	?	?	?	?	?	?	?	?
5	สุวรรณสาส์น	?	?	?	?	?	?	?	?	Y	Y	?	?	?
6	บริษัท ก.จักรวาล (2008)	Y	Y	Y	Y	?	Y	Y	?	Y	Y	?	?	Y
7	รุ่งทรัพย์เปเปอร์	Y	?	?	?	?	?	?	?	Y	Y	?	?	?
8	ร้าน[illegible]เทรดดิ้ง	Y	?	?	Y	?	?	?	?	Y	Y	?	?	Y
9	ห้างหุ้นส่วนจำกัด[illegible]	Y	?	?	?	?	?	?	?	Y	?	?	?	?
10	ร้านช้าง 2	Y	?	Y	?	?	?	?	?	?	?	?	Y	Y
11	ร้านสหกรณ์โรงพยาบาล[illegible]	?	?	?	Y	?	?	?	?	Y	Y	?	?	?
12	เอส วาย เครื่องเขียน	Y	?	Y	Y	?	?	?	?	?	Y	?	?	?
13	สุวรรณภูมิ	?	?	Y	?	?	?	?	?	Y	Y	?	?	?
14	วิทยาภัณฑ์	?	Y	?	?	?	?	?	?	?	?	?	?	?
15	ร้านโรจนาภัณฑ์	?	?	?	?	?	?	?	?	?	Y	?	?	?
16	ฉลองเครื่องเขียน	Y	Y	?	Y	?	Y	?	?	Y	?	?	?	?
17	ห้างหุ้นส่วนจำกัด[illegible]	Y	Y	Y	Y	?	?	?	Y	Y	Y	?	?	Y
18	[illegible]	Y	Y	?	?	?	?	Y	?	Y	?	?	?	Y
19	บริษัท หาดใหญ่[illegible]	?	?	?	Y	?	?	?	?	?	?	?	?	?
20	อาร์.เค.เทรดดิ้ง	Y	?	?	?	?	?	?	?	?	Y	?	?	?
21	ห้างหุ้นส่วนจำกัด[illegible] ซัพพลาย	?	?	?	?	?	?	?	?	Y	?	?	?	?
22	กันเองเครื่องเขียน	?	Y	?	Y	?	?	?	?	Y	Y	Y	Y	?
23	บริษัท [illegible]พาณิชย์	Y	?	Y	?	?	?	?	?	?	?	?	?	Y

รูปภาพที่ 4.2 การแปลงข้อมูลในการเข้าโปรแกรม Weka ด้วยเทคนิค Association Rule

ในรูปภาพที่ 4.2 ภาพนี้เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมจะเข้าโปรแกรม Weka ในการทำเทคนิค Association Rule โดยประกอบข้อมูล 3 ส่วน ได้แก่

1. ข้อมูลรายชื่อร้านค้าหรือข้อมูลลูกค้า
2. ข้อมูลผลิตภัณฑ์สินค้าในการจำหน่าย โดยแบ่งเป็นกลุ่มใหญ่ๆของสินค้า
3. ข้อมูลการขายสินค้า

โดยค่าข้อมูลที่ใส่ประกอบด้วย 2 รูปแบบ คือ

ค่า Y คือค่าที่ลูกค้าได้ซื้อสินค้าประเภทนั้น

ค่า ? คือค่าที่ลูกค้าไม่มีการซื้อสินค้าประเภทนั้น

ตารางที่ 4.1 รายชื่อประเภทสินค้า

Name	รายชื่อลูกค้า
Product A	สินค้าประเภทเทปผ้ากาว
Product B	สินค้าประเภทเทปกระดาษกาวย่น
Product C	สินค้าประเภทเทปกาวย่น
Product D	สินค้าประเภทเทป OPP
Product E	สินค้าประเภทคราฟท์เทป
Product F	สินค้าประเภทเทปใสกาวยาง
Product G	สินค้าประเภทเทปใสกาวน้ำ
Product H	สินค้าประเภทFilament Tape
Product I	สินค้าประเภทสติ๊กเกอร์
Product J	สินค้าประเภทเทปกาวสองหน้า
Product K	สินค้าประเภทเทปพันสายไฟ
Product L	สินค้าประเภทเทปอลูมิเนียม
Product M	สินค้าประเภทเทปกาวโฟมสองหน้า

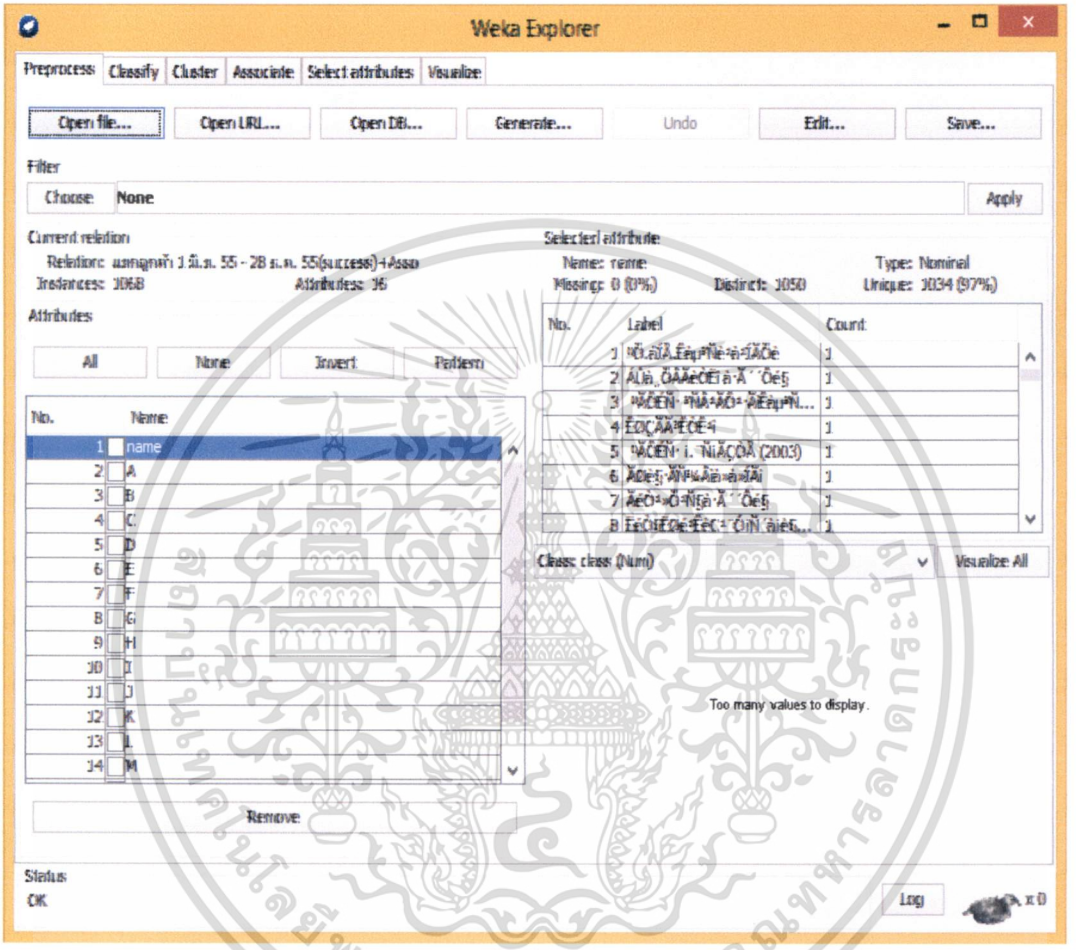
โดยค่าข้อมูลที่ใส่ประกอบด้วย 2 รูปแบบ คือ

a) ค่า Y คือค่าที่ลูกค้าได้ซื้อสินค้าประเภทนั้น

b) ค่า ? คือค่าที่ลูกค้าไม่มีการซื้อสินค้าประเภทนั้น

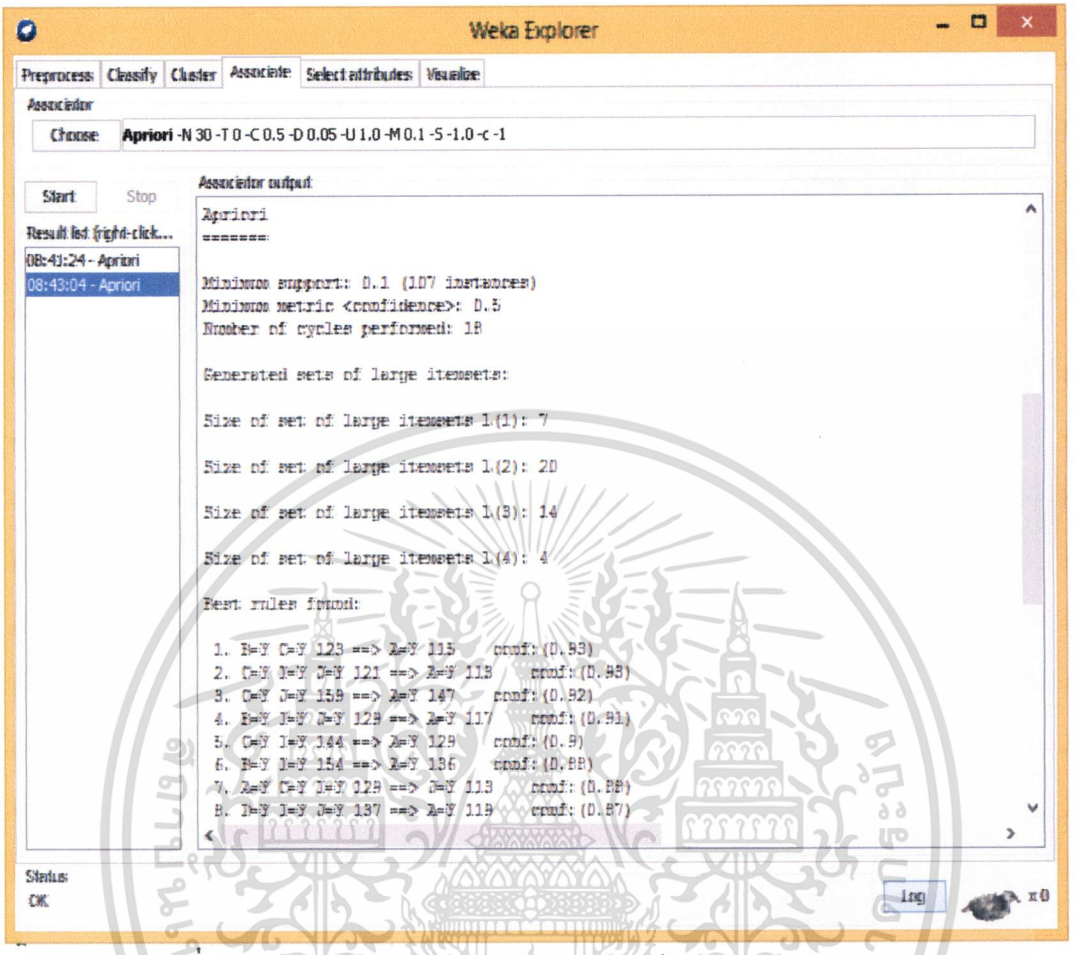
การนำเข้าข้อมูลในโปรแกรม Weka

เนื่องจากรูปแบบข้อมูลของซอฟต์แวร์ที่ใช้ในการประมวลผลยอมรับได้นั้นจะต้องอยู่ในรูปไฟล์นามสกุล .arff หรือ .crv ดังนั้นผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในรูปแบบ .crv เพื่อความเหมาะสมในการนำเข้าข้อมูล และสามารถแก้ไขข้อมูล ดังภาพที่ (ข้างบน)



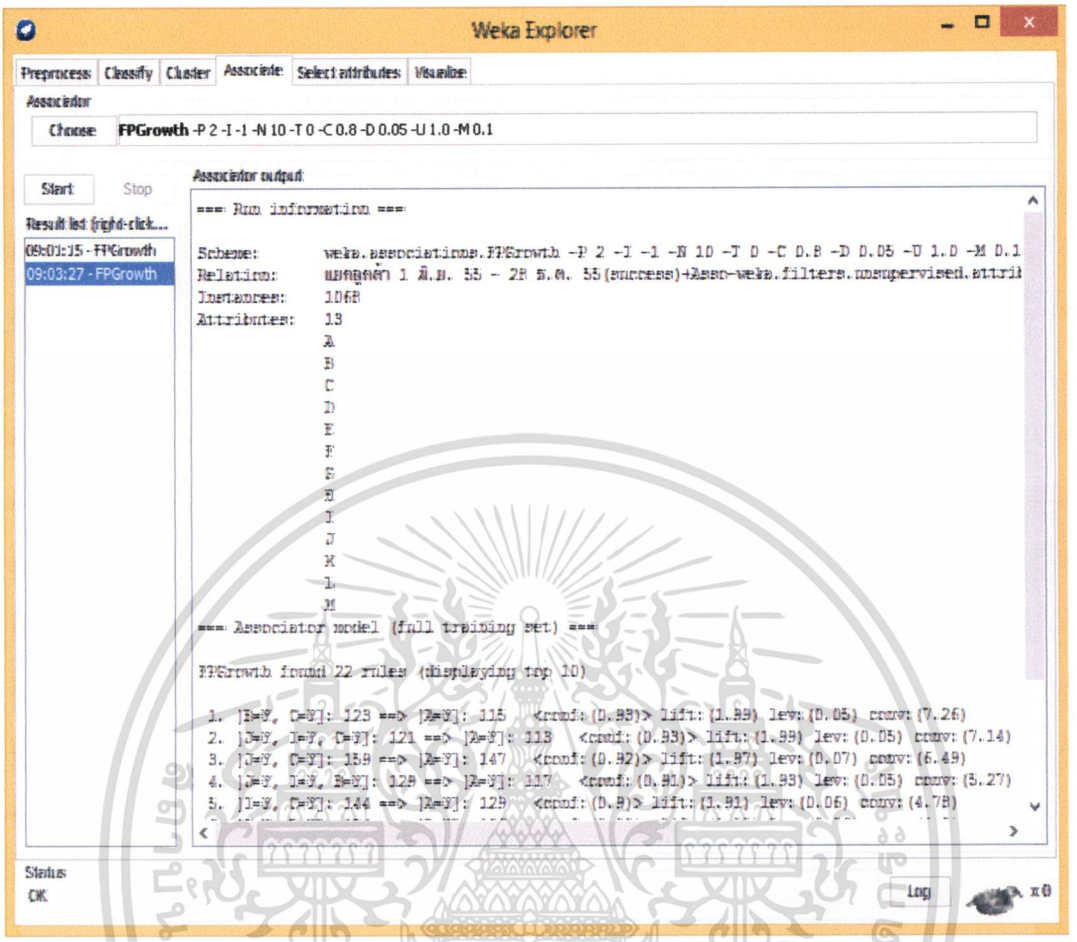
รูปภาพที่ 4.3 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Association Rule

รูปภาพที่ 4.3 รูปภาพการนำเข้าข้อมูลเข้าสู่โปรแกรม Weka ด้วยเทคนิค Association Rule ซึ่งข้อมูลจะแบ่งออกเป็นกลุ่มๆของสินค้าแต่ละชนิดตามรูปภาพ



รูปภาพที่ 4.4 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Association Rule โดยใช้อัลกอริทึม Apiori

รูปภาพที่ 4.4 รูปภาพแสดงผลลัพธ์ ด้วยเทคนิค Association Rule โดยใช้อัลกอริทึม Apiori ซึ่งผลลัพธ์ที่ได้จะมีการบอกเปอร์เซ็นต์ความน่าเชื่อถือในการแสดงผล



รูปภาพที่ 4.5 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Association Rule โดยใช้อัลกอริทึม FP Growth

รูปภาพที่ 4.5 รูปภาพแสดงผลลัพธ์ ด้วยเทคนิค Association Rule โดยใช้อัลกอริทึม FP Growth ซึ่งผลลัพธ์ที่ได้จะมีการบอกเปอร์เซ็นต์ความน่าเชื่อถือในการแสดงผล

4.1.2 ผลการดำเนินงาน

ในการทดสอบนี้ในกฎการความสัมพันธ์ เราจะใช้ 2 อัลกอริทึม

อัลกอริทึม Apiori

จากการทดสอบจะสรุปได้ว่า กฎที่มีค่าความเชื่อมั่นสูงกว่าค่าความเชื่อมั่นขั้นต่ำ 0.8 และมีค่าสนับสนุนสูงกว่าค่าสนับสนุนขั้นต่ำ 0.1 มีจำนวน 22 กฎ โดยจะได้ผลลัพธ์ข้อมูลดังต่อไปนี้

```
=== Run information ===
Scheme:     weka.associations.Apriori -N 30 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:   แยกลูกค้า 1 มิ.ย. 55 - 28 ธ.ค. 55(success)+Asso-
weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R14-15
Instances:  1068
```

Attributes: 13

Product A

Product B

Product C

Product D

Product E

Product F

Product G

Product H

Product I

Product J

Product K

Product L

Product M

=== Associator model (full training set) ===

Apriori

Minimum support: 0.1 (107 instances)

Minimum metric <confidence>: 0.8

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 20

Size of set of large itemsets L(3): 14

Size of set of large itemsets L(4): 4

Best rules found:

1. B=Y C=Y 123 ==> A=Y 115 conf:(0.93)
2. C=Y I=Y J=Y 121 ==> A=Y 113 conf:(0.93)
3. C=Y J=Y 159 ==> A=Y 147 conf:(0.92)
4. B=Y I=Y J=Y 129 ==> A=Y 117 conf:(0.91)
5. C=Y I=Y 144 ==> A=Y 129 conf:(0.9)
6. B=Y I=Y 154 ==> A=Y 136 conf:(0.88)
7. A=Y C=Y I=Y 129 ==> J=Y 113 conf:(0.88)
8. D=Y I=Y J=Y 137 ==> A=Y 119 conf:(0.87)
9. C=Y 207 ==> A=Y 179 conf:(0.86)
10. A=Y B=Y I=Y 136 ==> J=Y 117 conf:(0.86)
11. D=Y I=Y 168 ==> A=Y 143 conf:(0.85)
12. J=Y M=Y 134 ==> A=Y 114 conf:(0.85)
13. B=Y D=Y J=Y 147 ==> A=Y 124 conf:(0.84)
14. C=Y I=Y 144 ==> J=Y 121 conf:(0.84)
15. B=Y I=Y 154 ==> J=Y 129 conf:(0.84)

16. B=Y J=Y 197 ==> A=Y 164 conf:(0.83)
17. A=Y D=Y I=Y 143 ==> J=Y 119 conf:(0.83)
18. I=Y J=Y 234 ==> A=Y 194 conf:(0.83)
19. A=Y B=Y D=Y 150 ==> J=Y 124 conf:(0.83)
20. A=Y C=Y 179 ==> J=Y 147 conf:(0.82)
21. D=Y I=Y 168 ==> J=Y 137 conf:(0.82)
22. M=Y 196 ==> A=Y 157 conf:(0.8)

FP GROWTH

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1

Relation: แยกลูกค้า 1 มิ.ย. 55 - 28 ธ.ค. 55(success)+Asso-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R14-15

Instances: 1068

Attributes: 13

Product A
Product B
Product C
Product D
Product E
Product F
Product G
Product H
Product I
Product J
Product K
Product L
Product M

=== Associator model (full training set) ===

FPGrowth found 22 rules (displaying top 10)

1. [B=Y, C=Y]: 123 ==> [A=Y]: 115 <conf:(0.93)> lift:(1.99) lev:(0.05) conv:(7.26)
2. [J=Y, I=Y, C=Y]: 121 ==> [A=Y]: 113 <conf:(0.93)> lift:(1.99) lev:(0.05) conv:(7.14)
3. [J=Y, C=Y]: 159 ==> [A=Y]: 147 <conf:(0.92)> lift:(1.97) lev:(0.07) conv:(6.49)
4. [J=Y, I=Y, B=Y]: 129 ==> [A=Y]: 117 <conf:(0.91)> lift:(1.93) lev:(0.05) conv:(5.27)
5. [I=Y, C=Y]: 144 ==> [A=Y]: 129 <conf:(0.9)> lift:(1.91) lev:(0.06) conv:(4.78)
6. [I=Y, B=Y]: 154 ==> [A=Y]: 136 <conf:(0.88)> lift:(1.88) lev:(0.06) conv:(4.3)
7. [A=Y, I=Y, C=Y]: 129 ==> [J=Y]: 113 <conf:(0.88)> lift:(2.23) lev:(0.06) conv:(4.61)
8. [D=Y, J=Y, I=Y]: 137 ==> [A=Y]: 119 <conf:(0.87)> lift:(1.85) lev:(0.05) conv:(3.83)
9. [C=Y]: 207 ==> [A=Y]: 179 <conf:(0.86)> lift:(1.84) lev:(0.08) conv:(3.79)
10. [A=Y, I=Y, B=Y]: 136 ==> [J=Y]: 117 <conf:(0.86)> lift:(2.19) lev:(0.06) conv:(4.13)

ทำการเปรียบเทียบ 2 อัลกอริทึม และ สรุปผล

ตารางที่ 4.2 ทำการเปรียบเทียบ 2 อัลกอริทึม และ สรุปผล

Apiori	FP Growth
1. B=Y C=Y 123 ==> A=Y 115 conf:(0.93)	1. [B=Y, C=Y]: 123 ==> [A=Y]: 115 <conf:(0.93)
2. C=Y I=Y J=Y 121 ==> A=Y 113 conf:(0.93)	2. [J=Y, I=Y, C=Y]: 121 ==> [A=Y]: 113 <conf:(0.93)>
3. C=Y J=Y 159 ==> A=Y 147 conf:(0.92)	3. [J=Y, C=Y]: 159 ==> [A=Y]: 147 <conf:(0.92)>
4. B=Y I=Y J=Y 129 ==> A=Y 117 conf:(0.91)	4. [J=Y, I=Y, B=Y]: 129 ==> [A=Y]: 117 <conf:(0.91)>
5. C=Y I=Y 144 ==> A=Y 129 conf:(0.9)	5. [I=Y, C=Y]: 144 ==> [A=Y]: 129 <conf:(0.9)>
6. B=Y I=Y 154 ==> A=Y 136 conf:(0.88)	6. [I=Y, B=Y]: 154 ==> [A=Y]: 136 <conf:(0.88)>
7. A=Y C=Y I=Y 129 ==> J=Y 113 conf:(0.88)	7. [A=Y, I=Y, C=Y]: 129 ==> [J=Y]: 113 <conf:(0.88)>
8. D=Y I=Y J=Y 137 ==> A=Y 119 conf:(0.87)	8. [D=Y, J=Y, I=Y]: 137 ==> [A=Y]: 119 <conf:(0.87)>
9. C=Y 207 ==> A=Y 179 conf:(0.86)	9. [D=Y, J=Y, I=Y]: 137 ==> [A=Y]: 119 <conf:(0.87)>
10. A=Y B=Y I=Y 136 ==> J=Y 117 conf:(0.86)	10. [A=Y, I=Y, B=Y]: 136 ==> [J=Y]: 117 <conf:(0.86)>

4.1.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

จากผลการทดสอบ โดยผ่านการประมวลผลทั้ง 2 อัลกอริทึม คือ Apiori และ FP Growth นั้น โดยผลค่าการทดสอบที่ได้มานั้นใน 10 ค่าที่มีค่า confidence มากกว่า 0.8 โดยทั้งสองอัลกอริทึมนั้นได้ผลลัพธ์มาเหมือนกัน และมีค่า เปอร์เซ็นต์ความน่าเชื่อถือนั้น เหมือนกัน ซึ่งผู้วิจัยสรุปได้ว่าข้อมูลที่นำเข้าสู่โปรแกรม Weka ในการทดสอบนั้น สามารถให้ผลในการทดสอบ และค่าความ

น่าเชื่อถือที่สอดคล้องกัน จึงแสดงให้เห็นว่าข้อมูลในการนำเข้ามีความน่าเชื่อถือและเหมาะสมในการหาค่าของกฎ ความสัมพันธ์ และสามมารถตอบโจทย์ที่เราตั้งคำถาม ได้อย่างเหมาะสม ตัวอย่างการแปลผลลัพธ์ที่ได้จากหัวข้อกรณีศึกษา 4 ตัวอย่าง

1. B=Y C=Y 123 ==> A=Y 115 conf:(0.93)

2. C=Y I=Y J=Y 121 ==> A=Y 113 conf:(0.93)

3. C=Y J=Y 159 ==> A=Y 147 conf:(0.92)

4. B=Y I=Y J=Y 129 ==> A=Y 117 conf:(0.91)

1. ถ้าลูกค้าซื้อสินค้า B (เทปกระดาษกาวย่น) และซื้อสินค้าชนิด C(เทปกาวย่น) ลูกค้าจะซื้อสินค้าชนิด A(เทปผ้ากาว) โดยมีค่าความเชื่อมั่นถึง 0.93 หรือ 93%

2. ถ้าลูกค้าซื้อสินค้า C(เทปกาวย่น) ซื้อสินค้าชนิด I(สติ๊กเกอร์) และซื้อสินค้าชนิด J(เทปกาวสองหน้า) ลูกค้าจะซื้อสินค้าชนิด A(เทปผ้ากาว) โดยมีค่าความเชื่อมั่นถึง 0.93 หรือ 93% เช่นกัน

3. ถ้าลูกค้าซื้อสินค้า C(เทปกาวย่น) และซื้อสินค้าชนิด J(เทปกาวสองหน้า) ลูกค้าจะซื้อสินค้าชนิด A(เทปผ้ากาว) โดยมีค่าความเชื่อมั่นถึง 0.92 หรือ 92% เช่นกัน

4. ถ้าลูกค้าซื้อสินค้า B(เทปกระดาษกาวย่น) ซื้อสินค้าชนิด I(สติ๊กเกอร์) และซื้อสินค้าชนิด J(เทปกาวสองหน้า) ลูกค้าจะซื้อสินค้าชนิด A(เทปผ้ากาว) โดยมีค่าความเชื่อมั่นถึง 0.91 หรือ 91% เช่นกัน เป็นต้น

4.2 การแบ่งกลุ่ม (Clustering)

ในรูปแบบนี้เราต้องการวิเคราะห์กลุ่มข้อมูลลูกค้าได้แก่“วิเคราะห์พฤติกรรมลูกค้ากับการซื้อสินค้า (แยกตามภูมิศาสตร์ และรูปแบบองค์กรลูกค้า)” จึงจำเป็นต้องใช้การแบ่งกลุ่มข้อมูลในการทำดาต้าไมน์นิ่ง เพื่อหาคำตอบที่เหมาะสม โดยกฎความสัมการแบ่งกลุ่มข้อมูลเราต้องการทราบความสัมพันธ์ของข้อมูล ซึ่งผู้วิจัยจะใช้ อัลกอริทึม 2 แบบ ได้แก่

1. K means
2. Hierachicalcluster

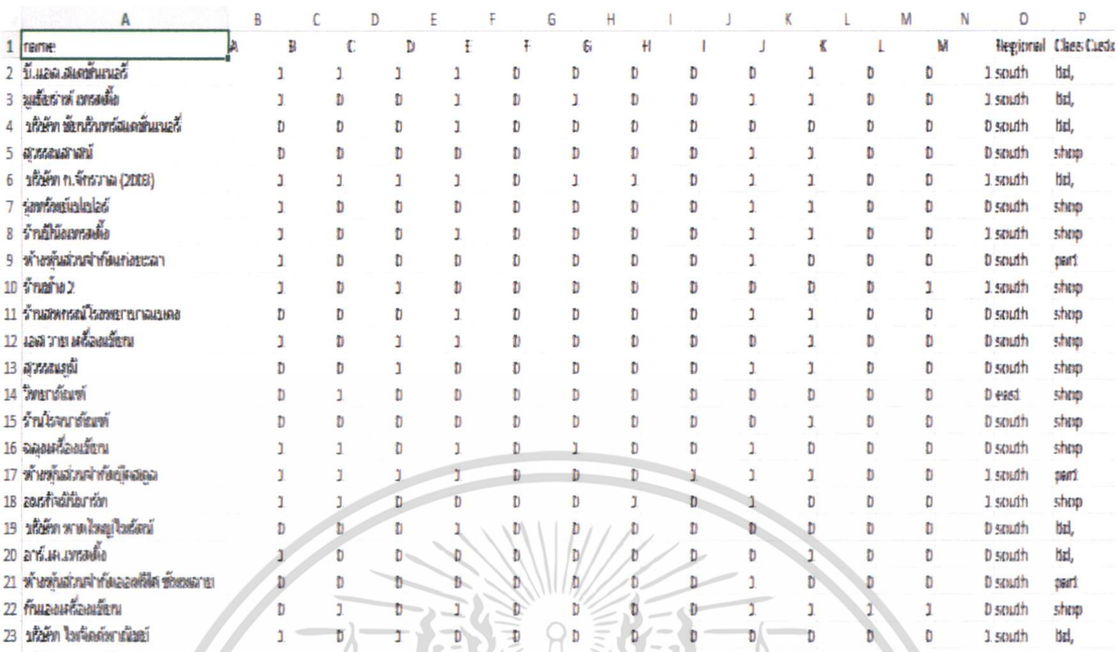
4.2.1 ขั้นตอนการนำเข้าระบบ

การแปลงข้อมูล

รูปแบบข้อมูลที่จะนำมาใช้ในการทดสอบในการแบ่งกลุ่มข้อมูลนี้ ซึ่งรูปแบบข้อมูลที่จะนำมาใช้งานจะประกอบด้วยข้อมูลหลากหลายแบบ ซึ่งบางข้อมูลสามารถนำมาใช้ได้ แต่บางข้อมูลไม่มีความจำเป็นที่นำเข้ามาใช้ในการทดสอบ เราจึงจำเป็นต้องมีการแปลงข้อมูล หรือการเลือกนำข้อมูลที่มีความเหมาะสม และถูกต้องที่จะใช้ในการทดสอบเพื่อในการหาคำตอบ และให้สอดคล้อง

กับรูปแบบในแต่ละเทคนิคของการทำคาต้าไมน์นิ่ง ซึ่งผู้วิจัยได้กำหนดกลุ่มข้อมูลตามลักษณะภูมิศาสตร์ของประเทศไทย โดยเราสามารถแบ่งกลุ่มข้อมูลออกมา ได้ 6 กลุ่มข้อมูล และผู้วิจัยได้ทำการแบ่งกลุ่มข้อมูลลูกค้า ซึ่งเราแบ่งกลุ่มข้อมูลประเภทธุรกิจลูกค้า ได้ 4 รูปแบบ

รูปภาพที่ 4.6 ลำดับขั้นตอนการทำงานด้วยเทคนิค Cluster



	name	A	B	C	D	E	F	G	H	I	J	K	L	M	Regional	Class Custo
2	[illegible]	1	1	1	1	0	0	0	0	0	1	0	0	1	south	ltd,
3	[illegible]	1	0	0	1	0	1	0	0	1	1	0	0	1	south	ltd,
4	[illegible]	0	0	0	1	0	0	0	0	0	0	0	0	0	south	ltd,
5	[illegible]	0	0	0	0	0	0	0	0	1	1	0	0	0	south	shop
6	บริษัท ก.จักรวาล (2008)	1	1	1	1	0	1	1	0	1	1	0	0	1	south	ltd,
7	[illegible]	1	0	0	0	0	0	0	0	1	1	0	0	0	south	shop
8	[illegible]	1	0	0	1	0	0	0	0	1	1	0	0	1	south	shop
9	[illegible]	1	0	0	0	0	0	0	0	1	0	0	0	0	south	part
10	[illegible]	1	0	1	0	0	0	0	0	0	0	0	1	1	south	shop
11	[illegible]	0	0	0	1	0	0	0	0	1	1	0	0	0	south	shop
12	[illegible]	1	0	1	1	0	0	0	0	0	1	0	0	0	south	shop
13	[illegible]	0	0	1	0	0	0	0	0	1	1	0	0	0	south	shop
14	[illegible]	0	1	0	0	0	0	0	0	0	0	0	0	0	east	shop
15	[illegible]	0	0	0	0	0	0	0	0	0	1	0	0	0	south	shop
16	[illegible]	1	1	0	1	0	1	0	0	1	0	0	0	0	south	shop
17	[illegible]	1	1	1	1	0	0	0	1	1	1	0	0	1	south	part
18	[illegible]	1	1	0	0	0	0	1	0	1	0	0	0	1	south	shop
19	[illegible]	0	0	0	1	0	0	0	0	0	0	0	0	0	south	ltd,
20	[illegible]	1	0	0	0	0	0	0	0	0	1	0	0	0	south	ltd,
21	[illegible]	0	0	0	0	0	0	0	0	1	0	0	0	0	south	part
22	[illegible]	0	1	0	1	0	0	0	0	1	1	1	1	0	south	shop
23	[illegible]	1	0	1	0	0	0	0	0	0	0	0	0	1	south	ltd,

รูปภาพที่ 4.7 การแปลงข้อมูลในการเข้าโปรแกรม Weka ด้วยเทคนิค Cluster

ในรูปภาพที่ 4.7 ภาพนี้เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมจะเข้าโปรแกรม Weka ในการทำเทคนิค Cluster โดยประกอบข้อมูล 4 ส่วน ได้แก่

1. ข้อมูลรายชื่อร้านค้าหรือข้อมูลลูกค้า
2. ข้อมูลผลิตภัณฑ์สินค้าในการจำหน่าย โดยแบ่งเป็นกลุ่มใหญ่ๆของสินค้า
3. ข้อมูลการขายสินค้า

   โดยค่าข้อมูลที่ใส่ประกอบด้วย 2 รูปแบบ คือ

   ค่า 1 คือค่าที่ลูกค้าได้ซื้อสินค้าประเภทนั้น

   ค่า 0 คือค่าที่ลูกค้าไม่มีการซื้อสินค้าประเภทนั้น
4. ข้อมูลที่ตั้งภูมิศาสตร์ของลูกค้า และข้อมูลประเภทลูกค้า

ตารางที่ 4.3 ตารางข้อมูลในการแบ่งกลุ่มข้อมูลพื้นที่กับประเภทธุรกิจของลูกค้า

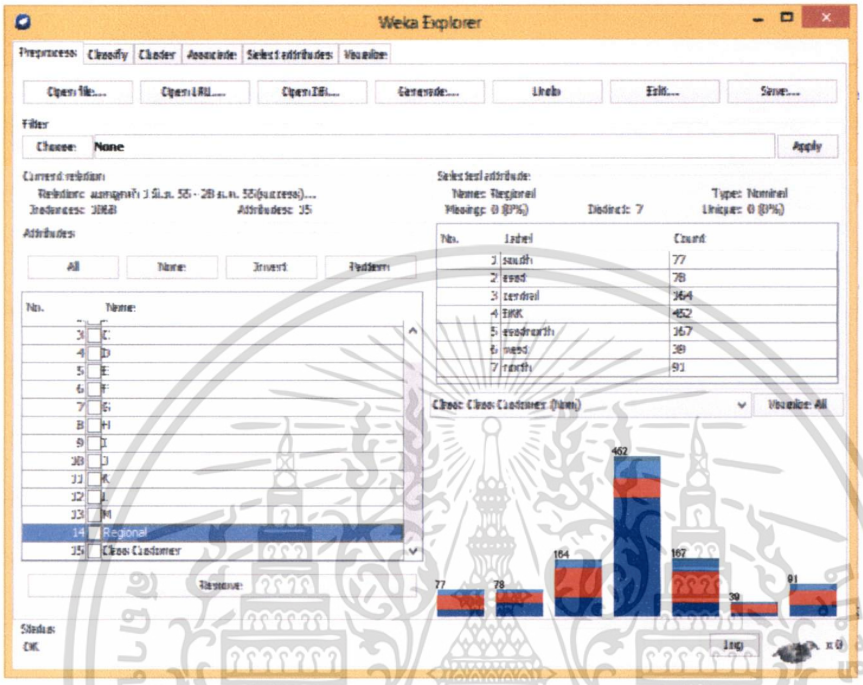
Name	รายชื่อลูกค้า
Product A	สินค้าประเภทเทปผ้ากาว
Product B	สินค้าประเภทเทปกระดาษกาวย่น
Product C	สินค้าประเภทเทปกาวย่น
Product D	สินค้าประเภทเทป OPP
Product E	สินค้าประเภทคราฟท์เทป
Product F	สินค้าประเภทเทปใสกาวยาง
Product G	สินค้าประเภทเทปใสกาวน้ำ
Product H	สินค้าประเภทFilament Tape
Product I	สินค้าประเภทสติ๊กเกอร์
Product J	สินค้าประเภทเทปกาวสองหน้า
Product K	สินค้าประเภทเทปพันสายไฟ
Product L	สินค้าประเภทเทปอลูมิเนียม
Product M	สินค้าประเภทเทปกาวโฟมสองหน้า
Regional	ภูมิภาคของประเทศไทย แบ่งออกเป็น 7 กลุ่ม คือ BKK = กรุงเทพและปริมณฑล Central = ภาคกลาง East = ภาคตะวันออก Eastnorth = ภาคตะวันออกเฉียงเหนือ North = ภาคเหนือ West = ภาคตะวันตก South = ภาคใต้
Class Customer	ประเภทธุรกิจของลูกค้า แบ่งออกเป็น 4 รูปแบบ คือ Ltd, = บริษัท จำกัด Part = ห้างหุ้นส่วน จำกัด Shop = ร้านค้า Personal = บุคคลธรรมดา

โดยค่าข้อมูลที่ใส่จะประกอบด้วย 2 รูปแบบ คือ

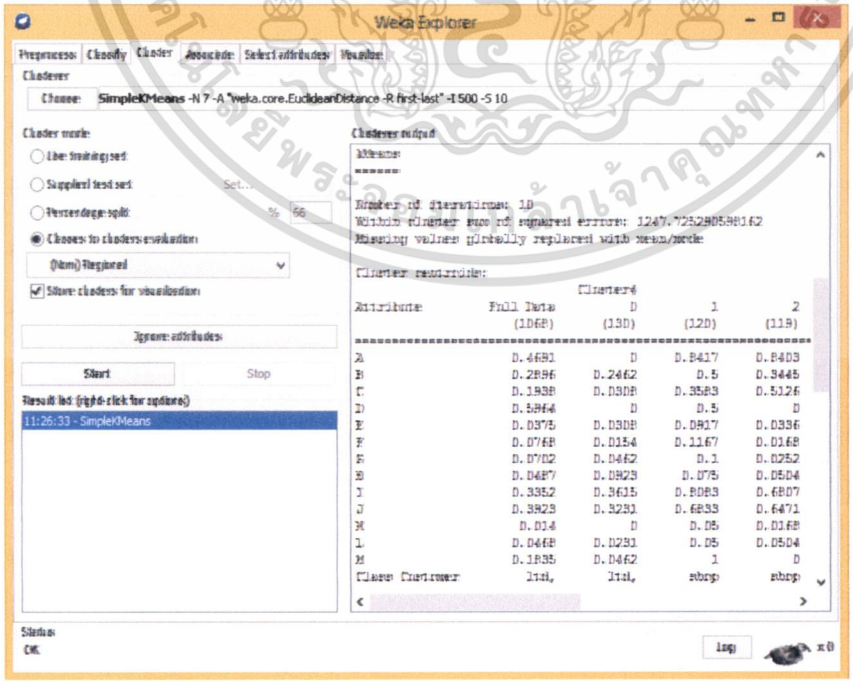
1. ค่า 1 คือค่าที่ลูกค้าได้ซื้อสินค้าประเภทนั้น
2. ค่า 0 คือค่าที่ลูกค้าไม่มีการซื้อสินค้าประเภทนั้น

3. **การนำเข้าข้อมูลในโปรแกรม Weka**

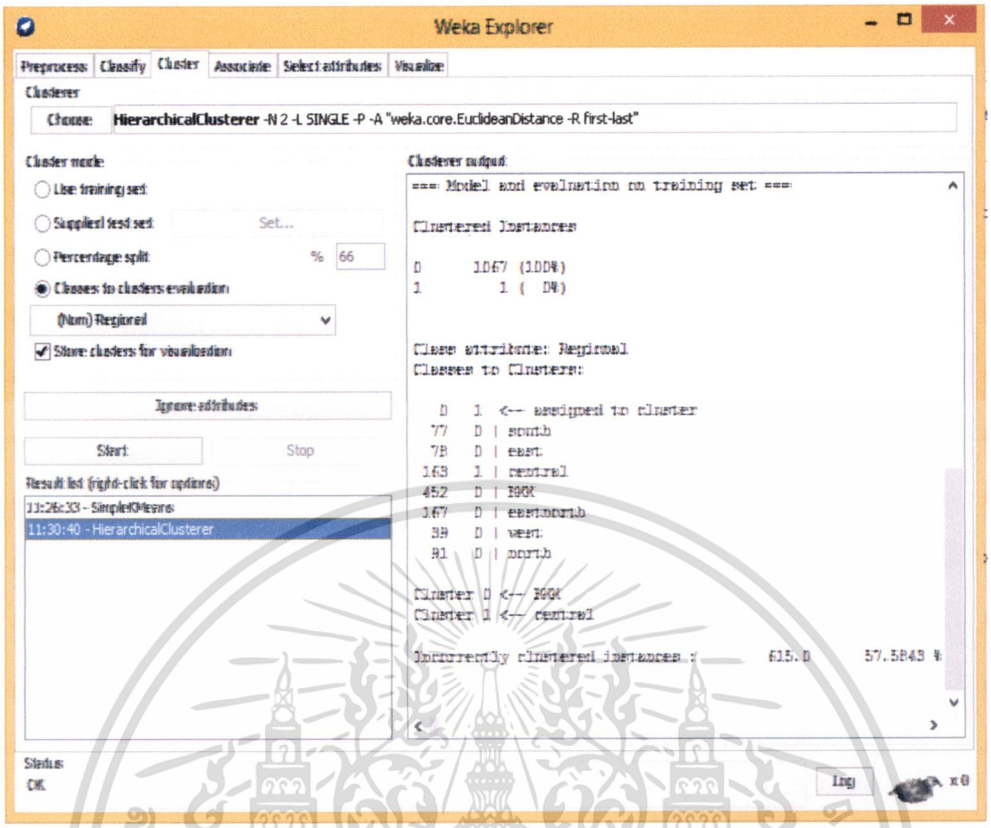
เนื่องจากรูปแบบข้อมูลของซอฟต์แวร์ที่ใช้ในการประมวลผลยอมรับได้นั้นจะต้องอยู่ในรูปไฟล์นามสกุล .arff หรือ .crv ดังนั้นผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในรูปแบบ .crv เพื่อความเหมาะสมในการนำเข้าข้อมูล และสามารถแก้ไขข้อมูล ดังภาพที่ (ข้างบน)



รูปภาพที่ 4.8 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Association Rule



รูปภาพที่ 4.9 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Clustering โดยใช้อัลกอริทึม K Means



รูปภาพที่ 4.10 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิค Clustering โดยใช้อัลกอริทึม Hierachicalcluster

4.2.2 ผลการดำเนินงาน

ในการทดสอบนี้ในกฎการความสัมพันธ์ เราจะใช้ 2 อัลกอริทึม

อัลกอริทึม K Means

จากการทดสอบโดยกำหนดตัวแปร คือ regional (ภาค) สามารถกำหนดการแบ่งกลุ่มลูกค้าได้ทั้งหมด 7 กลุ่ม เพื่อที่เราจะสามารถหาคำตอบในหัวข้อศึกษา "**วิเคราะห์พฤติกรรมลูกค้ากับการซื้อสินค้า (แยกตามภูมิศาสตร์ และรูปแบบองค์กรลูกค้า)**" โดยในหัวข้อกรณีการศึกษานี้ เราสามารถหาผลการศึกษาได้ดังนี้

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 7 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: แยกลูกค้า 1 มิ.ย. 55 - 28 ธ.ค. 55(success)+regional+class_customer-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1068

Attributes: 15

A

B

C

D

E

F

G

H

I

J

K

L

M

Class Customer

Ignored: Regional

Test mode:Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

Number of iterations: 10

Within cluster sum of squared errors: 1247.725290598162

Cluster centroids Cluster#

Incorrectly clustered instances : 683.0 63.9513 %

Attribute	Full Data	0	1	2	3	4	5	6
	(1068)	(130)	(120)	(119)	(81)	(115)	(137)	(366)
A	0.4691	0	0.8417	0.8403	0.9012	0.7826	1	0
B	0.2896	0.2462	0.5	0.3445	0.9136	0.5565	0.0751	0.0765
C	0.1938	0.0308	0.3583	0.5126	0.6049	0.3739	0.0292	0.0082
D	0.5964	0	0.5	0	0.9012	1	0.1679	1
E	0.0375	0.0308	0.0917	0.0336	0.1111	0.0348	0.0073	0.0191
F	0.0768	0.0154	0.1167	0.0168	0.3951	0.1565	0.0438	0.0219
G	0.0702	0.0462	0.1	0.0252	0.1852	0.0696	0.0511	0.0656
H	0.0487	0.0923	0.075	0.0504	0.0988	0.087	0.0219	0.0109
I	0.3352	0.3615	0.8083	0.6807	0.4321	0.6522	0.1387	0.0109
J	0.3923	0.3231	0.6833	0.6471	0.9012	0.8174	0.1387	0.0874
K	0.014	0	0.05	0.0168	0.0247	0.0348	0	0.0027
L	0.0468	0.0231	0.05	0.0504	0.3086	0.0261	0.0146	0.0137
M	0.1835	0.0462	1	0	0.6667	0	0.0657	0.0191
Class Customer	ltd,	ltd,	Shop	Shop	ltd,	shop	ltd,	Ltd

รูปภาพที่ 4.11 รูปภาพผลลัพธ์ที่ได้จากการใช้เทคนิคการแบ่งกลุ่มข้อมูลด้วย K Means

Hierachicalcluster

=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 7 –L MEAN –P -A

"weka.core.EuclideanDistance -R first-last"

Relation:แยกลูกค้า1มิ.ย.55-28ธ.ค.55

weka.filters.unsupervised.attribute.Remove-R1

Instances: 1068

Attributes: 15

A

B

C

D

E

F

G

H

I

J

K

L

M

Time taken to build model (full training data) : 24.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 592 (55%)

1 89 (8%)

2 60 (6%)

3 69 (6%)

4 147 (14%)

5 73 (7%)

6 38 (4%)

Incorrectly clustered instances : 641.0 60.0187 %

BKK	South	Eastnorth	East	North	Central	No Class
24	11	9	10	11	8	4
38	5	5	11	8	7	4
74	21	9	15	19	18	8
343	12	5	10	56	23	3
62	20	23	15	24	8	15
18	5	3	4	8	1	0
33	15	6	4	21	8	4

รูปภาพที่ 4.12 รูปภาพผลลัพธ์ที่ได้จากการใช้เทคนิคการแบ่งกลุ่มข้อมูลด้วย Hierachical Cluster

4.2.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

จากผลการทดสอบ โดยผ่านการประมวลผลทั้ง 2 อัลกอริทึม คือ K Means และ Hierachicalcluster นั้น โดยผลค่าการทดสอบที่ได้มานั้นใน เราสามารถทำการประเมินผลทดสอบ

และสามารถแปลค่าผลทดสอบที่ได้นั้น จะพบว่า อัลกอริทึม K Means นั้นสามารถแสดงผลการทดสอบ และแสดงค่าผลการทดสอบ ในหัวข้อที่จะทำการศึกษาได้อย่างชัดเจน และถูกต้อง โดยให้ค่าความถูกต้องในการแบ่งกลุ่ม อยู่ที่ 63.9513 % แต่ถ้าเป็นอัลกอริทึม Hierachicalcluster นั้น ถึงแม้จะให้ค่าความถูกต้องอยู่ที่ 60.0187% แต่ไม่สามารถแปลลัพธ์ รวมถึงประเมินผลในการหาคำตอบได้อย่างถูกต้องตัวอย่างการแปลผลลัพธ์ที่ได้จากหัวข้อกรณีของอัลกอริทึม K Means ในกรณีศึกษา 2 ตัวอย่าง คือ

1.ลูกค้ากลุ่มที่6 คือลูกค้าที่อยู่ในบริเวณกรุงเทพและปริมณฑล จะซื้อสินค้าประเภทเทป OPP โดยมีค่าความเชื่อมั่นคือ 1 และ ส่วนมากลูกค้าที่อยู่บริเวณกรุงเทพและปริมณฑลจะมีลักษณะรูปแบบองค์กร คือ เป็นบริษัทจำกัด เป็นต้น

2. .ลูกค้ากลุ่มที่5 คือลูกค้าที่อยู่ในบริเวณภาคตะวันตก จะซื้อสินค้าประเภทเทปผ้ากาว โดยมีค่าความเชื่อมั่นคือ 1 และ ส่วนมากลูกค้าที่อยู่บริเวณภาคตะวันตก จะมีลักษณะรูปแบบองค์กร คือ เป็นบริษัทจำกัด เป็นต้น เช่นกัน

4.3 การทำนาย (prediction)

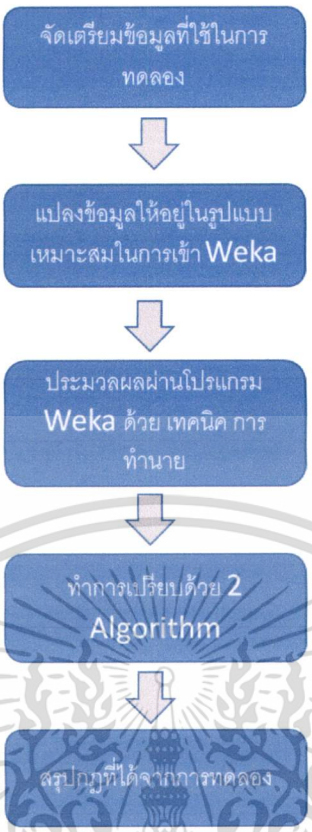
ในรูปแบบนี้เราต้องการทำนายการซื้อของลูกค้าได้แก่ "ทำนายยอดปริมาณสินค้าในการสั่งซื้อล่วงหน้า" จึงจำเป็นต้องใช้ข้อมูลย้อนหลังในการทำดาต้าไมน์นิ่ง เพื่อหาคำตอบที่เหมาะสม โดยข้อมูลที่ใช้นั้นจะเป็นข้อมูลจริงซึ่งมีระยะเวลา 12 เดือน และเป็นข้อมูลที่ได้จากการประมาณการของผู้วิจัยเอง ย้อนหลังเป็นระยะเวลา 24 เดือน โดยกฎของอนุกรมเวลา ในการหาผลลัพธ์ ซึ่งผู้วิจัยจะใช้ อัลกอริทึม 2 แบบ ได้แก่

1. Linear regression
2. Neural Network

4.3.1 ขั้นตอนการนำเข้าระบบ

การแปลงข้อมูล

รูปแบบข้อมูลที่จะนำมาใช้ในการทดสอบในการแบ่งกลุ่มข้อมูลนี้ ซึ่งรูปแบบข้อมูลที่จะนำมาใช้งานจะประกอบด้วยข้อมูลหลากหลายแบบ ซึ่งบางข้อมูลสามารถนำมาใช้ได้ แต่บางข้อมูลไม่มีความจำเป็นที่นำเข้ามาใช้ในการทดสอบ เราจึงจำเป็นต้องมีการแปลงข้อมูล หรือการเลือกนำข้อมูลที่มีความเหมาะสม และถูกต้องที่จะใช้ในการทดสอบเพื่อในการหาคำตอบ และให้สอดคล้องกับรูปแบบในแต่ละเทคนิคของการทำดาต้าไมน์นิ่ง ซึ่งผู้วิจัยทำการรวบรวมข้อมูลจำนวนสินค้าของแต่ละชนิดในการขายแต่ละเดือน จำแนกออกมาเป็นแต่ละประเภท และได้ทำการกำหนดข้อมูลให้มีระยะเวลาทั้งสิ้น 36 เดือน เพื่อที่จะสามารถมีข้อมูลเพียงพอต่อการในการทำนายผลลัพธ์



รูปภาพที่ 4.13 ลำดับขั้นตอนการทำงานด้วยเทคนิคในการทำนาย

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Month	Product A	Product B	Product C	Product D	Product E	Product F	Product G	Product H	Product I	Product J	Product K	Product L	Product M
2	ม.ค.-13	30000	7000	8000	4000	1000	1000	20000	1000	9000	15000	1500	1000	12000
3	ก.พ.-13	15000	4000	6000	3000	1500	1200	1000	1200	8000	8000	1000	800	10000
4	มี.ค.-13	20000	5000	5000	2000	1400	2000	1000	1000	6000	7000	700	500	5000
5	เม.ย.-13	10000	3000	5000	2000	1200	1300	1000	1000	5000	5000	500	500	5000
6	พ.ค.-13	10000	1000	6000	2000	1000	1500	1000	1000	5000	4000	500	500	4000
7	มิ.ย.-13	15000	2000	7000	6000	1200	1700	1000	1000	6000	4000	600	500	8000
8	ก.ค.-13	30000	6000	9000	6000	3000	2000	1400	2000	7000	4000	600	500	6000
9	ส.ค.-13	40000	4000	8000	5000	2000	2500	1800	1200	4000	6000	800	700	6000
10	ก.ย.-13	30000	3000	6000	5000	2000	2000	1500	1200	4000	5000	900	700	6000
11	ต.ค.-13	20000	3000	7000	3000	1000	2000	1000	1000	5000	8000	1000	800	8000
12	พ.ย.-13	20000	2000	6000	3000	2000	1000	1000	2000	7000	9000	1100	900	10000
13	ธ.ค.-13	40000	3000	6000	4000	2000	3000	2000	1000	8000	12000	1500	1300	12000
14	1-ธ.ค.	18000	5000	6000	3000	500	600	14000	700	6000	10000	900	800	10000
15	2-ธ.ค.	15000	3000	4000	3000	1500	1200	1000	1200	6000	9000	1000	800	10000
16	3-ธ.ค.	20000	4000	5000	2000	1400	2000	2000	1000	6000	7000	400	500	5000
17	4-ธ.ค.	20000	3000	4000	2000	1200	1300	1000	1000	2500	5000	500	300	4000
18	5-ธ.ค.	18000	3000	6000	2000	1000	1500	1000	1200	5000	4000	500	500	4000
19	6-ธ.ค.	15000	2000	7000	6000	1200	1700	1500	1000	4000	2500	600	500	8000
20	7-ธ.ค.	30000	6000	9000	5000	3000	2000	1400	2000	7000	4000	600	500	5000
21	8-ธ.ค.	20000	4000	8000	2500	4000	2500	1800	1200	3000	3000	600	700	6000
22	9-ธ.ค.	30000	3000	6000	5000	2000	2000	1500	1200	4000	5000	900	350	4000

รูปภาพที่ 4.14 การแปลงข้อมูลในการเข้าโปรแกรม Weka ด้วยเทคนิค การทำนาย

ตารางที่ 4.4 ตารางข้อมูลประเภทสินค้า

Month	เดือนและปีที่ทำการจัดเก็บข้อมูล
Product A	สินค้าประเภทเทปผ้ากาว
Product B	สินค้าประเภทเทปกระดาษกาวย่น
Product C	สินค้าประเภทเทปกาวย่น
Product D	สินค้าประเภทเทป OPP
Product E	สินค้าประเภทคราฟท์เทป
Product F	สินค้าประเภทเทปใสกาวยาง
Product G	สินค้าประเภทเทปใสกาวน้ำ
Product H	สินค้าประเภทFilament Tape
Product I	สินค้าประเภทสติ๊กเกอร์
Product J	สินค้าประเภทเทปกาวสองหน้า
Product K	สินค้าประเภทเทปพันสายไฟ
Product L	สินค้าประเภทเทปอลูมิเนียม
Product M	สินค้าประเภทเทปกาวโฟมสองหน้า

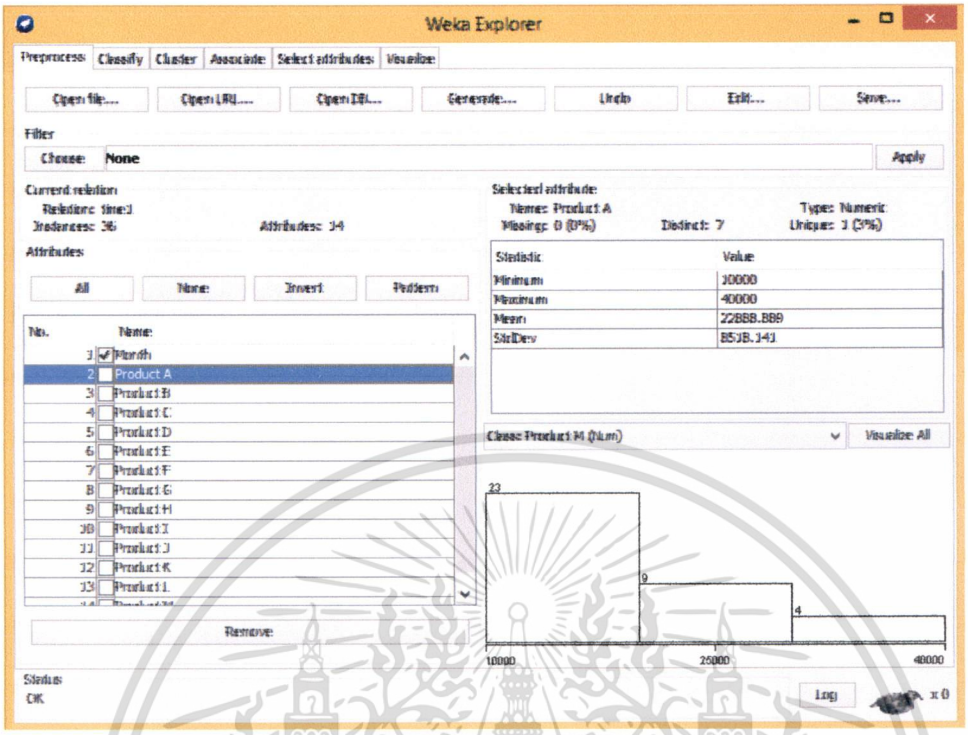
โดยค่าข้อมูลที่ใส่จะประกอบด้วย 2 รูปแบบ คือ

ข้อมูลที่เป็นรูปแบบเดือนและปี โดยมีทั้งสิ้น 36 เดือน

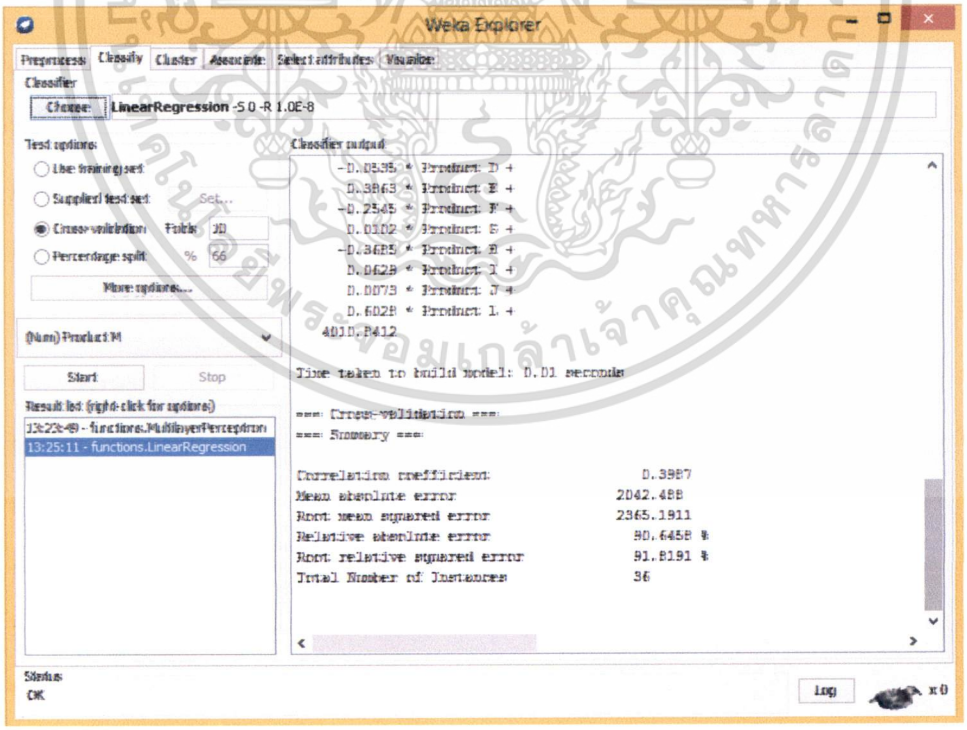
ข้อมูลประเภทสินค้า โดยค่าที่จัดเก็บจะเป็นจำนวนรวมของการขายสินค้าในแต่ละเดือน

การนำเข้าข้อมูลในโปรแกรม Weka

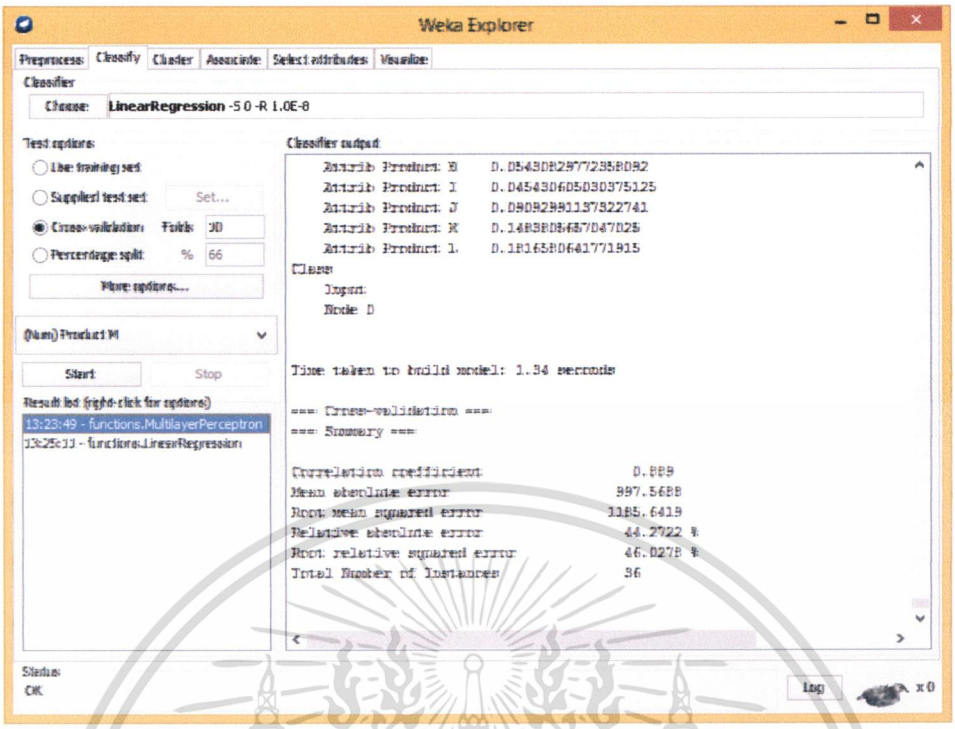
เนื่องจากรูปแบบข้อมูลของซอฟต์แวร์ที่ใช้ในการประมวลผลยอมรับได้นั้นจะต้องอยู่ในรูปไฟล์นามสกุล .arff หรือ .crv ดังนั้นผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในรูปแบบ .crv เพื่อความเหมาะสมในการนำเข้าข้อมูล และสามารถแก้ไขข้อมูล ดังภาพที่ (ข้างบน)



รูปภาพที่ 4.15 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิคการทำนาย



รูปภาพที่ 4.16 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิคการทำนาย
โดยใช้อัลกอริทึม Linear Regression



รูปภาพที่ 4.17 การนำเข้าข้อมูลในโปรแกรม Weka ด้วยเทคนิคการทำนาย โดยใช้อัลกอริทึม Neural Network

4.3.2 ผลการดำเนินงาน

ในการทดสอบนี้ในกฎการความสัมพันธ์ เราจะใช้ 2 อัลกอริทึม

อัลกอริทึม Linear Regression

จากการทดสอบโดยกำหนดตัวแปร คือ Product(สินค้า) โดยมีรูปแบบของกลุ่มสินค้าออกเป็น 13 ประเภท เพื่อที่เราจะสามารถหาคำตอบในหัวข้อศึกษา “ทำนายยอดปริมาณสินค้าในการสั่งซื้อล่วงหน้า” โดยในหัวข้อกรณีการศึกษานี้ เราสามารถหาผลการศึกษาได้ดังนี้

=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: time1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 36

Attributes: 13

Product A

Product B

Product C

Product D

Product E

Product F

Product G

Product H

Product I

Product J

Product K

Product L

Product M

Test mode:evaluate on training data

=== Classifier model (full training set) ===

Linear Regression Model

Product M =

Time taken to build model: 0.01 seconds

=== Predictions on training set ===

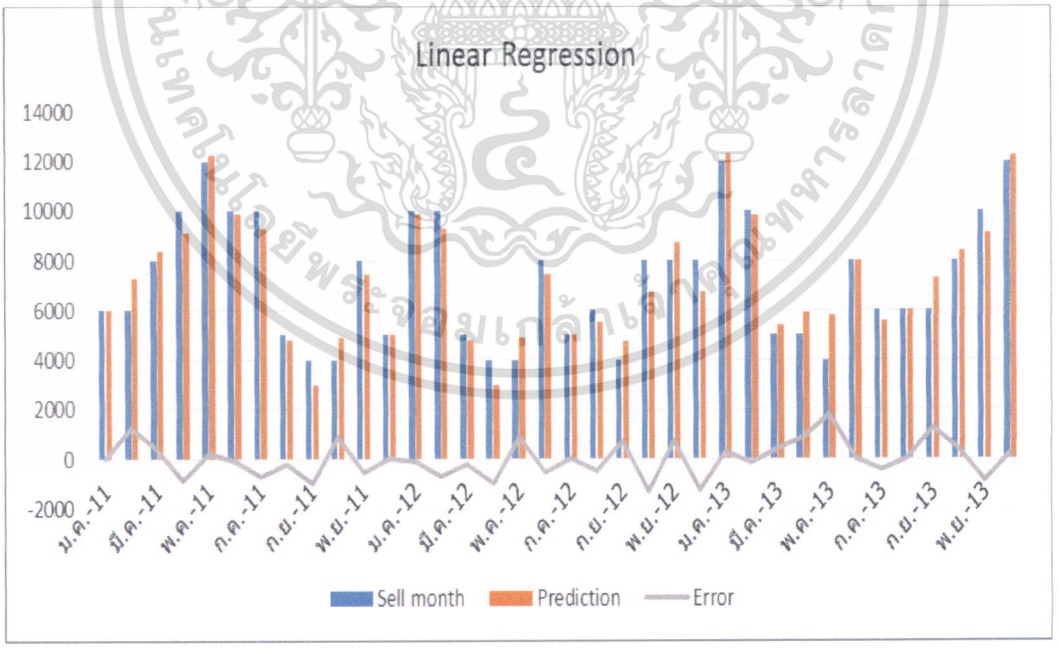
Instant	Actual	Predicted	Error
1	12000	12284.054	284.054
2	10000	9835.428	-164.572
3	5000	5388.482	388.482
4	5000	5917.15	917.15
5	4000	5801.584	1801.584
6	8000	7992.094	-7.906
7	6000	5572.804	-427.196
8	6000	6014.379	14.379
9	6000	7282.647	1282.647
10	8000	8416.532	416.532
11	10000	9121.806	-878.194
12	12000	12234.464	234.464
13	10000	9870.461	-129.539
14	10000	9280.832	-719.168
15	5000	4810.614	-189.386
16	4000	2990.442	-1009.558
17	4000	4903.571	903.571
18	8000	7437.498	-562.502
19	5000	5015.424	15.424
20	6000	5531.844	-468.156
21	4000	4750.863	750.863
22	8000	6710.159	-1289.841
23	8000	8729.407	729.407
24	8000	6728.795	-1271.205
25	6000	6014.379	14.379
26	6000	7282.647	1282.647

Instant	Actual	Predicted	Error
27	8000	8416.532	416.532
28	10000	9121.806	-878.194
29	12000	12234.464	234.464
30	10000	9870.461	-129.539
31	10000	9280.832	-719.168
32	5000	4810.614	-189.386
33	4000	2990.442	-1009.558
34	4000	4903.571	903.571
35	8000	7437.498	-562.502
36	5000	5015.424	15.424

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.9555
Mean absolute error	589.1985
Root mean squared error	742.6909
Relative absolute error	26.5139 %
Root relative squared error	29.4972 %
Total Number of Instances	36



รูปภาพที่ 4.18 รูปภาพการทำนายยอดขายด้วย Linear Regression

ในรูปแบบการวิเคราะห์ผลลัพธ์ด้วย Linear Regression นั้นค่าผลลัพธ์ที่ได้นั้นในแต่ละเดือนจะเห็นว่ามีค่า Error หรือค่าความคลาดเคลื่อนจากความเป็นจริงนั้นแตกต่างกัน เช่น ค่าในเดือน มิถุนายน 2013 นั้น จำนวนยอดขายในเดือนนั้นของสินค้า Produck M นั้น มียอดขายอยู่ที่ 8000 บาท ค่าผลลัพธ์ที่ได้จากการทำนายด้วย Linear Regression อยู่ที่ 7992.094 บาท มีค่าความคลาดเคลื่อน หรือค่า Error อยู่ที่ -7.906 บาท จะเห็นว่ามีค่าความคลาดเคลื่อนที่น้อยหรือมีค่าใกล้เคียงกับค่าความเป็นจริง แต่ถ้าในปีเดียวกัน แต่เป็นเดือน พฤษภาคม 2013 นั้น มียอดขายจริงอยู่ที่ 4000 บาท ค่าผลลัพธ์ที่ได้จากการทำนายด้วย Linear Regression อยู่ที่ 5801.584 บาท มีค่าความคลาดเคลื่อน หรือค่า Error อยู่ที่ 1801.584บาท ซึ่งมีค่าความคลาดเคลื่อนที่สูงมาก โดยเราสามารถดูค่า Error ของผลลัพธ์ที่ได้คือ Root mean squared error 742.6909 เป็นค่าผลลัพธ์ความคลาดเคลื่อนโดยรวมของผลลัพธ์ที่ได้จากการใช้อัลกอริทึม Neuron Network จะเห็นได้ว่ามีค่าความคลาดเคลื่อนโดยรวมของผลลัพธ์ที่สูงมาก

Neural Network

=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Relation: time1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 36

Attributes: 13

Product A

Product B

Product C

Product D

Product E

Product F

Product G

Product H

Product I

Product J

Product K

Product L

Product M

Test mode:evaluate on training data

=== Classifier model (full training set) ===

Input

Node 0

Time taken to build model: 0.11 seconds

=== Predictions on training set ===

Product M

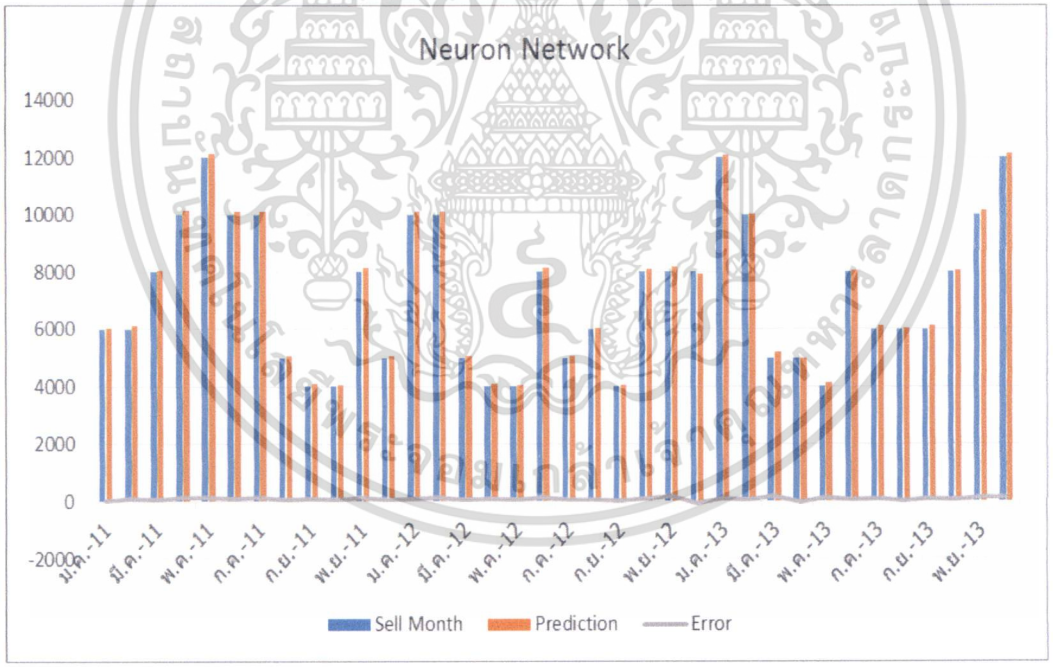
Instant	Actual	Predicted	Error
1	12000	12106.545	106.545
2	10000	10043.595	43.595
3	5000	5189.293	189.293
4	5000	4979.999	-20.001
5	4000	4148.416	148.416
6	8000	8047.06	47.06
7	6000	6112.765	112.765
8	6000	6037.108	37.108
9	6000	6115.445	115.445
10	8000	8061.095	61.095
11	10000	10163.542	163.542
12	12000	12143.846	143.846
13	10000	10110.043	110.043
14	10000	10129.214	129.214
15	5000	5077.275	77.275
16	4000	4098.217	98.217
17	4000	4077.246	77.246
18	8000	8144.847	144.847
19	5000	5058.808	58.808
20	6000	6048.158	48.158
21	4000	4038.122	38.122
22	8000	8103.542	103.542
23	8000	8174.607	174.607
24	8000	7948.947	-51.053
25	6000	6037.108	37.108
26	6000	6115.445	115.445
27	8000	8061.095	61.095
28	10000	10163.542	163.542
29	12000	12143.846	143.846
30	10000	10110.043	110.043
31	10000	10129.214	129.214

Instant	Actual	Predicted	Error
32	5000	5077.275	77.275
33	4000	4098.217	98.217
34	4000	4077.246	77.246
35	8000	8144.847	144.847
36	5000	5058.808	58.808

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.9998
Mean absolute error	97.6813
Root mean squared error	107.4242
Relative absolute error	4.3957 %
Root relative squared error	4.2665 %
Total Number of Instances	36



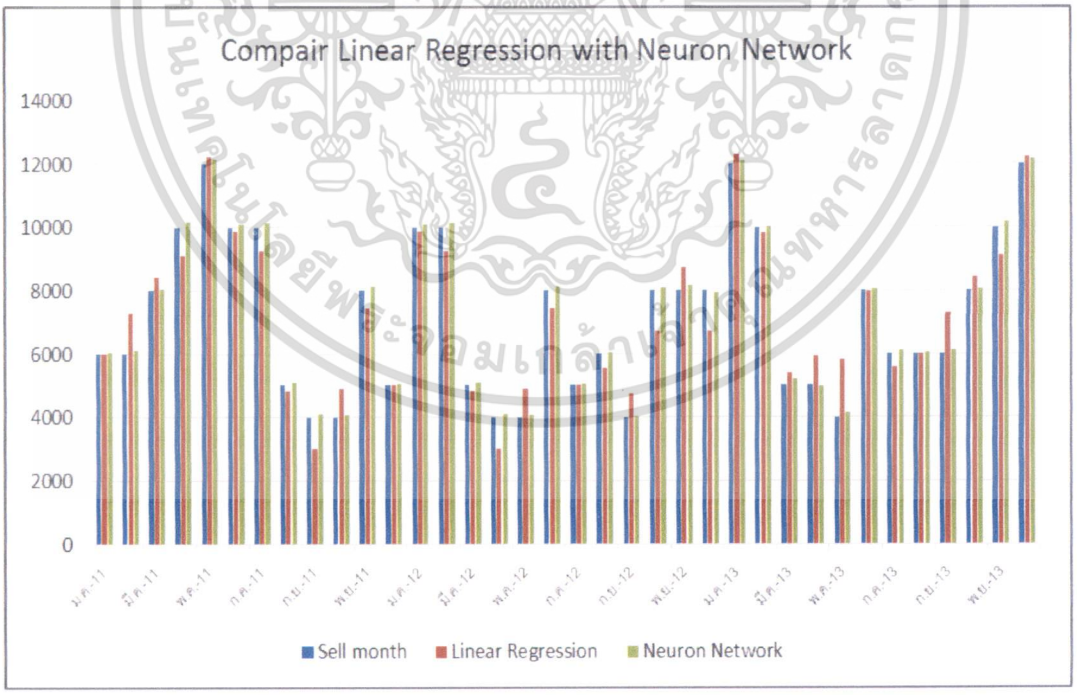
รูปภาพที่ 4.19 รูปภาพการทำนายยอดขายด้วย Neuron Network

ในรูปแบบการวิเคราะห์ผลลัพธ์ด้วย Neuron Network นั้นค่าผลลัพธ์ที่ได้นั้นในแต่ละเดือน จะเห็นว่ามีค่า Error หรือค่าความคลาดเคลื่อนจากความเป็นจริงนั้นมีค่าที่ไม่แตกต่างกัน เช่น ค่าในเดือน มิถุนายน 2013 นั้น จำนวนยอดขายในเดือนนั้นของสินค้า Produck M นั้น มียอดขายอยู่ที่

8000 บาท ค่าผลลัพธ์ที่ได้จากการทำนายด้วย Neuron Network อยู่ที่ 8047.06บาท มีค่าความคลาดเคลื่อน หรือค่า Error อยู่ที่ 47.06บาท แต่ถ้าในปีเดียวกัน แต่เป็นเดือน พฤษภาคม 2013 นั้นมียอดขายจริงอยู่ที่ 4000 บาท ค่าผลลัพธ์ที่ได้จากการทำนายด้วย Neuron Network อยู่ที่ 4148.416 บาท มีค่าความคลาดเคลื่อน หรือค่า Error อยู่ที่ 148.416 บาท เป็นต้น ซึ่งจะเห็นจากรูปภาพว่า ค่าความคลาดเคลื่อนที่ผ่านการวิเคราะห์ด้วย Neuron Network นั้นจะมีค่าความคลาดเคลื่อนที่ไม่สูงมากนัก โดยเราสามารถดูค่า Error ของผลลัพธ์ที่ได้คือ Root mean squared error 107.4242 เป็นค่าผลลัพธ์ความคลาดเคลื่อนโดยรวมของผลลัพธ์ที่ได้จากการใช้อัลกอริทึม Neuron Network

4.3.3 ผลการทดสอบเพื่อวัดประสิทธิภาพ

จากผลการทดสอบ โดยผ่านการประมวลผลทั้ง 2 อัลกอริทึม คือ Linear Regession และ Neural Network นั้น โดยผลค่าการทดสอบที่ได้มานั้นใน เราสามารถทำการประเมินผลทดสอบและสามารถแปลค่าผลทดสอบที่ได้นั้น จะพบว่า อัลกอริทึม Linear Regession นั้นสามารถแสดงผลการทดสอบที่ได้ออกมานั้นจะพบว่า มีค่าความคลาดเคลื่อนที่แตกต่างกัน และมีค่าความคลาดเคลื่อนที่สูงมาก แต่ผลลัพธ์ของ Neuron Network นั้น จะมีค่าความคลาดเคลื่อน ที่ไม่แตกต่างกันมากนัก และไม่มีค่าความคลาดเคลื่อนที่แตกต่างกันอย่างชัดเจน โดยดูผลลัพธ์ได้จากรูปภาพเปรียบเทียบ ค่าผลลัพธ์ในการทำนายของทั้งสองอัลกอริทึม ได้จากรูปภาพ 4.19



รูปภาพที่ 4.20 รูปภาพเปรียบเทียบการทำนายยอดขายด้วย Linear Regression และ Neuron Network

บทที่ 5
สรุป อภิปรายผลและข้อเสนอแนะ

จากผลการดำเนินงานวิจัยการวิเคราะห์ความสัมพันธ์ของสินค้าเทปกาวกับพฤติกรรมการซื้อของลูกค้า สามารถสรุปส่วนสำคัญต่างๆ ได้ดังนี้

5.1 สรุปผลการการทดสอบ

5.2 อภิปรายผล

5.3 ปัญหาและอุปสรรค

5.4 ข้อเสนอแนะ

5.1 สรุปผลกรณีศึกษา

จากการวิเคราะห์ความสัมพันธ์ของพฤติกรรมลูกค้า กับการซื้อสินค้าเทปกาว จากกรณีศึกษาข้อมูลการขายสินค้าเทปกาวของบริษัทที่จำหน่ายเทปกาวในปี พ.ศ. 2555 โดยใช้เทคนิคการทำเหมืองข้อมูล 3 เทคนิคคือเทคนิคกฎความสัมพันธ์และเทคนิคการแบ่งกลุ่มผู้บริโภค และเทคนิคการทำนาย ซึ่งผลการวิจัยการในการตอบคำถามในกรณีศึกษานี้ตั้งไว้ 3 ข้อ คือ

5.1.1 ลูกค้าส่วนใหญ่เมื่อซื้อสินค้าอย่างหนึ่งแล้วมักจะซื้อสินค้าคู่กับสินค้าอะไร

จากการวิเคราะห์ด้วย เทคนิคกฎความสัมพันธ์ โดยใช้อัลกอริทึม 2 แบบ คือ Apiori ได้ค่าของกฎ ออกมาที่ 23 กฎ โดยกำหนดค่า confidence: 0.8 ในส่วน อัลกอริทึม FP Growth ได้ค่าของกฎ ออกมาที่ 23 กฎ โดยกำหนดค่า confidence: 0.8 เช่นกัน และผลลัพธ์ที่ได้ใน 10 ลำดับแรกมีค่าเหมือนกัน จึงสามารถทำให้เราทราบว่า เมื่อลูกค้าซื้อสินค้าอย่างหนึ่งแล้วมักจะซื้อสินค้าชนิดใด เช่น ลูกค้าซื้อ Product B=Y และซื้อProduct C=Y 123 ==> ลูกค้าจะซื้อ Product A=Y 115 โดยได้ค่า confidence:0.93% จึงสามารถตอบโจทย์ในการนำข้อมูลไปวิเคราะห์พฤติการณ์ซื้อของลูกค้าในธุรกิจเทปกาวได้

5.1.2 วิเคราะห์พฤติกรรมลูกค้ากับการซื้อสินค้า (แยกตามภูมิศาสตร์ และรูปแบบองค์กรลูกค้า)

ในการวิเคราะห์พฤติกรรมลูกค้ากับการซื้อสินค้าเทปกาว โดยแยกตามภูมิศาสตร์ลักษณะที่ตั้ง และรูปแบบองค์กรของลูกค้า จากผลการทดสออบด้วยเทคนิคการแยกกลุ่ม โดยใช้อัลกอริทึม 2 แบบ คือ K Means ได้ผลการทดสอบออกมาถูกต้องตามที่ได้ตั้งคำถามในการทดสอบ โดยสามารถแยกกลุ่มของลูกค้าตามรูปแบบลักษณะองค์กร และ ยังสามารถแยกกลุ่มตามลักษณะภูมิประเทศ และสามารถให้คตอบในการทราบพฤติกรรมการซื้อสินค้าของลูกค้าแต่ละกลุ่ม [illegible] ส่วนอัลกอริทึมที่ 2 คือ Hierachicalclustering ผลการทดสอบที่ได้ออกมา จะเป็นเชิงข้อมูลเ..

ลักษณะตัวเลขแบบกลุ่ม โดยผลลัพธ์ที่ได้ในการทดสอบนั้น เราจะได้ผลลัพธ์ในการแบ่งกลุ่มข้อมูลตามลักษณะภูมิประเทศเพียงอย่างเดียวโดยไม่สามารถให้ผลลัพธ์ในรูปแบบองค์กรของลูกค้าหรือรูปแบบการซื้อสินค้าของลูกค้าได้

5.1.3 ทำนายยอดปริมาณสินค้าในการสั่งซื้อล่วงหน้า

ในการทำนายการสั่งซื้อสินค้าของลูกค้าล่วงหน้านั้นโดยใช้อัลกอริทึม 2 แบบ คือ Linear Regression และ Neural Network นั้น โดยผลลัพธ์ที่ได้จาก Linear Regression โดยผลค่าการทดสอบที่ได้มานั้นใน เราสามารถทำการประเมินผลทดสอบ และสามารถแปลค่าผลทดสอบที่ได้นั้น จะพบว่า อัลกอริทึม Linear Regession นั้นสามารถแสดงผลการทดสอบที่ได้ออกมานั้นจะพบว่า มีค่าความคลาดเคลื่อนที่แตกต่างกัน และมีค่าความคลาดเคลื่อนที่สูงมากของค่ายอดขายที่แท้จริง แต่ผลลัพธ์ของ Neuron Network นั้น จะมีค่าความคลาดเคลื่อนที่ไม่แตกต่างกันมากนัก โดยค่าที่ได้จากการทำนายยอดขายนั้นในแต่ละเดือน จะมีค่าความคลาดเคลื่อนจากยอดขายจริงไม่แตกต่ามากนัก ซึ่งสามารถสรุปได้ว่า อัลกอริทึม Nueron Network นั้นมีความเหมาะสมกับข้อมูลชุดนี้ที่เรานำมาทดสอบมากกว่า Linear Regression เพราะค่าที่ได้จากอัลกอริทึมนี้มีค่าความคลาดเคลื่อนสูง

5.2 อภิปรายผล

งานวิจัยนี้นำเสนอการวิเคราะห์พฤติกรรมลูกค้ากับธุรกิจเทปกาวด้วยเทคนิคการค้นหากฎความสัมพันธ์ เทคนิคการแบ่งกลุ่ม และเทคนิคการทำนาย โดยผลการจากกฎหาความสัมพันธ์ของข้อมูล จากผลการทดลองสรุปได้ว่าลูกค้ากับพฤติกรรมการซื้อสินค้าเทปกาวนั้นจะมีพฤติกรรมการซื้อสินค้า คือ ลูกค้าเมื่อมีการซื้อสินค้าชนิดเทปกระดาษกาวย่น และเทปผ้ากาวย่น แล้วนั้น โอกาสที่ลูกค้านั้นจะซื้อสินค้าชนิดเทปผ้ากาวนั้น มีค่าความเชื่อมั่นถึง 0.93% ถัดมาคือ ลูกค้าเมื่อมีการซื้อสินค้าชนิดเทปกาวย่น ซื้อสินค้าสติ๊กเกอร์ และซื้อสินค้าเทปกาวสองหน้านั้น โอกาสที่ลูกค้านั้นจะซื้อสินค้าชนิดเทปผ้ากาว มีค่าความเชื่อมั่นถึง 0.93% เช่นกัน โดยผลลัพธ์ที่ได้จากการทดสอบนั้นทั้ง 2 อัลกอริทึมให้ค่าผลลัพธ์และความเชื่อมั่นออกมาเหมือนกัน ส่วนเทคนิคการแบ่งกลุ่มนั้น เราได้ผลลัพธ์ที่ถูกต้องและครบถ้วน คือ อัลกอริทึม K Means กล่าวคือ ผลลัพธ์ที่ได้การการทดสอบของ K Means นั้นสามารถตอบหัวข้อของการศึกษาโดยใช้วิธีแบ่งกลุ่มได้ว่า ลูกค้าที่อยู่ในบริเวณกรุงเทพและปริมณฑล มีทั้งหมด จะซื้อสินค้าประเภทเทป OPP โดยมีค่าความเชื่อมั่นคือ 1 และส่วนมากลูกค้าที่อยู่บริเวณกรุงเทพและปริมณฑลจะมีลักษณะรูปแบบองค์กร คือ เป็นบริษัทจำกัด เป็นต้น แต่ในส่วนอัลกิริทึม Hierachicalcluster นั้นผลลัพธ์ที่ได้ไม่สามารถตอบโจทย์หัวข้อของการศึกษาในรูปแบบการแบ่งกลุ่มได้ชัดเจน ส่วนรูปแบบการทำนายผล เราสามารถเห็นได้ว่าข้อมูลที่เรานำมาทำการศึกษานี้ จะประกอบด้วยข้อมูลที่เป็นข้อมูลจริงและข้อมูลที่อ้างอิงขึ้น โดยผลลัพธ์

ที่ได้นั้นจะเห็นว่า ค่าผลลัพธ์ที่ได้จากอัลกอริทึม Neuron Network นั้นจะให้ผลลัพธ์ที่ใกล้เคียงกับยอดขาย และค่าที่ได้จากการทำนายนั้นจะมีค่าที่ไม่แตกต่างจากค่าความเป็นจริงมาก ซึ่งต่างกับอัลกอริทึม Linear Regression ที่ให้ค่าผลลัพธ์ในการทำนายที่มีค่าความคลาดเคลื่อนแตกต่างจากยอดขายจริง ดังนั้นเราสามารถนำโมเดลที่ได้จากผลการทดสอบทั้งสามกรณี มาทำโมเดลในการศึกษากับธุรกิจเทปกาวได้อย่างเหมาะสมและถูกต้อง

5.3 ปัญหาและอุปสรรค

5.3.1 เนื่องจากรายการสินค้ามีความหลากหลาย ทำให้ยากต่อการจัดกลุ่มสินค้าให้เหมาะสมกับการทดลอง

5.3.2 จากข้อมูลที่ได้รับมาให้ทำการทดสอบนั้นคือมีข้อมูล 12 เดือน ไม่เพียงพอต่อการทำการทดสอบในหัวข้อการศึกษาบางหัวข้อและไม่เพียงต่อการนำมาใช้ในบางอัลกอริทึม ของเทคนิคดาต้าไมน์นิ่ง เช่น เทคนิค Time Serie หรือเทคนิคที่ใช้ข้อมูลในการทำนาย เนื่องจากเทคนิคในการทำนายจำเป็นต้องใช้ข้อมูลในการทำนายเป็นจำนวนมาก

5.4 ข้อเสนอแนะ

5.4.1 การศึกษาครั้งนี้นำข้อมูลที่ได้รับมาเป็นระยะเวลา 12 เดือนมาวิเคราะห์หาแบบจำลองพฤติกรรมของลูกค้า ซึ่งพฤติกรรมลูกค้าอาจเปลี่ยนแปลงได้ตามกาลเวลา ดังนั้นจึงมีข้อมูลในการวิเคราะห์อย่างเพียงพอและครบถ้วนและข้อมูลที่ได้รับควรเป็นข้อมูลมีความถูกต้องและทันต่อเวลาเพื่อให้ผลที่ได้ทันต่อสภาพแวดล้อมที่เปลี่ยนไป ซึ่งจะทำให้ผลการวิเคราะห์สามารถทำนายได้อย่างแม่นยำมากขึ้น

5.4.2 การทำเหมืองข้อมูลมีเทคนิค และวิธีการทำเหมืองข้อมูลจำนวนมาก ดังนั้นควรศึกษาเทคนิคและวิธีการแบบอื่นด้วยเพื่อให้เหมาะสมกับข้อมูลที่ใช้ในการทดลองที่สุด

5.4.3 สามารถนำวิธีการอื่นมาเปรียบเทียบประสิทธิภาพหรือทำงานร่วมกัน เพื่อให้ได้ผลลัพธ์ที่ถูกต้องและชัดเจนในการทดสอบหรือหัวข้อในการศึกษา

5.4.5 ควรศึกษาข้อมูลเพิ่มเติมเกี่ยวกับพฤติกรรมของลูกค้าด้านอื่นๆ เพิ่มเติม ไม่ว่าจะเป็นโปรโมชั่น ประเภทธุรกิจของลูกค้าหรือ ข้อมูลอื่นๆเพิ่มเติมเพื่อสามารถทำการวิเคราะห์ข้อมูลและพฤติกรรมลูกค้าในธุรกิจเทปกาวได้อย่างถูกต้อง

บรรณานุกรม

ภาษาไทย

กฤษณะ ไวยมัย 2549. **สถาปัตยกรรมพื้นฐานของดาต้าไมน์นิ่ง**

กรกต เกรียงพันธุ์อมร 2543. **โมเดลในการทำดาต้าไมน์นิ่ง**

ธนพงษ์ นิติการุณ และปราการ อัศวชีวันธรกุล 2545. **งานของดาต้าไมน์นิ่งขั้นพื้นฐาน**

ดลใจ กระแสเสน 2547. **การหาพฤติกรรมของลูกค้าที่จะมีแนวโน้มการยกเลิกโทรศัพท์พื้นฐาน**

ชลนิศา สาระ 2550. **วิธีการจำแนกกลุ่มสถานภาพการสำเร็จการศึกษาโดยใช้แบบจำลองต้นไม้ตัดสินใจ**

บวร น้อยแสง 2549. **การใช้เทคนิคการทำเหมืองข้อมูลเพื่อช่วยในการวิเคราะห์การขาย**