บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

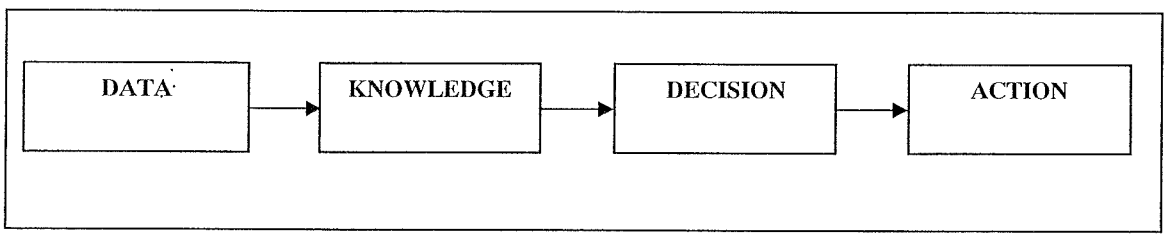
การสร้างโมเดลระบบบริหารความสัมพันธ์นักศึกษาโดยใช้เทคนิคเหมืองข้อมูล โดยนำข้อมูลต่างๆ มาใช้ในการบริหารความสัมพันธ์ของนักศึกษากับมหาวิทยาลัย เพื่อสร้างความพึงพอใจให้กับนักศึกษาในการเข้ารับการศึกษาในสถาบันการศึกษาต่างๆ จึงจำเป็นอย่างยิ่งที่ต้องศึกษาหาข้อมูลในส่วนของทฤษฎีและผลงานวิจัยต่างๆ ที่เกี่ยวข้อง ซึ่งมีรายละเอียดดังนี้คือ

2.1 การทำเหมืองข้อมูล (Data Mining)

2.1.1. ความหมายของการทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) หรืออาจเรียกว่าการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database - KKD) เป็นเทคนิคเพื่อค้นหารูปแบบ ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่องและการรู้จำแบบ หรือในอีกนิยามหนึ่งของการทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูล (โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

การทำเหมืองข้อมูล (Data Mining) คือ การค้นหาความสัมพันธ์ และรูปแบบทั้งหมดซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก การทำเหมืองข้อมูลจะทำการสำรวจ และวิเคราะห์อย่างอัตโนมัติ ในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆที่มีประโยชน์ในฐานข้อมูล



รูปที่ 2.1 การแสดงข้อมูลสู่การตัดสินใจและปฏิบัติ

ส่วนวิเคราะห์และทำนายนี้จะใช้เทคนิคการเหมืองข้อมูล(Data Mining) ซึ่งเป็นกระบวนการในการขุดค้น วิเคราะห์ เพื่อที่จะค้นหารูปแบบหรือกฎเกณฑ์ที่เป็นประโยชน์ และน่าสนใจออกมาจากข้อมูลที่มีจำนวนมากมายมหาศาลโดยวิธีแบบอัตโนมัติ หรือ กึ่งอัตโนมัติ

การทำเหมืองข้อมูล จำแนกออกเป็นประเภทของรูปแบบได้ 2 รูปแบบ ดังนี้

(1) การทำเหมืองข้อมูลเพื่อหาหรือกำหนดลักษณะหรือคุณสมบัติของข้อมูลเป็นการวิเคราะห์เพื่อหาลักษณะหรือคุณสมบัติของกลุ่มข้อมูล การหาลักษณะของกลุ่มข้อมูลทั่วไปในฐานข้อมูลซึ่งจะหาคุณลักษณะ 3 แบบคือ หาลักษณะคุณลักษณะที่ตรงกัน หาลักษณะคุณลักษณะที่ต่างกัน และหาลักษณะที่มีทั้งความต่างและเหมือนกัน

(2) การทำเหมืองข้อมูลเพื่อใช้ในการทำนายสิ่งที่จะเกิดขึ้น เป็นการวิเคราะห์หาโมเดลเพื่อทำนายแนวโน้มหรือคุณสมบัติของข้อมูลที่ยังไม่ปรากฏ เช่น การหายอดขายสินค้าของเดือนถัดไป

ประโยชน์ของการทำเหมืองข้อมูล

ประโยชน์ของการทำเหมืองข้อมูล นั้นมีหลายอย่างเริ่มตั้งแต่การวิเคราะห์ผลิตภัณฑ์ การวิเคราะห์บัตรเครดิต การวิเคราะห์ลูกค้า การวิเคราะห์การขาย พาณิชย์อิเล็กทรอนิกส์(E-Commerce)

กระบวนการค้นพบความรู้จากฐานข้อมูล

กระบวนการค้นพบความรู้จากฐานข้อมูลขนาดใหญ่มาก เป็นกระบวนการสร้างแบบจำลองหรือรูปแบบกฎเกณฑ์จากกลุ่มของข้อมูล ทำให้เกิดความเข้าใจใจลักษณะรูปแบบความเกี่ยวข้องสัมพันธ์กันของกลุ่มข้อมูล แล้วแนวโน้มเพื่อใช้ในการทำนายข้อมูลนั้นๆ โดยมี กระบวนการตามทฤษฎี รวม 4 ขั้นตอนดังนี้

(1) การเตรียมข้อมูล (Data preparation)

ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนสำคัญ และใช้เวลามากที่สุด เนื่องจากบ่อยครั้งเลือกข้อมูลมาไม่เหมาะสม และไม่ถูกต้อง หรือการนำข้อมูลมาจากหลายแหล่งที่มารวมเข้ากันเพื่อพิจารณาความสัมพันธ์ของข้อมูล ซึ่งส่งผลให้เกิดความผิดพลาด ดังนั้นขั้นตอนการเตรียมข้อมูลจึงถือเป็นหัวใจของงาน การเตรียมข้อมูลสามารถแบ่งออกได้เป็น 3 ขั้นตอนย่อย คือ

- การคัดเลือกข้อมูล (Data Selection)

จุดประสงค์หลัก คือ การระบุลักษณะข้อมูลที่ต้องการแล้ว ทำการคัดเลือกข้อมูลทำการคัดเลือกข้อมูลที่ต้องการ ซึ่งข้อมูลที่ได้จะแตกต่างกันไปตามจุดประสงค์ ของแต่ละธุรกิจ

- การกลั่นกรองข้อมูล (Data Cleaning)

จุดประสงค์ เพื่อมั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นถูกต้องและเหมาะสม เนื่องจากข้อมูลที่ถูกเลือกมาจากกระบวนการเลือกข้อมูลนั้นอาจมีข้อมูลไม่ถูกต้อง

- การแปลงรูปข้อมูล (Data Transformation)

จุดประสงค์เพื่อแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์ตามหลักอัลกอริทึมของการทำเหมืองข้อมูลที่เลือกใช้ เช่น การแบ่งช่วงอายุให้เป็นกลุ่มๆ หรือ กำหนดตัวเลขให้กับแต่ละกลุ่มของข้อมูล

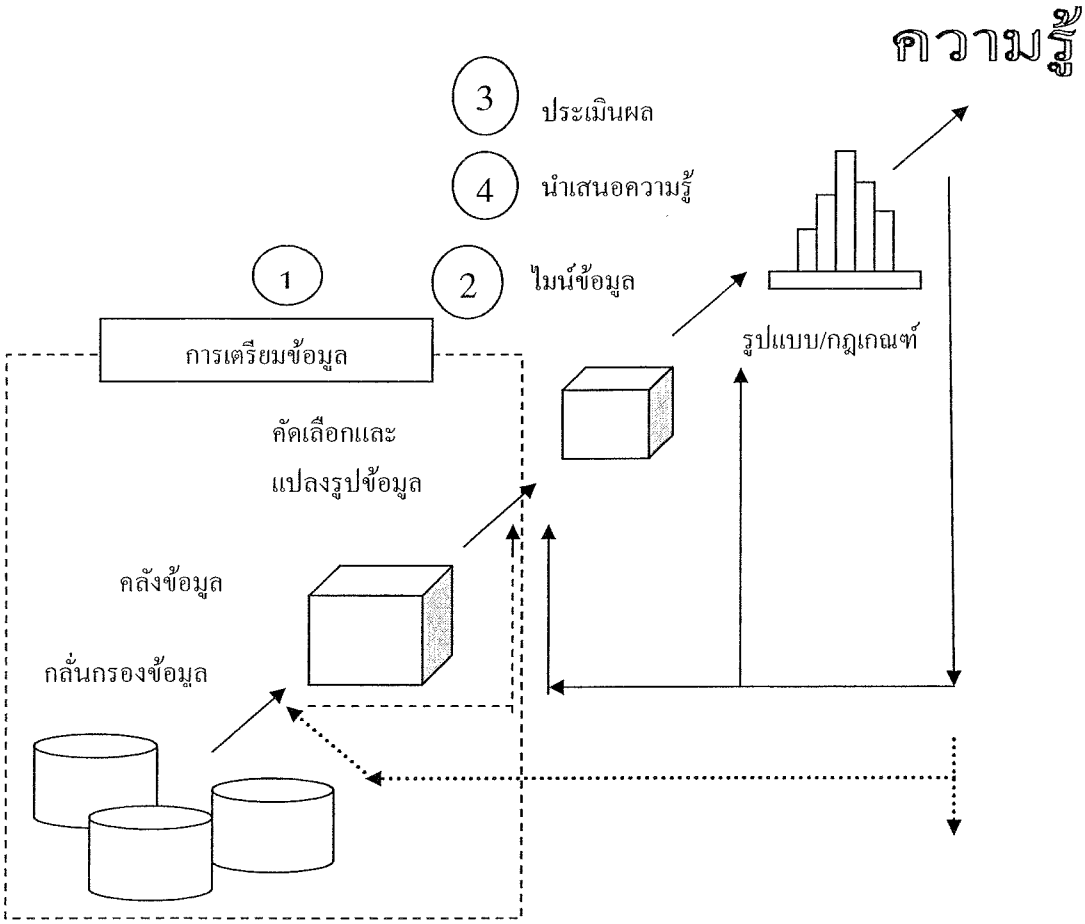
(2) การทำเหมืองข้อมูล (Data Mining)

เป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนที่ผ่านมา โดยเมื่อทำขั้นตอนนี้แล้วอาจต้องย้อนกลับไปทำขั้นตอนการเตรียมข้อมูลใหม่

(3) การประเมินรูปแบบ/กฎเกณฑ์ที่ได้ (Pattern evaluation)

เป็นขั้นตอนการวิเคราะห์ และประเมินผลของรูปแบบหรือกฎเกณฑ์ ที่หาได้จากขั้นตอนการหาความรู้จากข้อมูล การทำงานในส่วนนี้จำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลทางธุรกิจเข้าช่วย

(4) การนำเสนอความรู้ (Knowledge presentation)นำความรู้ที่ค้นพบไปประยุกต์ใช้งานจริงต่อไป



รูปที่ 2.2 กระบวนการค้นพบความรู้จากฐานข้อมูล

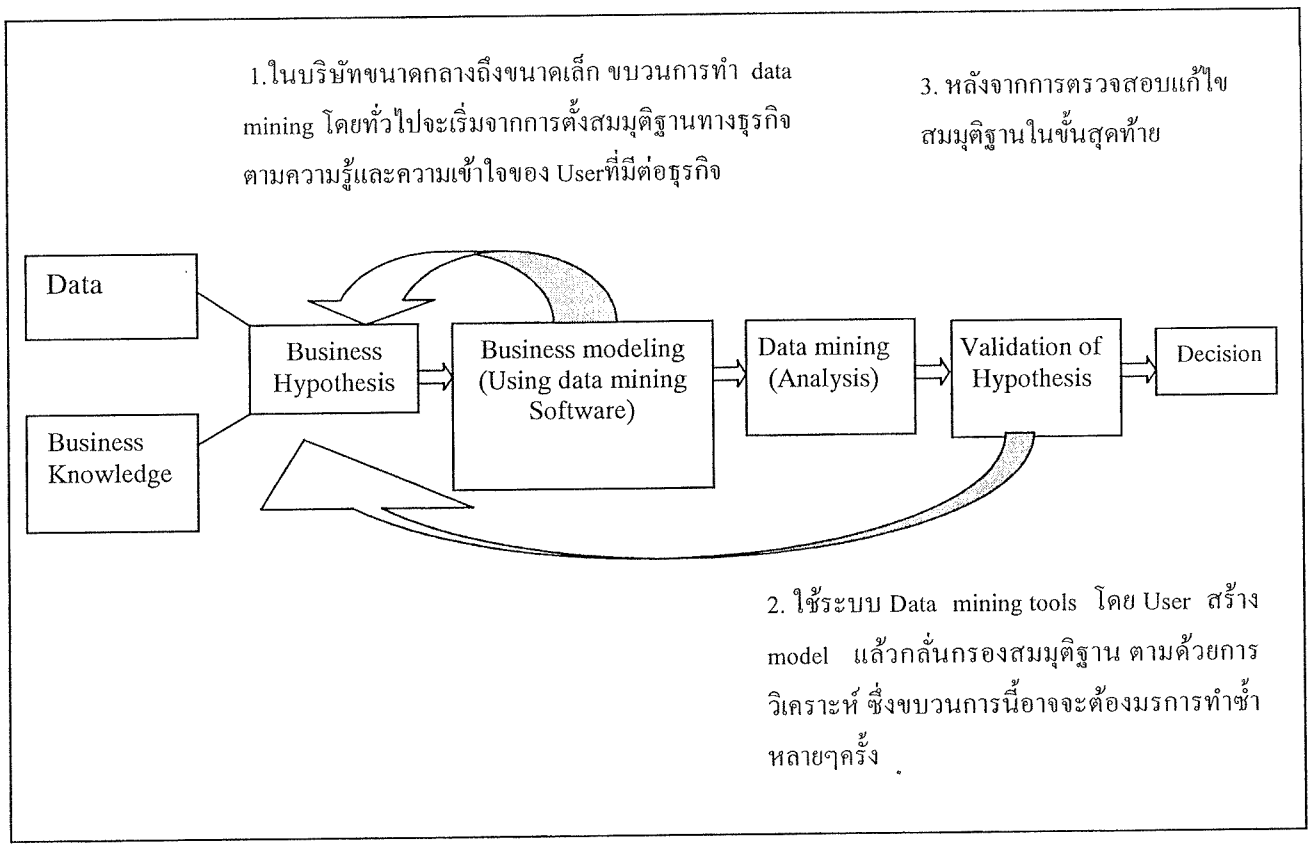
เทคนิคในการทำเหมืองข้อมูล

เทคนิคที่ใช้ในการขุดค้นข้อมูลหรือไมน์ข้อมูลอาศัยหลักการทางสถิติในการวิเคราะห์ข้อมูล โดยมี 3 เทคนิคสำคัญที่เป็นที่รู้จักแพร่หลายในปัจจุบันดังต่อไปนี้

(1) เทคนิคการค้นหากฎความสัมพันธ์ของข้อมูล(Association Rule Discovery)เป็นการค้นหากฎความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆ

(2) เทคนิคการจัดข้อมูลเข้ากลุ่มและการทำนาย(Classification and Prediction) เทคนิคการจัดข้อมูลเข้ากลุ่มเป็นกระบวนการการสร้างโมเดลจัดการข้อมูลขนาดใหญ่ให้อยู่ในกลุ่มที่กำหนดมาให้ และทำการจัดข้อมูลเข้ากลุ่มตามโมเดลที่สร้างขึ้นนั้น ส่วนเทคนิคการทำนายเป็นการวิเคราะห์ทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่

(3) เทคนิคการจำแนกลุ่มข้อมูล(Clustering) เป็นการวิเคราะห์หาชั้นหรือกลุ่มของข้อมูล ให้กับข้อมูลโดยที่ไม่รู้ชื่อคลาสมาก่อนซึ่งต่างกับเทคนิคการจัดการข้อมูล และการทำนายที่วิเคราะห์หาคลาสให้กับข้อมูลโดยมีการกำหนดคลาสไว้ให้ก่อนแล้ว



รูปที่ 2.3 การแสดงขบวนการตัดสินใจ

ตัวอย่างเช่น การค้นหากฎความสัมพันธ์(Association Rule) ของสินค้าในห้างสรรพสินค้า พบว่าลูกค้าร้อยละ 90 ที่ซื้อเบียร์ จะซื้อผ้าอ้อมเด็กด้วย ซึ่งเป็นข้อมูลให้ทางห้างสรรพสินค้าคิดรายการส่งเสริมการขายใหม่ๆ หรือธนาคารอาจพบว่าคนทั่วไปที่มีอายุ 20-29 ปี และมีรายได้ในช่วง 20,000-30,000 บาท มักซื้อเครื่องเล่น MP3 ธนาคารอาจจะเสนอให้คนกลุ่มนี้ทำบัตรเครดิตโดยแถมเครื่องเล่นดังกล่าว เป็นต้น

ในปัจจุบันองค์กรส่วนใหญ่จะเผชิญกับปัญหาของ "ข้อมูลดิบจำนวนมาก แต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย ดาต้าไมนิ่ง(Data Mining) จึงเป็นสาขาที่คาดว่าจะเป็นที่รู้จัก และนำมาใช้ประยุกต์ได้อย่างแพร่หลาย เนื่องจากการทำเหมืองข้อมูล(Data Mining) สามารถดึงความรู้ออกมาจากข้อมูลจำนวนมากที่ถูกเก็บสะสมไว้

ในโลกของธุรกิจ บริษัทต่างๆ จะพยายามหาเทคนิคที่สามารถนำความสำเร็จมาสู่บริษัท เช่น ในโลกธุรกิจขนาดย่อมจะสร้างความสัมพันธ์กับลูกค้า โดยสังเกตจากความต้องการ ความชอบความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากผลสะท้อนในอดีตว่า จะทำอย่างไรให้ การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต หรือบริษัทที่เป็นผู้ออกบัตรเครดิต และธนาคารต่างๆ จะมีขบวนการทำเหมืองข้อมูล(Data Mining) ให้เป็นประโยชน์ในการตัดสินใจว่ากลุ่มลูกค้าใดเป็นกลุ่มที่ดีทำความเข้าใจลูกค้าช่วยในการแยกประเภทของลูกค้า และจะทำนายกลุ่มของธุรกิจประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต เป็นต้น อย่างไรก็ตามการเรียนรู้นั้นต้องมากกว่าการเก็บสะสมข้อมูลอย่างตรงไปตรงมา ซึ่งจะทำให้การทำงานไม่เป็นประสิทธิภาพ

2.1.2 วัฏจักรขั้นตอนการทำงานของการทำเหมืองข้อมูล (Data Mining)

วัฏจักรขั้นตอนการทำงานของการทำเหมืองข้อมูล (Data Mining) ประกอบด้วย 4 ขั้นตอนหลักๆ

(1) **การระบุโอกาสทางธุรกิจหรือการระบุปัญหาที่เกิดขึ้นกับธุรกิจ** เป็นการระบุขอบเขตของข้อมูลที่จะนำมาทำการวิเคราะห์ เพื่อหาความได้เปรียบทางการตลาดเพื่อนำมาทำการแก้ไขปัญหา

(2) **ส่วนของการทำเหมืองข้อมูล** เป็นการนำเทคนิคของการทำเหมืองข้อมูล (Data Mining) ไปใช้ถ่ายทอด หรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลที่จะนำไปใช้ได้จริงในทางธุรกิจ

(3) **การปฏิบัติตามข้อมูล** คือ การนำเอาข้อมูลที่เป็นผลลัพธ์ของส่วน Data Mining มาลองปฏิบัติจริงกับธุรกิจ

(4) **การวัดประสิทธิภาพจากผลลัพธ์** การวัดประสิทธิภาพของเทคนิค ในการทำเหมืองข้อมูล (Data Mining) ที่จะนำมาช้ากผลลัพธ์ ซึ่มสามารถตรวจสอบได้หลายทาง เช่นวัดจากส่วนแบ่งของตลาด วัดจากปริมาณลูกค้า หรือ วัดจากกำไรสุทธิ เป็นต้น จากทั้ง 4 ขั้นตอนที่กล่าว

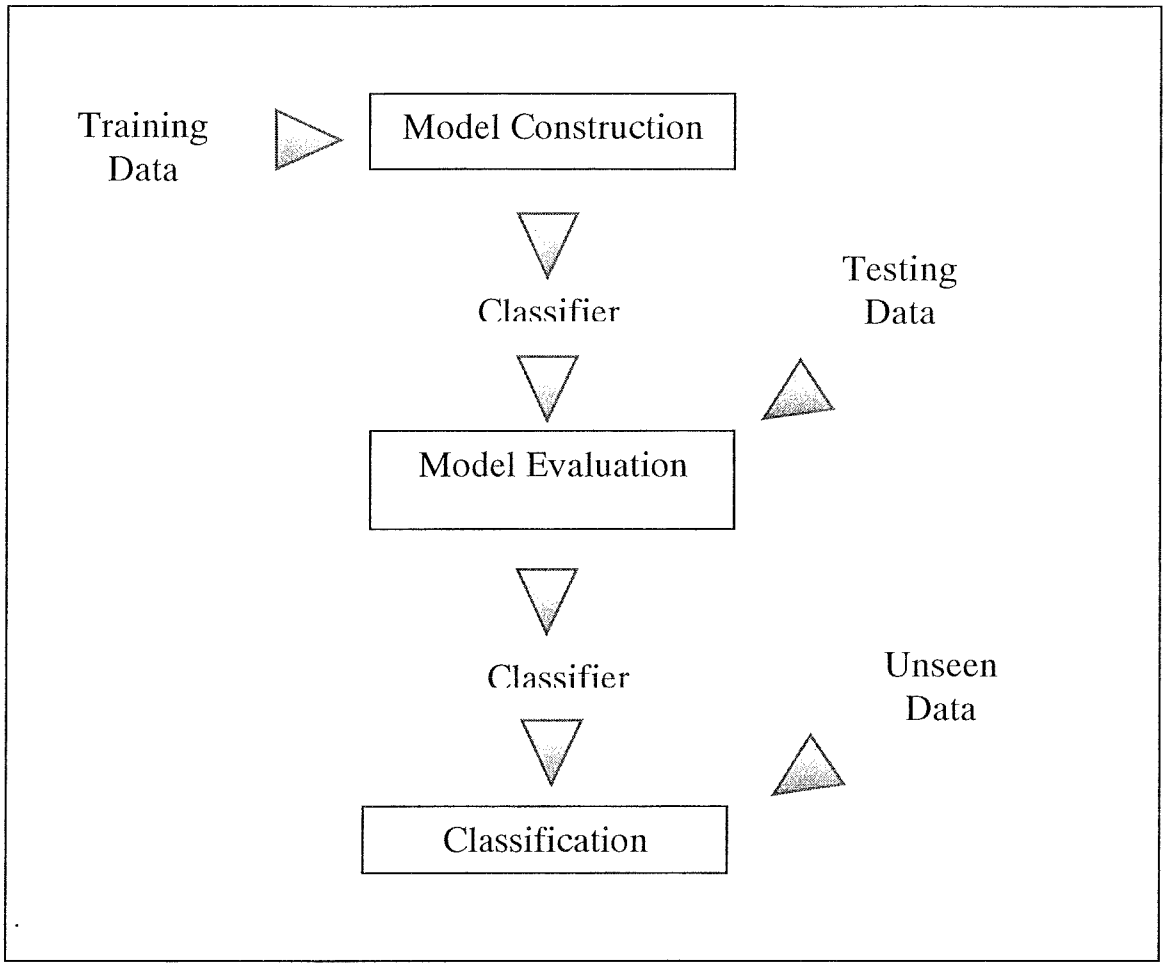
มาข้างต้น คือ การนำเอาการทำเหมืองข้อมูล (Data Mining) ไปใช้กับระบบทางธุรกิจ โดยแต่ละขั้นตอนจะพึ่งพาอาศัยกัน ผลลัพธ์จากขั้นตอนหนึ่งจะกลายมาเป็น Input จากอีกขั้นตอนต่อไปซึ่งการทำเหมืองข้อมูล (Data Mining) จะเปลี่ยนข้อมูลดิบให้เป็นข้อมูลประยุกต์ ดังนั้น การระบุแหล่งข้อมูลที่ถูกต้องจึงเป็นสิ่งที่สำคัญอย่างยิ่งต่อผลลัพธ์ที่ได้จากการวิเคราะห์

2.1.3 งานของการทำเหมืองข้อมูล (Task of data mining)

ในทางปฏิบัติการทำเหมืองข้อมูล (Data Mining) จะประสบความสำเร็จกับงานบางกลุ่มเท่านั้น และต้องอยู่ภายใต้ภาวะที่จัดปัญหาความเหมาะสมกับการใช้เทคนิคของการทำเหมืองข้อมูล (Data Mining) จะเป็นปัญหาที่ต้องใช้เหตุผลในการแก้ปัญหาเกี่ยวกับเศรษฐศาสตร์และการเงิน ซึ่งจะสามารถจัดรูปแบบของธุรกิจให้อยู่ในรูปแบบของงานทั้ง 6 งานได้ ดังนี้

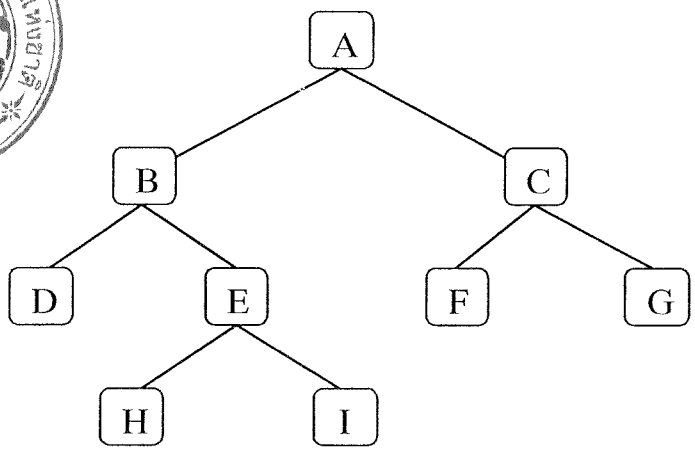
1. การแบ่งประเภท (Classification)

การแบ่งประเภทถือว่าเป็นงานธรรมดาทั่วไปของการทำเหมืองข้อมูล(Data Mining) เพราะการทำความเข้าใจ และการติดต่อสื่อสารต่างๆ ที่เกี่ยวข้องกับการแบ่งเป็นประเภท การจัดแยกประเภท และการแบ่งแยกชนิดโดยการแบ่งประเภท ประกอบด้วยการสำรวจจุดเด่นของวัตถุที่ปรากฏออกมา และทำการกำหนดจุดเด่นนั้นๆ เป็นตัวที่ใช้แบ่งประเภทของงานในการแบ่งประเภท คือ การบ่งบอกลักษณะ โดยการอธิบายจุดเด่นที่เป็นที่รู้จักดีในประเภทนั้นและเทรนนิ่งเซต (Training Set) ของตัวอย่างในแต่ละประเภท ซึ่งมีภาระหน้าที่ในการสร้างโมเดลของบางชนิดที่ไม่สามารถจะจัดประเภทของข้อมูลได้ ให้สามารถจัดเป็นประเภทได้ เช่น การจัดประเภทของผู้ยื่นบัตรเครดิต เป็นระดับต่ำ ระดับกลาง และระดับสูงของความเสี่ยงที่จะได้รับดังนั้นการแบ่งประเภทข้อมูล (Data Classification) เป็นปัญหาพื้นฐานของการเรียนรู้แบบมีผู้สอนโดยปัญหา คือ การทำนายประเภทของวัตถุจากคุณสมบัติต่างๆ ของวัตถุ ซึ่งการเรียนรู้แบบมีผู้สอนจะสร้างฟังก์ชั่นเชื่อมโยง ระหว่างคุณสมบัติของวัตถุกับประเภทของวัตถุ จากตัวอย่างสอนแล้วจึงใช้ฟังก์ชั่นทำนายประเภทของวัตถุที่ไม่เคยพบ เครื่องมือหรือขั้นตอนวิธีที่ใช้สำหรับการแบ่งประเภทของข้อมูล เช่น โครงข่ายประสาทเทียม(Neural Induction) ต้นไม้ตัดสินใจ(Decision Tree)



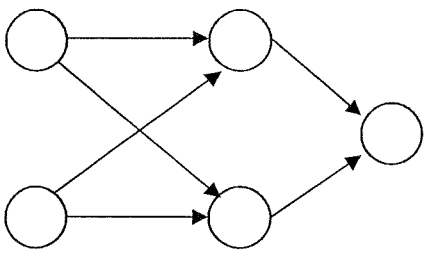
รูปที่ 2.4 กระบวนการ Classification

ต้นไม้ตัดสินใจ (Decision Tree) มีลักษณะคล้ายโครงสร้างต้นไม้ ซึ่งประกอบด้วย โหนดภายใน(Internal node) จะแสดงลักษณะ (Attribute) ของข้อมูล โดยจุดที่เริ่มต้นของต้นไม้เรียกว่าโหนดรากแต่ละกิ่งแสดงค่าของคุณลักษณะของแต่ละโหนด และลีฟโหนดแสดงกลุ่ม (Class) ซึ่งเป็นผลลัพธ์ที่สามารถแยกแยะได้



รูปที่ 2.5 โครงสร้างต้นไม้ตัดสินใจ

เครือข่ายใยประสาท (Neural Network) เป็นเทคโนโลยีที่มีมาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อใช้ในการคำนวณค่าฟังก์ชั่นจากกลุ่มข้อมูล วิธีการของเครือข่ายใยประสาทเป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบแล้วฝึก ให้ระบบได้รู้จักคิดแก้ปัญหาที่กว้างขึ้นได้ในโครงสร้างของเครือข่ายใยประสาทจะประกอบด้วยโหนด (node) สำหรับ Input-Output ปละการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ input layer , output layer และ hidden layer การประมวลผลของเครือข่ายใยประสาทจะอาศัยการทำงานผ่านโหนดต่างๆใน layer เหล่านี้ ตัวอย่างโครงสร้างแบบใยประสาท



Input Hidden Processing Layer Output

รูปที่ 2.6 โครงสร้างเครือข่ายประสาท

2. การประเมินค่า (Estimation)

การประเมินค่าทางธุรกิจอย่างต่อเนื่อง จะก่อให้เกิดผลลัพธ์ที่มีประโยชน์กับธุรกิจ การป้อนข้อมูลที่เรามีอยู่เข้าไป เพื่อใช้ในการประเมินสิ่งต่างๆ ที่จะก่อให้เกิดประโยชน์หรือสำหรับตัวแปรที่เราไม่รู้ค่าแน่นอน เช่นรายได้จากการค้า จุดสูงสุดทางธุรกิจ หรือดุลยภาพของบัตรเครดิตในทางปฏิบัติการประเมินค่าจะถูกใช้ในการทำงานการจัดหมวดหมู่ ตัวอย่างของการประเมินค่าเช่นการประเมินรายได้รวมของครอบครัว หรือการประเมินจำนวนบุตรในครอบครัว

3. การทำนายล่วงหน้า (Prediction)

การทำนายล่วงหน้า เป็นงานที่ลักษณะคล้ายกับการจัดหมวดหมู่ หรือ การประเมินค่า ยกเว้นเพียงแต่จะใช้สถิติการบันทึกของการจัดหมวดหมู่ในการทำนายอนาคตของพฤติกรรม หรือการประเมินค่าที่จะเกิดขึ้นในอนาคต ตัวอย่างของการทำงานการทำนายล่วงหน้า เช่น การทำนายการเปลี่ยนแปลงพฤติกรรมของตลาด หรือการทำนายจำนวนลูกค้าที่จะออกจากธุรกิจของเรา
ใน 6 เดือนข้างหน้า เป็นต้น

4. การจัดกลุ่มโดยอาศัยความใกล้ชิด (Affinity Group)

งานในการจัดกลุ่มหรือการวิเคราะห์ตลาด คือ การตัดสินใจรวมสิ่งที่สามารถไปด้วยกันเข้าไว้ในกลุ่มเดียวกัน ตัวอย่างของการจัดกลุ่มโดยอาศัยความใกล้ชิดกัน หรือ การวิเคราะห์ของตลาด เช่น การตัดสินใจว่ามีสิ่งใดบ้างที่จะไปอยู่ด้วยกันอย่างสม่ำเสมอในรถเข็นในซุปเปอร์

5. การแบ่งกลุ่ม (Clustering)

การแบ่งกลุ่มข้อมูล เป็นขั้นตอนเบื้องต้นของการวิเคราะห์ข้อมูล ซึ่งใช้ในการเรียนรู้ของเครื่อง การทำเหมืองข้อมูลโดยจะแบ่งชุดข้อมูลออกเป็นกลุ่ม นำข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน หรือความใกล้ชิด โดยคำนวณจากการวัดระยะห่างเวกเตอร์ของข้อมูลเข้า โดยใช้การวัดระยะแบบต่างๆเช่น การวัดระยะแบบยูคลิด การวัดระยะแบบแมนฮัตตัน การวัดระยะแบบเชบิเชฟ เพื่อช่วยในการลดขนาดข้อมูล(แยกเป็นหลายๆ กลุ่ม และคัดเฉพาะบางกลุ่มเพื่อทำการวิเคราะห์ต่อไป หรือแยกการวิเคราะห์ออกเป็นสำหรับแต่ละกลุ่ม)ก่อนที่จะนำไปวิเคราะห์ด้วยวิธีการอื่นต่อไป

การแบ่งกลุ่มข้อมูลจะแตกต่างกันจากการแบ่งประเภทข้อมูล(Classification) โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดประเภทของข้อมูลไว้ก่อน

ขั้นตอนวิธีการแบ่งกลุ่ม ได้แก่

-k-means clustering

-hierarchical clustering

-self-organizing map(som)

อัลกอริทึมในการแบ่งกลุ่มข้อมูล โดยทั่วไปแบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ

1.การแบ่งแบบเป็นลำดับขั้น(Hierarchical) จะมีการแบ่งกลุ่มจากกลุ่มย่อยที่ถูกแบ่งไว้ก่อนหน้านั้นซ้ำหลายครั้ง

2.การแบ่งเป็นสัดส่วน(Partitional) การจัดการ การแบ่งจะทำเพียงครั้งเดียว การแบ่งแบบเป็นลำดับขั้น จะมี 2 ลักษณะ คือ

-แบบล่างขึ้นบน(bottom-up) หรือ เป็นการแบ่งแบบรวมกลุ่มจากกลุ่มย่อยให้ใหญ่ขึ้นไปเรื่อยๆ โดยเริ่มจากกลุ่มเล็กสุด คือ ในแต่ละกลุ่มมีข้อมูลเพียงตัวเดียว

-แบบบนลงล่าง(top-down) หรือเป็นการแบ่งกลุ่มจากกลุ่มใหญ่ให้ย่อยไปเรื่อยๆโดยเริ่มจากกลุ่มใหญ่ที่สุด คือ กลุ่มเดียวมีข้อมูลทุกตัวในกลุ่ม

6. การบรรยาย (Description)

ในบางครั้งวัตถุประสงค์ของการทำเหมืองข้อมูล (Data Mining) คือ ต้องการอธิบายความสับสนของฐานข้อมูลในทางที่จะเพิ่มความเข้าใจในส่วนของประชากร ผลิตภัณฑ์ หรือขบวนการให้มากขึ้น

เทคนิคการทำเหมืองข้อมูล (Data Mining) ส่วนใหญ่ต้องการเทรนนิ่งข้อมูลจำนวนมากที่ประด้วยหลายๆตัวอย่าง เพื่อจะสร้างกฎที่ใช้ในการจัดหมวดหมู่ กฎความสัมพันธ์คลัสเตอร์ การทำนายล่วงหน้า ดังนั้นชุดของข้อมูลขนาดเล็กจะนำไปสู่ความไม่น่าไว้วางใจของผลสรุปที่ได้

ไม่มีเทคนิคใดสามารถแก้ปัญหาของ (Data Mining) ได้ทุกปัญหา ดังนั้นความหลากหลายของเทคนิคจึงเป็นสิ่งที่จำเป็นในการไปสู่การแก้ปัญหาของ (Data Mining) ได้ดีที่สุด

2.1.4 เทคนิคของการทำเหมืองข้อมูล (Data Mining)

การแก้ปัญหาของงานชนิดต่างๆ โดยใช้วิธีการทำเหมืองข้อมูล (Data Mining) ในแต่ละงานก็จะมีเทคนิคของการทำเหมืองข้อมูล (Data Mining) ที่จะนำมาใช้ได้อย่างเหมาะสม โดยเทคนิคในการทำเหมืองข้อมูล (Data Mining) นั้นมีมากมาย ซึ่งเทคนิคที่ถูกใช้กันอย่างแพร่หลายมีดังนี้

1. นิวตรอนเน็ตเวิร์ค

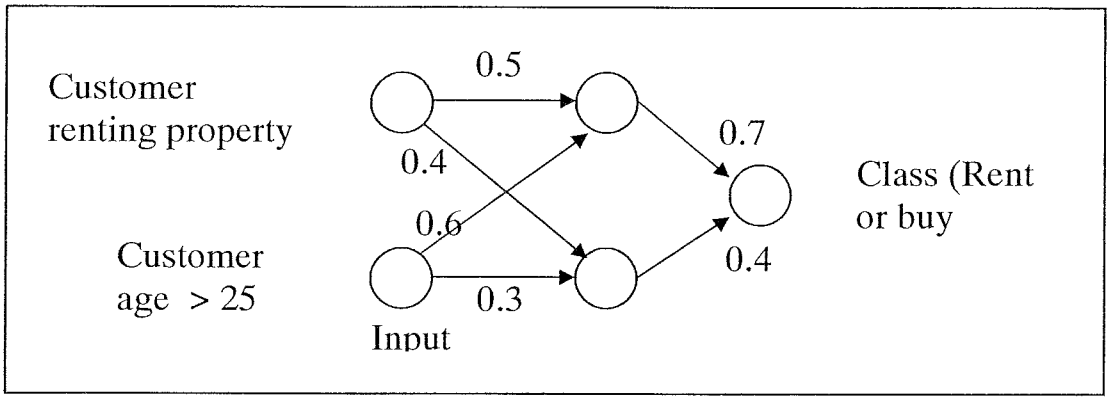
นิวตรอนเน็ตเวิร์ค คือ ระบบที่มีการประมวลผลข้อมูลซึ่งรวมคุณสมบัติของไบโอโลจิคอล โดยถูกพัฒนาขึ้นด้วยโมเดลทางคณิตศาสตร์ของกระบวนการเรียนรู้ของมนุษย์ และเรียนรู้จากชุดข้อมูลของชุดความรู้

นิวตรอนเน็ตเวิร์ค ประกอบด้วย หน่วยความจำจำนวนมากเรียกว่า นิวตรอนเซล หรือโหนด แต่ละนิวตรอนต่อกัน โดนคอนเน็กชั่นลิ้งค์ (Connection Link) ที่ค่าน้ำหนักของมันอยู่โดยค่าน้ำหนักจะแสดงรายละเอียดที่เน็ตเวิร์คใช้ในการแก้ปัญหา โดยนิวตรอนเน็ตเวิร์คถูกใช้ในการแก้ปัญหาอย่างกว้างขวาง เช่น การเก็บ และการเรียกข้อมูล การแยกประเภทของข้อมูล การ

เปลี่ยนรูปแบบของ input ให้อยู่ในรูปแบบของ output ความสามารถในการตรวจสอบรูปแบบของข้อมูลที่คล้ายคลึงกับความคิดมนุษย์ เป็นต้น แต่นิวตรอนเน็ตเวิร์คสามารถนำไปประยุกต์ใช้กับงานหลายๆชนิดได้อย่างมีประสิทธิภาพ แต่นิวตรอนเน็ตเวิร์คก็ยังมีข้อเสียอยู่บ้าง ดังนี้

1.เป็นวิธีที่ยากต่อการทำความเข้าใจในโมเดลที่ถูกผลิตออกมา

2.มีคุณสมบัติที่ไวต่อรูปแบบของ input ถ้าเราแทนข้อมูลด้วยรูปแบบที่แต่ต่างกันก็จะสามารถผลิตผลลัพธ์ที่แตกต่างกันออกมา ดังนั้นการกำหนดค่าเริ่มต้นให้กับข้อมูลจึงเป็นส่วนที่มีความสำคัญส่วนหนึ่ง



รูปที่ 2.7 นิวรอลเน็ต เพื่อการวิเคราะห์การเช่าและซื้อบ้าน

2. จีเนติก อัลกอริทึม (Genetic Algorithm : GA)

ในยีนส์ของมนุษย์นั้น จะมีการถ่ายทอดพันธุกรรมไปยังลูกหลานได้ ซึ่งจีเนติกอัลกอริทึม จะอาศัยหลักการนี้ ข้อมูลชุดหนึ่ง ซึ่งมีกฎของตัวมันเอง หากมีการนำข้อมูลทั้งสองชุดมารวมกันเป็นรูปแบบนี้ จะมีการสร้างกฎขึ้นมาโดยวิเคราะห์จากกฎที่มีอยู่ของทั้ง 2 รูปแบบข้อมูล

จีเนติก อัลกอริทึม เป็นทฤษฎีที่จำลองกระบวนการวิวัฒนาการทางธรรมชาติ คือการคัดเลือกทางธรรมชาติ โดยอาศัยพื้นฐานความคิดทางพันธุกรรมในการถ่ายทอดลักษณะต่างๆไปยังรุ่นถัดไปที่สามารถนำมาพัฒนาใช้ในการหาคำตอบที่เหมาะสมที่สุด ของแต่ละปัญหาจีเนติกอัลกอริทึมเป็นวิธีการหาคำตอบโดยการพิจารณา และดำเนินการจากกลุ่มของคำตอบของปัญหาที่ถูกสร้างขึ้นมาโดยการเข้ารหัส คือ การแปลงค่าตัวแปรหรือพารามิเตอร์ของปัญหาให้อยู่ในรูปโครงสร้างของโครโมโซมที่กำหนด เพื่อคัดเลือกโครโมโซมที่เหมาะสมสำหรับสร้างวิวัฒนาการให้ดีขึ้นตามกระบวนการทางพันธุศาสตร์ โดยการแลกเปลี่ยนค่าพารามิเตอร์ต่างๆ ระหว่างโครโมโซมที่ถูกคัดเลือก อันจะทำให้คำตอบของปัญหาถูกปรับปรุงให้ดีขึ้น จีเนติก อัลกอริทึม ใช้กระบวนการหลักๆ 3 กระบวนการในการหาคำตอบที่ใกล้เคียงหรือดีที่สุดของปัญหาดังนี้

1.การคัดเลือก (Selection) : คัดเลือกอันที่ดีที่สุดซึ่งจริงๆแล้วถ้าผ่านขั้นตอนต่อไปอาจจะเป็นค่าที่ใช้ไม่ได้ก็ได้

2.การครอสโอเวอร์ (Crossover) : ค่าจากขั้นตอนที่ 1 มาสับเปลี่ยนบิต(bit) เพื่อประเมินแล้วเลือก

3.การมิวเตชั่น (Mutation) : ค่าจากขั้นตอนที่ 3 นำกลับมาบิตจาก 0 เป็น 1 เพื่อประเมินแล้วกลับไปขั้นตอนที่ 1 อีกครั้ง ถึงแม้ว่าในปัจจุบัน จีเนติก อัลกอริทึม ยังเป็นวิธีการที่ไม่ได้แพร่หลาย แต่สาขาวิชาทางด้านจีเนติกอัลกอริทึม นับว่าเป็นอีกสาขาวิชาหนึ่งที่สนใจ และน่าจะเป็นวิธีที่ได้รับความนิยมในอีกไม่กี่ปีข้างหน้า เนื่องมาจากสามารถนำมาประยุกต์ใช้ได้กับหลายๆปัญหา รวมทั้งปัญหาทางการทำเหมืองข้อมูล

3. ดีซีซันทรี (Decision Tree)

เป็นแบบจำลองที่มีลักษณะคล้ายกับต้นไม้ จะมีการสร้างกฎต่างๆ ขึ้นเพื่อใช่ในการตัดสินใจ ดีซีซันทรี เป็นวิธีที่ได้รับความนิยม เนื่องจากความไม่ซับซ้อนของอัลกอริทึม ทำให้เครื่องมือที่ใช้ในการทำวางขายกันอยู่ในท้องตลาดต่างๆ ข้อดีของวิธีนี้คือ สามารถตีความและเข้าใจลักษณะของรูปแบบข้อมูลได้ง่าย เพราะมีการแยกออกเป็นกฎ หรือข้อกำหนดต่างๆ แต่ก็ยังคงมีปัญหาในเรื่องของการห้ามน้ำหนักความน่าเชื่อถือ หรือ การให้ค่าน้ำหนักในแต่ละโหนด ซึ่งถ้าให้น้ำหนักผิดไป อาจจะทำให้การตีความผิดไปได้

4. คลัสเตอร์ลิ่ง (Clustering)

วิธีคลัสเตอร์ลิ่ง นี้เป็นวิธีที่อาจเรียกว่าเป็นการทำเหมืองข้อมูลแบบ อ้อมๆ ได้ เนื่องจากการหาผลลัพธ์ในแต่ละครั้งนั้น แม้กระทั้งผู้หายังไม่อาจจะทราบว่าสิ่งที่ต้องการจะหานั้นคืออะไร จำเป็นต้องรอจนกว่าการค้นหาจะทำเสร็จสมบูรณ์จึงจะทราบข้อมูลที่ซ่อนอยู่ เปรียบเสมือนการมีข้อมูลจำนวนมากมายอยู่ในตะกร้า แล้วจากนั้นมีเวทย์มนต์มาจัดเรียงข้อมูลหน่วยนั้นให้อยู่เป็นกลุ่ม ซึ่งทำให้สังเกตลักษณะเด่นที่ซ่อนเร้นอยู่ภายในข้อมูลจำนวนมากหน่วยนั้น

ตัวอย่าง เช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1.กลุ่มผู้มีรายได้สูง (>$80,000)

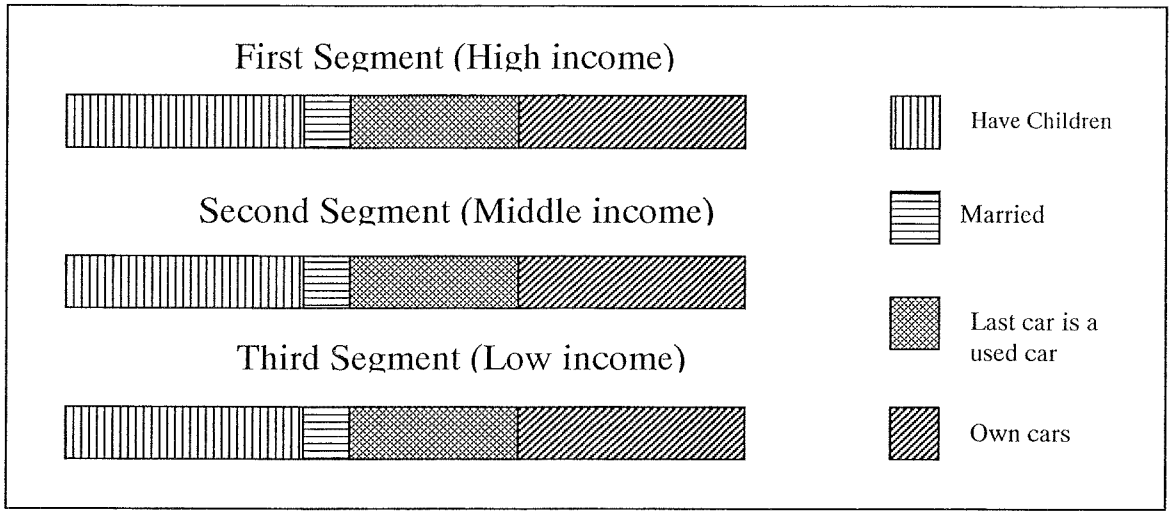
2.กลุ่มผู้มีรายได้ปานกลาง ($25,000 to 80,000)

3.กลุ่มผู้มีรายได้ต่ำ(less than $25,000)

และภายในแต่ละกลุ่มยังแยกออกเป็น

-Have Children

-Married



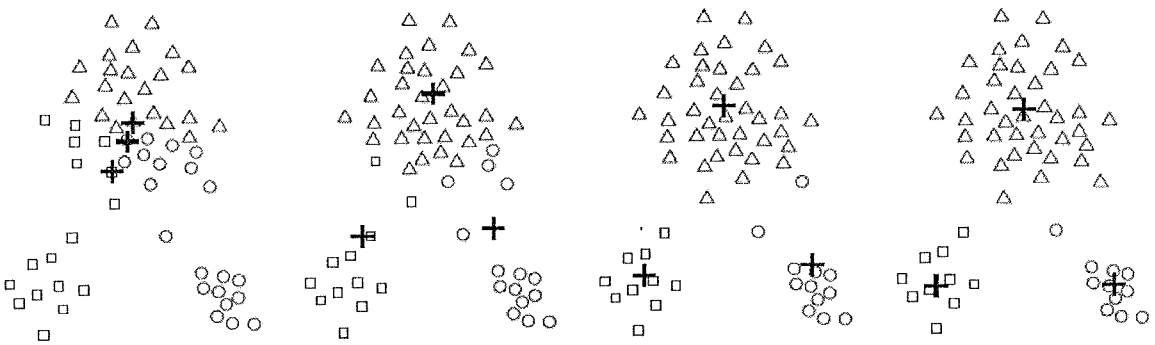
รูปที่ 2.8 ตัวอย่าง Clustering

จากข้อมูลข้างต้น ทำให้ทางบริษัทรู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัท ควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

การแบ่งกลุ่มข้อมูล (Data Clustering) เป็นศาสตร์แขนงหนึ่งทางเหมืองข้อมูล (Data Mining) เป็นวิธีการแบ่งกลุ่มข้อมูลโดยไม่อาศัยตัวอย่างในการเรียนรู้ (unsupervised learning) โดยจะทำการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ เรียกว่าคลัสเตอร์ (cluster) โดยพิจารณาจากความคล้ายคลึงกันของข้อมูลเป็นเกณฑ์หลักในการแบ่งกลุ่มข้อมูล ดังนั้นกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันจะถูกจัดให้อยู่ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะแตกต่างกันจะถูกจัดไว้ต่างกลุ่มกัน วิธีการแบ่งกลุ่มนี้มีประโยชน์อย่างมากในการวิเคราะห์ความสัมพันธ์กันของชุดข้อมูล เทคนิคการแบ่งกลุ่มในปัจจุบันมีหลากหลายเทคนิค แต่ที่ใช้กันแพร่หลายได้แก่ การแบ่งกลุ่มแบบพาร์ทิชั่น (partition based) การแบ่งกลุ่มแบบใช้ความหนาแน่น (density based) การแบ่งกลุ่มโดยใช้กริด (grid based) และการแบ่งกลุ่มแบบมีลำดับชั้น (hierarchical based) ซึ่งในงานวิจัยนี้ใช้วิธีการแบ่งกลุ่มแบบพาร์ทิชั่น ด้วยอัลกอริทึม K-Means ซึ่งจะกล่าวถึงรายละเอียดในส่วนถัดไป

4.1 K-Mean Clustering

เทคนิคการจัดกลุ่ม K-means เป็นวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่ง่ายที่สุด หลักการของการจัดกลุ่มด้วย K-Means นั้น อัลกอริทึมจะทำการแบ่งข้อมูล (Partition) ออกเป็นกลุ่ม จำนวน K กลุ่ม (Cluster)ตามที่ผู้ใช้งานกำหนด โดยในแต่ละกลุ่มจะมีข้อมูลจุดศูนย์กลางประจำกลุ่ม (centroid) ซึ่งได้จากการคำนวณค่าเฉลี่ยของกลุ่ม และการพิจารณาว่าข้อมูลหนึ่งๆ ควรอยู่กลุ่มใดนั้น จะทำการพิจารณาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางประจำกลุ่มแต่ละกลุ่ม ข้อมูลที่มีระยะห่างจากจุดศูนย์กลางประจำกลุ่มใดน้อยที่สุดจะถูกจัดอยู่ในกลุ่มนั้น เมื่อข้อมูลทุกตัวถูกพิจารณาจัดกลุ่มแล้ว อัลกอริทึมจะทำการคำนวณหาค่าเฉลี่ยจุดศูนย์กลางประจำกลุ่มใหม่ หากจุดศูนย์กลางของแต่ละกลุ่มมีการเปลี่ยนตำแหน่ง อัลกอริทึมจะทำการวนซ้ำเพื่อจัดกลุ่มข้อมูลตามจุดศูนย์กลางกลุ่มที่เปลี่ยนไป และจะหยุดกระบวนการเมื่อจุดศูนย์กลางประจำกลุ่มของทุกกลุ่มไม่มีการเปลี่ยนแปลง สุดท้ายจะได้ข้อมูลที่มีลักษณะเหมือนหรือคล้ายกันอยู่ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะต่างกันอยู่คนละกลุ่มกัน



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

รูปที่ 2.9 แสดงขั้นตอนการจัดกลุ่มข้อมูลด้วยเทคนิค Clustering

จากรูปที่ 2.9 แสดงการจัดกลุ่มข้อมูล โดยรูปที่ 2.9(a) เริ่มจากกำหนดผู้ใช้งานทำการกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มข้อมูลเป็น 3 กลุ่ม จากนั้นอัลกอริทึมทำการสุ่มกำหนดจุดศูนย์กลางประจำกลุ่มเริ่มต้น แทนด้วยสัญลักษณ์ + อัลกอริทึม K-Mean ทำการคำนวณระยะห่างระหว่างข้อมูลและจุดศูนย์กลางกลุ่มแต่ละกลุ่ม หากข้อมูลห่างจากกลุ่มใดน้อยที่สุดจะถูกจัดว่าเป็นสมาชิกของกลุ่มนั้น รูปที่ 2.9 (b) และ รูปที่ 2.9 (c) อัลกอริทึมคำนวณจุดศูนย์กลางประจำกลุ่มใหม่ และพบว่ามีการเปลี่ยนแปลง จึงทำการวนซ้ำเพื่อคำนวณระยะห่างระหว่างข้อมูลและจุดศูนย์กลาง และพิจารณาการจัดกลุ่มข้อมูลใหม่ และเมื่อพบว่ากลุ่มข้อมูลไม่มีการเปลี่ยนแปลงแล้วจึงหยุดการทำงาน และได้ผลลัพธ์รูปที่ 2.9 (d)

อัลกอริทึมการจัดกลุ่มโดย K-means

1. ผู้ใช้งานกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มข้อมูลทั้งหมดออกเป็น K กลุ่ม
2. อัลกอริทึมทำการสุ่มกำหนดจุดศูนย์กลางประจำกลุ่มข้อมูลแต่ละกลุ่ม
3. อัลกอริทึมทำการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางประจำกลุ่มทุกกลุ่ม เพื่อพิจารณาความเหมือนหรือความคล้ายกันของข้อมูล
4. ข้อมูลที่มีระยะห่างจากจุดศูนย์กลางของกลุ่มใดน้อยที่สุด จะถูกจัดอยู่กลุ่มเดียวกับกลุ่มของจุดศูนย์กลางประจำกลุ่มนั้น
5. เมื่อข้อมูลทุกตัวถูกจัดกลุ่มครบแล้ว อัลกอริทึมจะคำนวณจุดศูนย์กลางประจำกลุ่ม K กลุ่มใหม่ โดยพิจารณาค่าเฉลี่ยของข้อมูลทุกตัวที่อยู่ในกลุ่ม
6. ทำซ้ำในข้อ c จนกระทั่งจุดศูนย์กลางประจำกลุ่มไม่เปลี่ยนแปลง

4.2 ฟังก์ชันวัดระยะห่างระหว่างข้อมูล (Dissimilarity Function)

ฟังก์ชันวัดระยะห่างของข้อมูลใช้ในการพิจารณาความเหมือน หรือความคล้ายกันของข้อมูล โดยฟังก์ชันที่นิยม สำหรับใช้ในเทคนิคการจัดกลุ่ม ได้แก่ ฟังก์ชันระยะห่างยูคลิเดียน (Euclidean Distance) ฟังก์ชันระยะห่างแมนฮัตตัน (Manhattan Distance) หรือ ซิตี้บลอค (City-Block) เป็นต้น

4.2.1 Euclidean Distance

เป็นฟังก์ชันระยะห่างที่นิยมใช้ในเทคนิค K-Mean โดยมีสูตรการคำนวณดังนี้

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

โดยที่ x คือข้อมูลใดๆ ที่ต้องการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางกลุ่มของ Cluster K

y คือข้อมูลที่เป็นจุดศูนย์กลางกลุ่มของ Cluster K

m คือจำนวนแอทริบิวต์

4.2.2 Manhattan Distance หรือ City-Block Distance

เป็นฟังก์ชันวัดระยะห่างโดยพิจารณาความแตกต่างของมิติค่าเฉลี่ย โดยมีสูตรการคำนวณดังนี้

$$d_{manhattan}(x,y) = \frac{1}{m}\sum_{i=1}^{m}|x_i - y_i|$$

โดยที่ $\mathbf{x}$ คือข้อมูลใดๆ ที่ต้องการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางกลุ่มของ Cluster K

y คือข้อมูลที่เป็นจุดศูนย์กลางกลุ่มของ Cluster K

m คือจำนวนแอทริบิวต์

2.1.5 เทคนิคการจำแนกประเภทข้อมูล JRip

เทคนิค JRip เป็นเทคนิคที่ใช้สำหรับจำแนกประเภทข้อมูล (Classification) ที่ถูกพัฒนามาจากเทคนิค IREP [1] ซึ่งเทคนิค JRip นี้จะประกอบด้วย 3 ขั้นตอน คือ

1. ขั้นการสร้างกฎเพื่อใช้จำแนกประเภทข้อมูล
2. ขั้นการปรับแต่งกฎ
3. ขั้นการลบกฎ

การทำงานของแต่ละขั้นมีดังต่อไปนี้

เริ่มแรกจะกำหนดให้โมเดลที่เก็บกฎหรือ RuleSet เป็นเซตว่างเสียก่อน หลังจากนั้น

ขั้นที่ 1 จะเป็นขั้นตอนที่ทำการสร้างกฎเพื่อเพิ่มเข้าไปใน RuleSet และกฎเหล่านี้จะใช้ในการจำแนกประเภทข้อมูล โดยการทำงานของขั้นนี้จะวนรอบทำจนข้อมูลที่อยู่ในคลาส Pos (สมมุติว่าข้อมูลมี 2 คลาส คือ คลาส Pos และ คลาส Neg) ถูกนำมาสร้างกฎทุกตัวแล้วหรือว่าค่าความผิดพลาด (error rate) ของโมเดลหรือ RuleSet ที่สร้างได้ มีค่าเกินกว่า 50% ในขั้นตอนที่ 1 นี้จะประกอบด้วย 2 ขั้นตอนย่อย คือ

1. *ขั้น Grow* ในขั้นนี้จะเป็นการเพิ่มแอตทริบิวต์ต่างๆ เข้าไปในส่วนเงื่อนไขของกฎ โดยการเพิ่มเข้าไปนี้จะเป็นการเพิ่มแบบ greedy คือ ถ้าเพิ่มแอตทริบิวต์เข้าไปแล้วทำให้ค่าความถูกต้องเพิ่มมากขึ้นก็จะเพิ่มแอตทริวบิวต์นั้นเข้าไปในกฎเรื่อยๆ ทำเช่นนี้จนกระทั่งได้กฎที่ดี (เช่น อาจจะได้กฎที่มีความถูกต้อง 100%) โดยค่าในแต่ละแอตทริบิวต์ที่เพิ่มเข้าไปจะต้องเป็นค่าที่ทำให้ค่า information gain มีค่ามากที่สุด

2. *ขั้น Prune* กฎที่สร้างได้จากขั้นตอนที่ผ่านมาจะถูกนำมา prune โดยจะลบบางแอตทริบิวต์ออกแล้วทำการตรวจสอบว่ากฎที่ถูก prune แล้วนี้ครอบคลุมข้อมูลของคลาส Pos และคลาส Neg เพิ่มมากขึ้นเท่าใดและการ prune นี้จะสิ้นสุดเมื่อกฎที่ prune ได้ครอบคลุมข้อมูลเป็นจำนวนลดลง หลังจากนั้นจึงนำกฎที่ prune ได้นี้เก็บเข้าไปใส่ไว้ใน RuleSet เพื่อใช้เป็นโมเดลสำหรับการจำแนก

ประเภทข้อมูลต่อไป และทำการลบข้อมูลที่กฎนี้ครอบคลุมอยู่

ขั้นที่ 2 เป็นขั้นการปรับแต่งกฎที่อยู่ใน RuleSet ซึ่งสร้างได้จากขั้นที่ 1 ในขั้นนี้จะสร้างกฎเพิ่มขึ้นมาอีก 2 แบบคือกฎแบบ replacement และกฎแบบ revision โดยแต่ละแบบมีลักษณะดังนี้

- กฎแบบ replacement จะใช้การสร้างกฎตามขั้นที่ 1.1 แต่ในขั้นตอนการ prune กฎจะพิจารณาผลรวมความถูกต้องของทุกกฎใน RuleSet แทน
- กฎแบบ revision จะทำการเพิ่มแอตทริบิวต์เข้าไปในกฎที่สร้างได้จากขั้นที่ 1 แล้ว และในขั้นตอนการ prune กฎจะพิจารณาผลรวมความถูกต้องของทุกกฎใน RuleSet

สิ่งที่แตกต่างกันระหว่างกฎแบบ replacement และกฎแบบ revision คือขั้นตอนการเพิ่มแอตทริบิวต์ซึ่งแบบแรกจะเพิ่มเข้าไปในกฎที่ว่างๆ ในขณะที่แบบหลังจะเพิ่มเข้าไปในกฎเดิมที่มีอยู่ จากกฎทั้ง 3 แบบ (กฎที่ได้จากขั้นที่ 1, กฎแบบ replacement และกฎแบบ revision) นี้กฎที่มีค่า description length น้อยที่สุดจะถูกเลือกเพื่อใช้เป็นโมเดลต่อไป และถ้ายังมีข้อมูลอยู่ในคลาส Pos อยู่ก็จะกลับไปทำขั้นที่ 1 ต่อไปอีก

ขั้นที่ 3 เป็นการลบกฎบางกฎซึ่งทำให้ description length ของ RuleSet เพิ่มขึ้นจึงทำการลบออกจาก RuleSet

2.1.6 ประเภทของข้อมูลที่สามารถทำเหมืองข้อมูล

ประเภทของข้อมูลที่สามารถทำเหมืองข้อมูล (Data Mining) มีดังต่อไปนี้

(1) Relational Database

เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถว และคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดยแบบจำลองฐานข้อมูลเชิงสัมพันธ์ (Entity-Relationship (E-R) Model)

(2) Data Warehouses

เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกัน และรวบรวมไว้ในที่เดียวกัน ซึ่งสะดวกต่อการนำข้อมูลไปไมน์ต่อไป

(3) Transactional Database

ประกอบด้วยข้อมูลที่แต่ละทรานเซกชันแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้า และรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

(4) Advance Database

เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น ข้อมูลแบบ Object-oriented ข้อมูลที่

เป็นเท็กซ์ไฟล์ ข้อมูลมัลติมีเดีย ข้อมูลในรูปของเว็บ เป็นต้น

2.1.7 ความรู้ที่ได้จากการทำเหมืองข้อมูล

ความรู้ที่ได้จากการทำเหมืองข้อมูลมีหลายรูปแบบ ได้แก่

(1) กฎเชื่อมโยง (Association Rule)

กฎเชื่อมโยง แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุ ที่เกิดขึ้นพร้อมกัน กฎเชื่อมโยงมักจะใช้ในการวิเคราะห์ข้อมูลการขายสินค้าโดยเก็บข้อมูลจากระบบ ณ จุดขาย หรือระบบขายสินค้าออนไลน์ และพิจารณาสินค้าที่อยู่ในตะกร้าเดียวกัน หรือผู้ซื้อมักจะซื้อพร้อมกันเช่น ถ้าพบว่าคนที่ซื้อเทปวีดีโอมักจะซื้อเทปกาวด้วย ร้านค้าอาจจะจัดร้านให้สินค้าทั้งสองอย่างอยู่ใกล้กันเพื่อเพิ่มยอดขาย หรือ อาจพบว่าคนที่ซื้อหนังสือ เอ แล้วหลังจากนั้นมักจะซื้อหนังสือ บีสามารถนำกฎนี้ไปแนะนำผู้ที่กำลังจะซื้อหนังสือ เอ ได้

(2) การแบ่งประเภทข้อมูล (Data Classification)

เพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น หาความสัมพันธ์ระหว่างผลการตรวจร่างกายต่างๆ กับการเกิดโรค โดยใช้ข้อมูลผู้ป่วย และการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วย หรือการวิจัยทางการแพทย์ ในทางธุรกิจจะใช้เพื่อดูคุณสมบัติของผู้ที่จะก่อหนี้ดีหรือหนี้เสีย เพื่อประกอบการพิจารณาการอนุมัติเงินกู้

(3) การแบ่งกลุ่มข้อมูล (Data Clustering)

การแบ่งกลุ่มข้อมูล แบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม แบ่งกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรค โดยพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

(4) จินตทัศน์ (Visualization)

สร้างภาพคอมพิวเตอร์กราฟิกที่สามารถนำเสนอข้อมูลมากมายอย่างครบถ้วนแทนการใช้ข้อความนำเสนอข้อมูลที่มากมาย เราอาจพบข้อมูลที่ซ้อนเร้นเมื่อดูข้อมูลชุดนั้นด้วยจินตทัศน์ ตัวอย่างเช่น ถ้าเรามีข้อมูลชุดหนึ่งที่มีข้อมูลอยู่จำนวน 100,000 รายการ แต่เราใช้รูปแบบการนำเสนอเป็นข้อความ ผลลัพธ์คงจะไม่เหมาะสม เพราะจะทำให้เราต้องสร้างรายงานจำนวนมหาศาล กล่าวคือถ้าเราสามารถใช้ภาพคอมพิวเตอร์กราฟิกมาช่วยในการนำเสนอข้อมูลชุดนี้แทนการใช้ข้อความผลลัพธ์คือเราไม่ต้องสร้างรายงานจำนวนมหาศาล การนำเสนอข้อมูลด้วยจินตทัศน์เพียง 1 ภาพ แต่สามารถแทนค่าของชุดข้อมูลจำนวนมหาศาลได้ หรือกล่าวอีกนัยหนึ่งคือ ภาพเพียง1 ภาพ อาจใช้แทนข้อมูลข่าวสารได้เป็น 100 คำ หรือ 1,000 คำ จึงเป็นแนวคิดของการทำจินตทัศน์ นอกจากนี้เรายังสามารถทำการโต้ตอบกับจินตทัศน์ได้อีกด้วย ซึ่งการนำเสนอข้อมูลในรูปแบบข้อความไม่สามารถโต้ตอบได้ข้อมูลที่มีปริมาณมหาศาลนำเสนอให้เห็นในแบบจินตทัศน์ อาจทำให้เราพบข้อมูลที่ซ่อนเร้นอยู่ ยกตัวอย่างเช่น ถ้าเรานำชุดข้อมูล IRIS ซึ่งเป็นข้อมูลตัวเลขของดอกไม้ชนิดหนึ่งจำนวน 150

รายการ นำชุดข้อมูลชุดนี้มาแสดงให้เห็นในแบบจินตทัศน์ โดยวิธี scatter plot เราจะเห็นว่าจุด plot ที่แสดงมีการเกาะกลุ่มกันจำนวน 3 กลุ่ม บางส่วนของกลุ่มที่ 1Intersection กับบางส่วนของกลุ่มที่ 2 ส่วนกลุ่มที่ 3 ไม่มีส่วนใด Intersection กับกลุ่มที่ 1 และ กลุ่มที่ 2 แสดงให้เห็นว่าชุดข้อมูลดอกไม้ IRIS นี้มีจำนวน 3 กลุ่มที่คล้ายคลึงกัน และบางสมาชิกของกลุ่มแรกมีความคล้ายคลึงกับบางสมาชิกของกลุ่มที่ 2 ส่วนสมาชิกของกลุ่มที่ 3 มีความแตกต่างกับสมาชิกในกลุ่มที่ 1และกลุ่มที่ 2 อย่างมีนัยสำคัญ

2.1.8 ประโยชน์ของการทำเหมืองข้อมูล

การทำเหมืองข้อมูลจำเป็นต้องอาศัยบุคลากรจากหลายฝ่าย และต้องอาศัยความรู้จำนวนมากถึงจะได้รับประโยชน์อย่างแท้จริง เพราะสิ่งที่ได้จากขั้นตอนวิธีเป็นเพียงตัวเลข และข้อมูลที่อาจจะนำไปใช้ประโยชน์ได้ หรือ ใช้ประโยชน์อะไรไม่ได้เลยก็เป็นไปได้ ผู้ที่ศึกษาการทำเหมืองข้อมูลจึงควรมีความรู้รอบด้าน และต้องติดต่อกับทุกๆฝ่าย เพื่อให้เข้าใจถึงขอบเขตของปัญหาโดยแท้จริงก่อน เพื่อให้การทำเหมืองข้อมูลเกิดประโยชน์อย่างแท้จริง ซึ่งสามารถแบ่งได้
ดังนี้

(1) ค้นหาข้อมูลโดยอาศัยเทคโนโลยีของเหมืองข้อมูล
(2) ใช้สถาปัตยกรรมแบบ Client/Server
(3) ผู้ใช้ระบบไม่จำเป็นต้องมีทักษะในการเขียนโปรแกรม
(4) ผู้ใช้ต้องกำหนดขอบเขต และเป้าหมายของระบบให้ชัดเจนเพื่อความรวดเร็วและถูกต้องตามความต้องการ
(5) การประมวลผลแบบขนานจะช่วยเพิ่มประสิทธิภาพ และความเร็วในการค้นหา

2.2 การสุ่มตัวอย่าง

การวิจัยทางสังคมศาสตร์จำเป็นต้องอาศัยวิธีวิทยาศาสตร์ในการคัดเลือกกลุ่มเป้าหมายที่จะทำการศึกษาซึ่งสามารถทำได้โดยการอาศัยการสุ่มตัวอย่าง (Sampling) การสุ่มตัวอย่างเป็นการคัดเลือกจากประชากรทั้งหมด โดยสุ่มตัวอย่างมาเพียงส่วนหนึ่ง เป็นตัวแทนของประชากรทั้งหมดเพื่อนำมาศึกษา

2.2.1 องค์ความรู้ในการสุ่มตัวอย่าง

1. ข้อมูลประชากร (Population) หมายถึง กลุ่มเป้าหมายที่ต้องการศึกษาทั้งหมด ซึ่งอาจจะเป็น คน สัตว์ พืช วัตถุ หรือปรากฏการณ์ต่างๆ เช่น ในการศึกษาความรู้ในการประกอบอาชีพด้านหม่อนไหมของเกษตรกรผู้ปลูกหม่อนเลี้ยงไหมในเขต ภาคอีสานตอนบน ประชากรในที่นี้คือ

เกษตรกร ที่มีภูมิลำเนาอยู่ในเขตจังหวัดต่างๆ ของภาคอีสานตอนบนในการวิจัยเชิงสังคมศาสตร์ ประชากรแบ่งออกได้ 2 ประเภทดังนี้

1.1 ประชากรที่มีจำนวนจำกัด (Finite Population) หมายถึงประชากรที่มีปริมาณซึ่สามารถนับออกมาเป็นตัวเลขได้ครบถ้วนเช่น ประชากรนิสิต หรือนักศึกษา ของมหาวิทยาลัยทุกแห่ง ประชากรของเกษตรกรในภาคกลาง ฯลฯ

1.2 ประชากรที่มีจำนวนไม่จำกัด (Infinite Population) หมายถึงประชากรที่มีปริมาณซึ่งไม่สามารถนับจำนวนออกมาเป็นตัวเลขได้ครบถ้วน เช่น ประชากรเมล็ดถั่วเหลืองที่จำหน่ายในจังหวัดขอนแก่น ฯลฯ

2. ขนาดตัวอย่าง (Sample Size) ขนาดตัวอย่างต้องมากพอที่จะเป็นตัวแทนได้ วิธีการประมาณขนาดตัวอย่างโดยใช้สูตรของ TARO YAMANE ดังนี้

$$\frac{n = N}{1 + Nd^2}$$

เมื่อ n = ขนาดของหน่วยตัวอย่างกลุ่มเป้าหมาย

N = ประชากรทั้งหมด

D = ระดับความมีนัยสำคัญ

ตัวอย่างเช่น N = 1,000 คน

D = 0.05

แทนค่า $$n = 1000\frac{1000}{1+1000(0.05)^2}$$

n = 285.7

n = 286

3. ประเภทและวิธีการสุ่มตัวอย่าง ทฤษฎีการสุ่มตัวอย่าง ได้แบ่งประเภทการสุ่มตัวอย่างออกเป็น 2 ประเภทใหญ่ๆ คือ

1. การสุ่มตัวอย่างในเชิงเป็นไปได้ (Probability Sampling) การสุ่มตัวอย่างแบบนี้เราสามารถกำหนดได้ว่าทุกภาคส่วนของประชากรมีโอกาสได้รับเลือกเป็นตัวอย่างเท่ากัน การสุ่มแบบนี้มีหลายวิธีดังนี้

1.1 การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) หมายถึง การสุ่มตัวอย่างที่ประชากรทุกภาคส่วนมีโอกาสเท่าเทียมกันที่จะได้รับการคัดเลือกเป็นตัวอย่างโดยวิธีการใช้ ตารางเลขสุ่ม นำจำนวนขนาดตัวอย่างไปสุ่มในตารางสำเร็จรูปที่นักสถิติจัดทำไว้แล้ว เพียงแต่นักวิจัยกำหนดหลักที่จะใช้ว่ามีกี่หลัก และจะนับไปซ้ายขวา ขึ้นบน ลงล่างอย่างไรต้องกำหนดไว้และปฏิบัติอย่างนั้น

ตลอด สุ่มโดยการชี้ตัวเลขเริ่มต้น เมื่อชี้ตรงไหนก็บอกว่าเป็นเลขประจำตัวของประชากรหรือไม่ถ้าไม่ใช่ให้ข้ามไป ทำการคัดเลือกไปเรื่อยๆ จนได้ตามจำนวนที่ต้องการ หรือ โดยวิธีการจักฉลากโดยการเขียนหมายเลขกำกับประชากรตัวอย่าง แต่ละรายการก่อนแล้วจึงจับฉลากขึ้นมา ซึ่งวิธีการจับฉลากอาจใช้ 2 แบบคือ

ก. ไม่สุ่มประชากรที่ถูกสุ่มแล้วขึ้นมาอีก (Simple Random Sampling with out Replacement) คือหยิบแล้วเอาออกได้เลยไม่ต้องใส่กลับลงไปอีก

ข. สุ่มประชากรที่ถูกสุ่มแล้วขึ้นมาได้อีก (Sample Random Sampling with Replacement) คือ หยิบขึ้นมาแล้วก็ใส่ลงไปใหม่เพื่อให้โอกาสแก่ประชากรทุกหน่วย มีโอกาสถูกเลือกขึ้นมาเท่าเดิม

1.2 การสุ่มตัวอย่างแบบมีระบบ (Systematic Sampling) การสุ่มแบบนี้นักวิจัยจะต้องอาศัยบัญชีรายชื่อ เกี่ยวกับประชากรกลุ่มเป้าหมาย โดยเลือกตามเลขที่ที่กำหนดไว้ เช่น ประชากรจำนวน 1,000 นักวิจัยต้องการตัวอย่างจำนวน 100 นักวิจัยจะต้องคัดเลือกทุกหน่วยที่ 10 เป็นต้น

1.3 การสุ่มตัวอย่างแบบแบ่งชั้น (Stratified Random Sampling) การสุ่มตัวอย่างแบบนี้ต้องแยกประเภทของประชากรเป็นกลุ่มย่อยหรือชั้นก่อน แล้วจึงสุ่มตัวอย่างแยกกันคนและกลุ่มโดยวิธี Simple Random Sampling หรือ Systematic Sampling ก็ได้ กลุ่มย่อยที่มีลักษณะเป็น Homegeneous คือมีลักษณะเหมือนกันภายในกลุ่มเช่น การแยกประเภทของประชากรตามสถานการณ์เป็นสมาชิกของกลุ่มเกษตรกร

1.4 การสุ่มตัวอย่างแบบกลุ่ม (Cluster Sampling) คือการสุ่มตัวอย่างประชากรโดยแบ่งประชากรออกเป็นกลุ่มๆ ให้แต่ละกลุ่มมีความเป็น Heterogeneous กัน คือมีความแตกต่างกันภายในกลุ่ม เช่น การสุ่มตัวอย่างโดยการแบ่งตามเขตการปกครอง

1.5 การสุ่มตัวอย่างในทุกขั้นตอน (Multi Stage Sampling) เช่น ต้องการจะทำการวิจัยโดยการสุ่มตัวอย่างประชากร โดยทำการสุ่มจังหวัดที่เป็นตัวอย่างก่อน ต่อไปก็สุ่มอำเภอ ตำบล หมู่บ้าน และครัวเรือนที่เป็นตัวอย่างตามลำดับ

2. การสุ่มตัวอย่างในเชิงเป็นไปไม่ได้ (Non-probability Sampling) คือ การสุ่มตัวอย่างโดยไม่อาจกำหนดได้ว่าทุกส่วนของประชากรมีโอกาสได้รับการคัดเลือกโดยเท่ากัน ซึ่งทำให้ไม่สามารถจะคาดคะเนหรือคำนวณหาความผิดพลาดในการสุ่มเลือกตัวอย่างได้ การสุ่มแบบนี้มีหลายวิธีคือ

2.1 การสุ่มตัวอย่างแบบบังเอิญ (Accidental Sampling) เช่น พบใครก็สัมภาษณ์ตามความพอใจของผู้วิจัย เช่น สุ่มนักท่องเที่ยวที่จะเข้าประเทศไทยที่สนามบินดอนเมือง

2.2 การสุ่มตัวอย่างโดยวิธีการจัดสรรโควตา (Quota Sampling) การสุ่มตัวอย่างเหล่านี้ต้องแบ่งกลุ่มของประชากรแล้วจัดสรรโควตาตัวอย่างไปให้แต่ละกลุ่มตามสัดส่วนของปริมาณประชากรในกลุ่มนั้นๆ ที่มีอยู่จากนั้นก็ทำการสุ่มจากแต่ละกลุ่มตามโควตาที่จัดสรร ทั้งนี้เพื่อให้ได้ตัวแทนจากกลุ่มต่างๆ อย่างเหมาะสม เช่น ชาย 80 คน หญิง 80 คน

2.3 การสุ่มตัวอย่างแบบเจาะจง(Purposive Sampling) โดยจะเลือกศึกษาจากประชากรที่มีลักษณะตรงตามวัตถุประสงค์ที่จะศึกษาเช่น เกษตรกรที่ปลูกหม่อน บร.60 เป็นต้น

2.4 การสุ่มตัวอย่างพิจารณาตามความสะดวก (Convenience Sampling) โดยจะเลือกศึกษากลุ่มประชากรที่เห็นว่าง่ายต่อการศึกษา เช่น ไม่อยู่ในแดนของผู้ก่อการร้าย หรือเลือกเฉพาะผู้เป็นสมาชิกของกลุ่มทางการเกษตร กลุ่มใดกลุ่มหนึ่ง

2.2.2 ปัจจัยที่ทำให้สำเร็จในการสุ่มตัวอย่าง (Key Success Factor)

1. ฐานข้อมูล/ประชากรต้องเป็นปัจจุบัน (Update Population)
2. วิธีการสุ่ม ต้องมีความน่าเชื่อถือ (มีแหล่งที่มาอ้างอิงได้)
3. ขนาดตัวอย่างต้องมีการกระจายตัวและครอบคลุมประชากรเพื่อให้มีความคลาดเคลื่อนน้อยที่สุด

2.3 ประเภทของสถิติ

นักคณิตศาสตร์ได้แบ่งสถิติในฐานะที่เป็นศาสตร์ออกเป็นสาขาใหญ่ ๆ 2 สาขาด้วยกัน คือ สถิติเชิงพรรณนา (Descriptive Statistics) และการอนุมานเชิงสถิติ หรือ สถิติเชิงอนุมาน (Inferential Statistics) ซึ่งแต่ละสาขามีรายละเอียดดังนี้

1. **สถิติพรรณนา** (Descriptive Statistics) หมายถึง การบรรยายลักษณะของข้อมูล (Data) ที่ผู้วิจัยเก็บรวบรวมจากประชากรหรือกลุ่มตัวอย่างที่สนใจ ซึ่งอาจจะแสดงในรูป ค่าเฉลี่ย มัธยฐาน ฐานนิยม ร้อยละ ส่วนเบี่ยงเบนมาตรฐาน ความแปรปรวน เป็นต้น
2. **สถิติเชิงอนุมาน** (Inferential Statistics) หมายถึง สถิติที่ว่าด้วยการวิเคราะห์ข้อมูลที่รวบรวมมาจากกลุ่มตัวอย่าง เพื่ออธิบายสรุปลักษณะบางประการของประชากร โดยมีการนำทฤษฎีความน่าจะเป็นมาประยุกต์ใช้ สถิติสาขานี้ ได้แก่ การประมาณค่าทางสถิติ การทดสอบสมมุติฐานทางสถิติ การวิเคราะห์การถดถอยและสหสัมพันธ์ เป็นต้น

2.4 ข้อมูล

ข้อมูล(Data) หมายถึง ข้อมูลหรือตัวเลขที่แสดงคุณสมบัติที่ผู้วิจัยต้องการศึกษา เช่น อายุ รายได้ ยอดขาย เป็นต้น

2.4.1 ประเภทของข้อมูล

ข้อมูลที่เก็บรวบรวมมา ไม่ว่าจะใช้ในงานวิจัย หรือวัตถุประสงค์อื่นใดก็ตาม ย่อมมีลักษณะแตกต่างกันไป โดยการจำแนกข้อมูล อาจจำแนกตามเกณฑ์ใหญ่ ๆ ได้ 3 ประการ คือ จำแนกตามแหล่งข้อมูล จำแนกตามลักษณะของข้อมูล และจำแนกตามมาตรการวัด เป็นต้น

จำแนกตามแหล่งข้อมูล

1. **ข้อมูลปฐมภูมิ** (Primary Data) หมายถึง ข้อมูลที่ผู้วิจัยเป็นผู้เก็บรวบรวมข้อมูลที่สนใจเอง โดยที่อาจจะใช้วิธีเก็บแบบสอบถาม แบบสัมภาษณ์ การทดลอง เป็นต้น
2. **ข้อมูลทุติยภูมิ** (Secondary Data) หมายถึง ข้อมูลที่ผู้วิจัยไม่ได้เป็นผู้เก็บรวบรวมข้อมูลที่สนใจเอง โดยนำข้อมูลที่ผู้อื่น ๆ เก็บมาใช้ เช่น ข้อมูลเกี่ยวกับการจ้างงานที่กระทรวงแรงงานรวบรวมไว้ เป็นต้น

จำแนกตามลักษณะของข้อมูล

1. **ข้อมูลเชิงปริมาณ** (Quantitative Data) หมายถึง ข้อมูลที่สามารถแสดงในรูปตัวเลขได้ เช่น น้ำหนัก อายุ คะแนน จำนวนสินค้า งบประมาณ จำนวนพนักงานในบริษัท เป็นต้น ข้อมูลเชิงปริมาณยังแบ่งได้ ๒ ประเภท คือ
   1.1. **ข้อมูลแบบต่อเนื่อง** (Continuous Data) หมายถึงข้อมูลที่มีค่าต่าง ๆ ทุกค่าต่อเนื่องกัน โดยแสดงได้ทั้งเศษส่วนหรือตัวเลขที่เป็นจำนวนเต็ม เช่น ส่วนสูง น้ำหนัก ความยาวของโต๊ะ
   1.2. **ข้อมูลแบบไม่ต่อเนื่อง** (Discrete Data) หมายถึงข้อมูลที่มีค่าเป็นจำนวนเต็มหรือจำนวนนับ เช่น ค่าใช้จ่าย จำนวนสินค้า งบประมาณ จำนวนพนักงานในบริษัท
2. **ข้อมูลเชิงคุณภาพ** (Qualitative Data) หมายถึง ข้อมูลที่ไม่สามารถแสดงในรูปตัวเลขได้ หรือ อาจจะแสดงในรูปตัวเลขได้แต่ไม่สามารถคำนวณในเชิงปริมาณได้ เนื่องจากตัวเลขเหล่านั้นไม่สามารถอธิบายได้ เช่น เพศ สถานภาพ วุฒิการศึกษา เป็นต้น

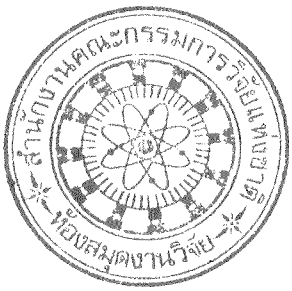
จำแนกตามมาตรการวัด

1. **นามบัญญัติ** (Nominal Scales) คือระดับของข้อมูลที่เป็นการกำหนดชื่อหรือแบ่งแยกประเภทของสิ่งต่าง ๆ เช่น เบอร์โทรศัพท์ ห้องเรียน อาคารเรียน บ้านเลขที่ เป็นต้น
2. **เรียงลำดับ** (Ordinal Scales) คือระดับของข้อมูลที่สามารถจัดลำดับความสำคัญของข้อมูลตามความแตกต่างได้ เช่น ชอบมาก ชอบปานกลาง ชอบน้อย ไม่ชอบ เป็นต้น
3. **อันตรภาค** (Interval Scales) คือระดับของข้อมูลที่สามารถบอกถึงปริมาณของความแตกต่างของข้อมูลได้ แต่ไม่มีศูนย์แท้ เช่น อุณหภูมิ คะแนนสอบวัดความรู้ ความสูง เป็นต้น
4. **อัตราส่วน** (Ratio Scales) คือระดับของข้อมูลที่เป็นการบอกถึงปริมาณความแตกต่างของข้อมูลที่มีรายละเอียดมากที่สุดและมีศูนย์แท้ เช่น ระยะทาง น้ำหนัก ความเร็ว เป็นต้น

2.5 ค่าเฉลี่ย ($\bar{x}$)

กรณีข้อมูลที่ไม่ได้แบ่งกลุ่ม $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ หรือ $\bar{x} = \frac{\sum x}{n}$

กรณีข้อมูลที่แบ่งกลุ่ม $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$ หรือ $\bar{x} = \frac{\sum fx}{n}$



2.6 ระบบบริหารความสัมพันธ์ลูกค้า (Customer Relationship Management)

ความสำคัญของระบบบริหารความสัมพันธ์ลูกค้า ช่วยให้องค์กรสามารถเพิ่มความสัมพันธ์อันดีให้กับลูกค้า เพิ่มรายได้ลดค่าใช้จ่ายต่างๆ โดยเฉพาะเรื่องค่าใช้จ่ายในการแสวงหาลูกค้า และ เพิ่มความพึงพอใจของลูกค้า (Customer Satisfaction) โดยการสร้างกระบวนการทำงานและพัฒนาผลิตภัณฑ์ตามความต้องการของลูกค้า จำนวนผู้ประกอบการ ที่เพิ่มขึ้นอย่างรวดเร็วในปัจจุบัน สามารถนำแนวทางการบริหารลูกค้าสัมพันธ์ เช่น แนวทางที่สำคัญในการจัดการระบบการบริหารงานและสร้างมาตรฐานการทำงานในบริษัท เช่น การรวบรวมการเกี่ยวกับข้อมูลของลูกค้า , การจัดการเกี่ยวกับช่องทางการสื่อสารและการพัฒนาสินค้า และบริการเพื่อสนองตอบสิ่งที่ลูกค้าต้องการ ประโยชน์ของการบริหารลูกค้าสัมพันธ์ต่อบริษัท

1. การเพิ่มรายได้จากการขาย (Sale Revenue Increase) การมุ่งเน้นการสร้างความสัมพันธ์อันดีกับลูกค้า เพื่อทำให้ลูกค้าเกิดความภักดีการใช้สินค้าหรือบริการ (Customer Loyalty) การนำหลักการบริหารลูกค้าสัมพันธ์ (CRM) มีรายได้ที่เพิ่มขึ้น ปรับปรุงกระบวนการทำงานในบริษัทลดรายจ่ายในการดำเนินงาน และต้นทุนการหาลูกค้า ใหม่ๆหรือดึงลูกค้ากลับมาใช้สินค้าหรือบริการอีกครั้ง

2. การบริหารของวงจรการทำธุรกิจของลูกค้า (Customer Life Cycle Management)

2.1 การหาลูกค้าใหม่เข้าองค์กร (Customer Acquisition) โดยการสร้างความเด่น (Differentiation) ของสินค้าหรือบริการที่ใหม่ (Innovation) และเสนอความความสะดวกสบาย (Convenience) ให้กับลูกค้า

2.2 การเพิ่มความพึงพอใจให้กับลูกค้าเพื่อทำการซื้อสินค้าและบริการ โดยผ่านขั้นตอนการทำงาน ที่กระชับเพื่อการสนองตอบความต้องการของลูกค้าได้รวดเร็วและถูกต้อง และการทำงานที่สนอง ตอบสิ่งที่ลูกค้าต้องการหรือเสนอสิ่งที่ลูกค้าต้องการ โดยผ่านหน่วยงาน ลูกค้าสัมพันธ์ (Customer Service)

2.3 การรักษาลูกค้า (Customer Retention) ให้อยู่กับองค์การนานที่สุด และการดึงลูกค้าให้กลับมา ใช้สินค้าหรือบริการ โดยฟังความคิดเห็นจากลูกค้าและพนักงานในองค์กร (Listening) รวมถึงการเสนอสินค้าและบริการใหม่ (New Product)

3. การเพิ่มประสิทธิภาพในกระบวนการตัดสินใจ (Improving Integration of Decision Making Process)

การเพิ่มการประสานงานในฝ่ายต่างๆของบริษัท โดยเฉพาะการใช้ระบบฐานข้อมูลของลูกค้าร่วมกัน และผู้บริหารสามารถดึงข้อมูลจากระบบต่างๆมาประกอบการตัดสินใจ เช่น รายละเอียดของลูกค้าที่ติดต่อเข้ามาในฝ่ายลูกค้าสัมพันธ์ (Call Center) รายละเอียดของการจ่ายเงินของลูกค้าจากฝ่ายขาย(Sales) กิจกรรมทางการตลาดที่เสนอให้ลูกค้าแต่ละกลุ่มจากฝ่ายการตลาด(Marketing) และ การควบคุมปริมาณของสินค้าในแต่ละช่วงจากฝ่ายสินค้าคงคลัง (Inventory Control) เป็นต้น

4. การเพิ่มประสิทธิภาพในการดำเนินงาน (Enhanced Operational Efficiency)

การบริหารลูกค้าสัมพันธ์ (CRM) จะช่วยเพิ่มประสิทธิภาพการทำงานของฝ่ายต่างๆ ของบริษัท โดยข้อมูลต่างนั้นได้มาจากช่องทางการสื่อสาร เช่น แฟกซ์ โทรศัพท์ และ อีเมล์ (Email) เป็นต้น

- ฝ่ายขาย: Sales, Cross-selling และ Up-selling

ระบบเทคโนโลยีสารสนเทศที่เกี่ยวข้องกับฝ่ายขาย เช่น ในการขายสินค้าแบบ Cross-selling และ Up-selling เพิ่มความสามารถในการคาดเดาแนวโน้มการซื้อสินค้าหรือบริการ รวมถึงการใช้ข้อมูลของลูกค้า เช่น ข้อสัญญา (Contract) ระหว่างองค์กรกับลูกค้า ระบบยังช่วยระบุรายละเอียดของสินค้าหรือบริการให้เหมาะสมกับลูกค้าแต่ละราย การเก็บข้อมูลทาง ด้านการขาย และการตรวจสอบสถานภาพของการส่งสินค้าให้กับลูกค้า

- ฝ่ายการตลาด (Marketing) ระบบการบริหารลูกค้าสัมพันธ์ (CRM) มีส่วนช่วยให้บริษัทสามารถวิเคราะห์ว่าวิธีใดที่ควรจัด จำหน่ายสินค้าผ่านช่องทางการขาย (Sales Channels) ต่างๆ เช่น ตัวแทนการขาย (Sales Representatives) และ ผ่านทางเว็บไซต์ (Website) ระบบการ บริหารลูกค้าสัมพันธ์ยังมีบทบาท สำคัญกับช่องทางการสื่อสาร (Communication Channels) เช่น ระบุช่องทางการสื่อสาร ที่เหมาะสมที่สุดสำหรับการขายสินค้าชนิดนั้นหรือลูกค้าแต่ละราย หรือ การระบุ พนักงานที่เหมาะสม ที่สุดในการให้บริการหรือติดต่อกับลูกค้ารายนั้นๆ

- ฝ่ายลูกค้าสัมพันธ์ (Customer Service) และฝ่ายสนับสนุน (Support)

ระบบการบริหารลูกค้าสัมพันธ์ในฝ่ายลูกค้าสัมพันธ์(Customer Service) และฝ่ายสนับสนุน (Support) ที่สำคัญคือด้านการดูแลลูกค้า (Customer Care Service) เช่น ระบบการจัดการเกี่ยวกับข้อมูล รายละเอียดของลูกค้าในองค์กร (Account management) และ ระบบแสดงรายละเอียดของสัญญาระหว่างองค์กรกับลูกค้า (Detail Service Agreement) นอกจากนี้แล้วระบบการจัดการด้าน

อีเมลล์ (Email Management System) ถือว่าเป็นส่วนสำคัญในการสร้างกลยุทธ์ทางด้านการบริหารลูกค้าสัมพันธ์ (CRM) เช่น สามารถย้อนหลังดูอีเมลล์ของลูกค้าในอดีตได้ และระบุผู้แทนฝ่ายขายที่เหมาะสมที่สุดกับลูกค้ารายนั้นได้โดยข้อมูลที่ใช้ อาจจะมาจากข้อมูลต่างๆที่ลูกค้าเคยติดต่อด้วย

- รายละเอียดของการชำระค่าสินค้าหรือบริการให้กับลูกค้า (Customer Billing)

ธุรกิจสามารถใช้ระบบการบริหารลูกค้าสัมพันธ์ (CRM) ในออกรายละเอียดการจ่ายเงินของลูกค้า (Bill Payment) และที่ผ่านการจ่ายเงินระบบอินเตอร์เนต (Electronic Bill) และการให้บริการนการตอบข้อสงสัยต่างๆผ่านช่องทางการสื่อสารต่างๆ เช่น ในระบบออนไลน์

- การขายและให้บริการในสถานที่ที่ลูกค้าต้องการ (Field Sales and Service)

การบริหารลูกค้าสัมพันธที่เกี่ยวข้องกับการขาย และให้บริการในสถานที่ที่ลูกค้าต้องการ (Field Sales and Service) ทำให้พนักงานสามารถช่วยในการดึงข้อมูลมาใช้ในขณะที่ทำการขายหรือการ ให้บริการกับลูกค้า โดยสามารถใช้ข้อมูลดังกล่าวร่วมข้อมูลขององค์การร่วมกันได้ การบริหารลูกค้าสัมพันธ์ยังมีส่วนการจัดการเกี่ยวกับการทำรายงานทางการขาย การสร้างใบเสนอ ราคาให้กับลูกค้า และเงื่อนไขพิเศษให้กับลูกค้าแต่ละรายแบบอัตโนมัติ การเสนอสินค้าที่มีความพิเศษเฉพาะตามต้องการของลูกค้าแต่ละราย (Customized Products) ระบบสินค้าคงคลัง (Inventory System),ระบบการสั่งซื้อ (Ordering System) การส่งและรับสินค้าหรือบริการ (Logistic System), การจัดตารางให้กับพนักงานที่จะให้บริการ การออกใบแจ้งหนี้และการจัดการระบบโควต้าในการขาย

- กิจกรรมที่สร้างความภักดีและการรักษาลูกค้า (Loyalty และ Retain Program)

การบริหารลูกค้าสัมพันธ์ ที่มีประสิทธิภาพขึ้นอยู่กับการแยกความแตกต่างเหล่านี้ตามกลุ่มลูกค้า(Customer Segmentation) เช่น การจำแนกประเภทของลูกค้าออกตามความต้องการของลูกค้า ประวัติส่วนตัวของลูกค้า และประวัติการซื้อ นอกจากนี้ยังสามารถกิจกรรมลูกค้าย้อนหลัง เพื่อบริษัทจะได้นำข้อมูลเหล่านี้ไปวิเคราะห์หาข้อมูลเชิงลึก เช่น ช่องทางการสื่อสารเหมาะสมที่สุดของลูกค้าแต่ละราย (Effective Communication Channel) พฤติกรรมการซื้อของลูกค้า (Customer Behavior) และสินค้าที่มีความพิเศษเฉพาะตัว (Customized Product) สำหรับลูกค้าแต่ละราย

5. เพิ่มความรวดเร็วในการให้บริการ (Speed of Service)การใช้หลักการบริหารลูกค้าสัมพันธสามารถปรับปรุงกระบวนการทำงานโดยมุ่งเน้นที่การตอบสนองความต้องการของลูกค้า จะต้องรวดเร็วและถูกต้อง โดยเฉพาะการตอบสนองแบบให้บริการ หรือตอบสนองกับลูกค้าทันที (Real Time) เช่น ระบบการสั่งสินค้ามีการเชื่อมโยงระบบต่างๆ ทั้งในฝ่ายรับการสั่งซื้อ (Order Fulfillment) ฝ่ายขาย (Sales Department) ฝ่ายบัญชี (Accounting Department) ฝ่ายสินค้าคงคลัง (Inventory) และฝ่ายที่เกี่ยวข้องกับการให้เครดิตกับลูกค้า (Credit Authorization)

6. การรวบรวมรายละเอียดต่างของลูกค้า (Gathering More Comprehensive Customer Profiles) การบริหารลูกค้าสัมพันธ์ได้ช่วยให้เพิ่มประสิทธิภาพการทำงานของฝ่ายต่างๆในบริษัท ได้

มากขึ้น เพราะว่าการบริหารลูกค้าสัมพันธ์ช่วยการจัดการเกี่ยวกับข้อมูลของลูกค้าที่มีอยู่ได้มากขึ้นทำให้ข้อมูลเก็บอย่างเป็นระบบอย่างเชื่อมโยงขึ้นบริษัทสามารถนำฐานข้อมูลนี้มาใช้ในระบบต่างๆได้

7. การลดต้นทุนในการขายและการจัดการ (Decrease General Sales and Marketing Administration Costs) การลดลงของต้นทุนการดำเนินงานนั้นมาจากใช้หลักการบริหารลูกค้าสัมพันธ์เนื่องจากบริษัทมีระบบการจัดการที่เน้นในเรื่องการสร้างความสัมพันธ์อันดีกับลูกค้าเข้าใจความต้องการของลูกค้าและ ตอบสนองความต้องการของลูกค้าได้มากขึ้น ทำให้บริษัทไม่สูญเสียต้นทุนในการดึงลูกค้ากลับเป็นลูกค้าขององค์กรอีก และตัดกระบวนการที่ไม่จำเป็น และกิจกรรมที่ไม่ก่อให้เกิดรายได้แก่บริษัท

8. การสร้างมูลค่าเพิ่ม (Value Added) ให้กับลูกค้าในปัจจุบันลูกค้านั้นพยายามแสวงหาความพึงพอใจสูงสุดจาก สินค้าและบริการ สิ่งที่ลูกค้าต้องการจึงไม่ใช่แค่คุณค่า (Value) อีกต่อไป แต่ต้องการคุณค่าเพิ่มที่ทำให้ลูกค้ามีความรู้สึกมากกว่าความพอใจ ซึ่งผู้ประกอบการควรสร้างคุณค่าเพิ่มในตัวสินค้าและบริการ โดยผ่าน Value Chain ทั้งในส่วนของคู่ค้า (Supply Chain) และในส่วนของความต้องการของลูกค้า (Demand Chain) เพื่อทำให้เกิดการบูรณาการที่ทำให้เกิดมูลค่าเพิ่มให้กับลูกค้าอย่างครบวงจรทั้งระบบ จากหลายหน่วยงานเข้ามาเกี่ยวข้องทั้งภายในองค์กร และภายนอกองค์กร (Internal and External Organization) นับตั้งแต่ผู้จำหน่ายวัตถุดิบ (Raw Materials Suppliers) กระบวนการข่าวเศรฐกิจที่เกี่ยวข้องกับวัตถุดิบ (Material Procurement) การออกแบบผลิตภัณฑ์ (Product Designers) การจัดหาอุปกรณ์ชิ้นส่วน(Spare Parts Suppliers) การขาย (Sales) และการตลาด (Marketing) ผู้ที่ทำการจัดจำหน่าย (Distributors) และ หน่วยงานลูกค้าสัมพันธ์ (Contact Center) เป็นต้น

ระบบบริหารความสัมพันธ์ลูกค้า (Customer Relationship Management :CRM) คือกลยุทธ์การบริหารจัดการอย่างหนึ่ง ซึ่งถูกออกแบบมาเพื่อช่วยองค์การให้สามารถจัดการกระบวนการต่างๆภายในให้ดำเนินการได้อย่างสอดคล้องและตอบสนองต่อความต้องการของลูกค้า เพื่อให้ลูกค้าเกิดความพอใจสูงสุด นำมาซึ่งความภักดีของลูกค้า รายได้ที่เพิ่มขึ้น และการทำกำไรในระยะยาว สำหรับบางคนที่เข้าใจว่า CRM เป็นซอฟท์แวร์ นั่นอาจเป็นเพราะส่วนหนึ่งของกลยุทธ์ CRM จำเป็นต้องอาศัยเทคโนโลยีเป็นเครื่องมือในการเก็บข้อมูลลูกค้า ซึ่งเป็นส่วนสำคัญที่จะทำให้องค์กรสามารถดำเนินงาน เพื่อสร้างความพึงพอใจให้ลูกค้าได้อย่างถูกต้อง มีประสิทธิภาพ และรวดเร็วขึ้น ซึ่งถ้าจะให้เห็นภาพชัดขึ้น ก็คงเปรียบเทียบฐานลูกค้าขององค์การเป็นเหมือนน้ำที่อยู่ในถัง ถ้ามีรูรั่วที่ก้นถังน้ำก็จะไหลออก เปรียบเทียบได้กับการที่องค์การจะต้องสูญเสียลูกค้าออกไปอยู่ตลอดเวลา และ CRM ก็คือเครื่องมือที่จะมาลดขนาดรอยรั่วขององค์กรให้เล็กลง เท่ากับองค์กรได้ลดอัตราการสูญเสียลูกค้าให้ต่ำลงนั่นเอง

2.7 งานวิจัยที่เกี่ยวข้อง

1) **อินเทลลิเวิร์ค (Intelliworks, 2004)** เอกสารงานวิจัย Challenges in Marketing and Student Relationship Management in Higher Education เป็นงานวิจัยจากบริษัท Intelliworks ว่าถึงความสำคัญขององค์กรทางธุรกิจ ที่มีการประยุกต์ใช้ระบบ CRM ในการนำพาองค์กรไปสู่ผลสำเร็จตามเป้าหมาย และกล่าวถึงการนำแนวคิดเดียวกันมาใช้กับองค์กรทางการศึกษา เพื่อจัดการความสัมพันธ์ของนักศึกษาภายในสถาบันการศึกษา เพื่อเพิ่มศักยภาพทางด้านการควบคุมอย่างมีประสิทธิภาพ โดยมีผลถึงการตอบรับต่อตลาดการศึกษาในองค์รวม นอกจากนั้นยังได้อธิบายถึงการใช้เทคโนโลยีการสื่อสารที่เหมาะสม สำหรับนำมาประกอบกับการสร้างระบบบริการการศึกษาให้กับนักศึกษา ซึ่งท้ายที่สุดได้ยกตัวอย่างการนำระบบเว็บมาใช้ เพื่อลดต้นทุนในการใช้จ่าย และเพิ่มประสิทธิภาพความรวดเร็วในการสื่อสาร อีกทั้งยังตอบรับกับการขยายตัวในอนาคตได้เป็นอย่างดี ด้วยการอิงหลักการของการใช้ผู้เรียนเป็นศูนย์กลาง ในการออกแบบและการจัดสร้าง

2) **เนจ เชียค์ (Naj Shaik 2005)** เอกสารงานวิจัย Service Center: A SRM Strategy to Promote Student Retention เป็นงานวิจัยที่กล่าวถึงการสร้างระบบศูนย์การการบริการ เพื่อสร้างกลไกในการบริการนักศึกษา ด้วยยุทธวิธีทาง SRM โดยผลของงานวิจัยทำให้ได้มาซึ่ง ระบบการบริการนักศึกษา ผ่านระบบเว็บ โดยมีการเก็บรักษาข้อมูล ทั้งในส่วนประวัตินักศึกษา เจ้าหน้าที่ หลักสูตร และข้อมูลบริหารจัดการ เพื่อให้บริการนักศึกษาแบบเบ็ดเสร็จ วัตถุประสงค์หลักของงานชิ้นนี้คือการสร้าง แอปพลิเคชั่นบริการนักศึกษา เพื่อใช้งานในการจัดการ เพื่อลดความซ้ำซ้อนของข้อมูล และลดค่าใช้จ่ายที่ไม่จำเป็นออกไป แต่ยังคงและเสริมประสิทธิภาพการทำงาน ความรวดเร็ว และความถูกต้องของคุณภาพข้อมูล

3) **อะแลน พอล (Alan Paul, 2008)** เอกสารงานวิจัยชื่อ A generic student lifecycle relationship management system เป็นเอกสารที่อธิบายถึงปัญหาต่างๆ ที่ก่อให้เกิดความยุ่งยากในการบริหาร และจัดการวงจรของการนักศึกษา ในสถาบันการศึกษาวงจร ทั้งที่มีเทคโนโลยีเพื่อก่อให้เกิดผลสำเร็จอยู่มากมาย นอกจากนั้นยังอธิบายถึงข้อมูลที่สามารถสนับสนุนการบริหารงานนักศึกษาอยู่มากมาย แต่อย่างไรนั้นภาพและมุมมองของการผสมผสานสิ่งต่างๆก็ยังคงมีปัญหา และความซับซ้อนแฝงอยู่ภายใน งานวิจัยชิ้นนี้นำปัญหาตัวอย่างที่เกิดขึ้นจากความผิดพลาดการดำเนินงาน มาจัดสร้างเป็นโมเดล ด้วยแนวคิดในเรื่องการตัดสินใจ (Decision Making) การตรวจสอบ (Monitoring) ที่มีส่วนสำคัญในการสร้างแบบจำลองให้มีประสิทธิภาพ โดยอ้างอิงโมเดลที่ใช้อยู่ในระบบธุรกิจที่ได้รับผลสำเร็จ นำมาเทียบเคียงและเปรียบเทียบ อีกทั้งยังมีข้อเสนอแนะและตัวอย่างต่างๆ

4) **ชุติมา อุตมะมุณีย์, 2552** งานวิจัยการพัฒนาตัวแบบระบบสนับสนุนการตัดสินใจอย่างชาญฉลาดแบบอัตโนมัติ สำหรับการเลือกสาขาวิชาเรียนของนักศึกษาระดับอุดมศึกษา The Development of Autonomous Intelligent Decision Support Systems for Managing The Study Plan of Students in Higher Education เป็นงานวิจัยที่อธิบายการวิเคราะห์เชิงเปรียบเทียบด้วยอัลกอริทึมสำหรับการเลือกสาขาวิชาเรียนของนักศึกษาระดับอุดมศึกษา และพัฒนาตัวแบบการตัดสินใจอย่างชาญฉลาดแบบอัตโนมัติ ด้วยการใช้เทคนิคการทำเหมืองข้อมูลซึ่งอาศัยการประยุกต์การเรียนรู้แบบเบย์ (Bayesian Learning) ร่วมกับขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) และกฎความสัมพันธ์ โดยมีวัตถุประสงค์เพื่อช่วยพัฒนาหลักสูตร และพัฒนาคุณภาพของนักศึกษา รวมถึงเพื่อผลักดันให้นักศึกษามีความกระตือรือร้นมากยิ่งขึ้น

5) **ปรีชา ยามันสะบีดีน, บุญเสริม กิจศิริกุล และ ประสงค์ ปราณีตพลกรัง ,2549** งานวิจัย เรื่อง การบริหารความสัมพันธ์กับนักศึกษาในสถาบันระดับอุดมศึกษาโดยการประยุกต์การทำเหมืองข้อมูล(Student Relationship Management in Higher Education Institutions with Application of Data Mining) งานวิจัยนี้เป็นการนำแนวความคิดของการทำเหมืองข้อมูลมาออกแบบและพัฒนาใช้สำหรับการบริหารความสัมพันธ์ (Student Relationship Management, SRM) โดยการใช้ข้อมูลทางสถิติมาใช้ในการทำเหมืองข้อมูล ซึ่งเป็นกรรมวิธีที่ช่วยให้สามารถวิเคราะห์ถึงพฤติกรรมของนักศึกษา ตลอดจนคาดคะเนถึงแนวโน้มของความเป็นไปได้ในอนาคตอันใกล้ ทั้งนี้เพื่อเป็นเหตุผลประกอบการตัดสินใจในการจัดการกับปัญหาของนักศึกษาที่มีแนวโน้มที่จะถูกคัดชื่อออก ดังนั้นการบริหารความสัมพันธ์กับนักศึกษาดังกล่าว สามารถช่วยในการรักษาสัมพันธ์ภาพที่ดีระหว่างนักศึกษาระดับอุดมศึกษา ผลลัพธ์ที่ได้คือข้อมูลทางสถิติที่เป็นประโยชน์และรูปแบบของพฤติกรรมนักศึกษา เพื่อนำไปใช้กำหนดภารกิจวัตถุประสงค์และวางแผนกลยุทธ์ของมหาวิทยาลัย

6) **รศ.ดร.กฤษณะ ไวยมัย และกัลยาภัทร อุดมสิน, 2550** งานวิจัยมีชื่อ คลังข้อมูลอัจริยะด้านการศึกษา (Intelligence Data Warehouse For Education) งานวิจัยนี้ได้จัดทำขึ้นโดยมีวัตถุประสงค์เพื่อนำข้อมูลต่างๆ ในสถานศึกษามาทำการวิเคราะห์ให้เกิดประโยชน์แก่สถานศึกษา และในภาพรวมของประเทศนับเป็นเรื่องที่สำคัญมาก งานวิจัยคลังข้อมูลอัจริยะด้านการศึกษา มีวัตถุประสงค์ที่สำคัญ 4 ประการ คือ ประการแรกเพื่อศึกษาวางรูปแบบและออกแบบคลังข้อมูลอัจฉริยะด้านการศึกษา ประการที่สองเพื่อจัดทำคลังข้อมูลของสถานศึกษา ประการที่สามเพื่อกำหนดดัชนีชี้วัดทางการศึกษาทั้ง 4 ด้าน และประการที่สี่เพื่อกำหนดรหัสมาตรฐานในการแลกเปลี่ยนข้อมูลด้านการศึกษา ซึ่งถ้าโครงการนี้สำเร็จลุล่วงตามวัตถุประสงค์ ก็จะมีส่วนช่วยให้สถานศึกษาของประเทศไทยได้มีเครื่องมือที่ใช้ในการจัดการข้อมูลต่างๆ ที่เกี่ยวข้องกับสถานศึกษา เพื่อที่จะสามารถนำข้อมูลเหล่านี้ไปใช้ใน

การวิเคราะห์เพื่อให้เกิดประโยชน์แก่สถานศึกษา มีศูนย์กลางข้อมูลด้านการศึกษาเพื่อใช้ในการแลกเปลี่ยนข้อมูล สามารถเปรียบเทียบข้อมูลระหว่างสถานศึกษาได้ ซึ่งจะก่อให้เกิดประโยชน์แก่สถานศึกษา และในภาพรวมของประเทศ

7) **Hengqing Tong, Xiaochuan Lu ,2009** งานวิจัยชื่อ Consumption Psychoanalysis and Customer Relationship Management Based on Association Rules Mining การวิเคราะห์พฤติกรรมการบริโภคของลูกค้า ถือเป็นสิ่งสำคัญอย่างมากสำหรับองค์กรที่ต้องการปรับปรุงการผลิตและพัฒนาตัวเองขึ้นไป และดัชนีความพึงพอใจของลูกค้าก็ถือเป็นส่วนสำคัญรูปแบบหนึ่งในแง่การบริโภค ความต้องการในการสร้างสมการเพื่อบ่งชี้ถึงดัชนีความพึงพอใจ ยังคงต้องมีการปรับปรุงและพัฒนาต่อๆไป วิธีการปรับปรุงอัลกอริทึมสมการโครงสร้าง คือการกำหนดค่าเริ่มต้นเพื่อประมวลผลซ้ำ และหาอัลกอริทึมในการเพิ่มอัตราการลู่เข้าสู่ความถูกต้องให้ได้มากที่สุด และเนื่องจากระบบการจัดการความสัมพันธ์ของลูกค้าเป็นปัจจัยในการระบบบริหารงานระดับใหญ่จึงมีการประยุกต์ใช้วิธีการใช้เหมืองข้อมูล ร่วมกับวิธีการหาความสัมพันธ์เป็นหากฎ (Rules) หรือรูปแบบ (Pattern) เพื่อเป็นพื้นฐานสำหรับขั้นตอนการตัดสินใจ(Decision) งานวิจัยนี้อธิบายอัลกอริทึมที่ถูกพัฒนาขึ้นเพื่อใช้ในซอฟต์แวร์ที่พร้อมสำหรับการใช้งานได้อย่างเป็นรูปธรรม

8) **Vasile Paul Bresfelean ,2007** งานวิจัยชื่อ "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment Decision Trees เป็นหลักการพื้นฐานเพื่อให้สามารถแยกแยะข้อมูล เพื่อการเรียนรู้โดยให้ผลลัพธ์ที่มีประสิทธิภาพซึ่งเป็นวิธีที่ใช้กับเหมืองข้อมูล สำหรับศึกษาข้อมูลและสร้างรูปแบบต้นไม้ร่วมกับกฎเกณฑ์เพื่อช่วยคำนวณและคาดคะเนผลลัพธ์ได้สิ่งหนึ่งที่ถูกท้าทาย คือการค้นหาความรู้ และรูปแบบของเหมืองข้อมูล เพื่อใช้พัฒนาความสามารถในการวิเคราะห์ข้อมูลจากโมเดลทั่วไป และเพื่อความเป็นเลิศในมหาวิทยาลัยที่ต้องพร้อมรองรับ ในการปรับเปลี่ยนความต้องการของหน่วยต่างๆ คุณภาพของระบบบริหารอาศัยผู้เชี่ยวชาญ และเทคโนโลยีชั้นสูง งานวิจัยนี้แสดงวิธีการใช้เครื่องมือในการวิเคราะห์ รวมกับข้อมูลที่ได้จากการสำรวจจากผู้เรียนในมหาวิทยาลัย เพื่อวัตถุประสงค์ในการคาดคะเนทางเลือกแนวทางการศึกษาในระดับต่อๆไป หลังจากสำเร็จการศึกษาออกไป