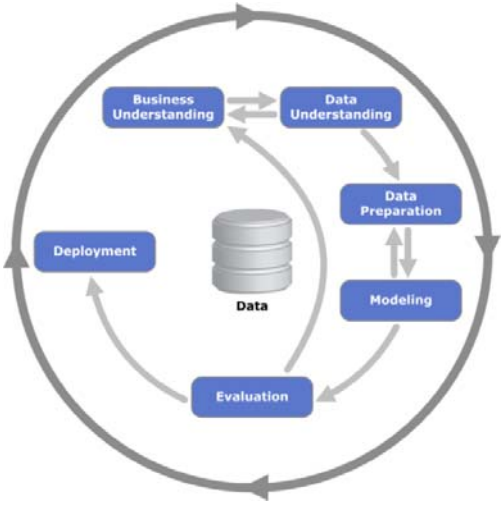
บทที่ 2
ทบทวนวรรณกรรมและเอกสารงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดเกี่ยวกับเหมืองข้อมูล

เหมืองข้อมูลหมายถึง การวิเคราะห์ข้อมูล เพื่อแยกประเภทจำแนกรูปแบบความสัมพันธ์ ของข้อมูล จากฐานข้อมูลขนาดใหญ่หรือคลังข้อมูล และนำสารสนเทศที่ไดไปใช้ในการตัดสินใจ การ ทำเหมืองข้อมูล ทำเพื่อค้นหาข้อมูลที่ถูกซ่อนอยู่ อาจค้นพบรูปแบบความสัมพันธ์ของข้อมูลที่ไม่เคยมีมาก่อน จนกลายเป็นองค์ความรู้ใหม่ (Knowledge discovery) การทำเหมืองข้อมูลคือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น (บุญเสริม กิจศิริกุล, 2552) (กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักษ์ธรรมานนท์, 2544) การจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บอย่างง่าย ๆ มาสู่การจัดเก็บในรูปแบบฐานข้อมูล ที่สามารถดึงค่าสารสนเทศของข้อมูลมาใช้จนถึงการทำเหมืองที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล สายสุนีย์ จับโจร (2547) ให้คำจำกัดความว่า การทำเหมืองข้อมูลคือกระบวนการเพื่อกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ โดยมองที่ความสัมพันธ์ของข้อมูล แนวโน้มของข้อมูลต่างๆ เพื่อให้สามารถกลั่นกรองข้อมูลและ นำไปใช้ประโยชน์ เป็นข้อมูลสนับสนุนในการตัดสินใจในเรื่องต่าง ๆ โดยความรู้ที่ได้เป็นความรู้ที่ ไม่เห็นเด่นชัด ความรู้ที่บ่งบอกเป็นนัย ส่วน Wenke Lee (2001) อธิบายว่า การทำเหมืองข้อมูลคือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น (Knowledge discovery)

ในปัจจุบันการทำเหมืองข้อมูลได้ ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเปรียบเสมือน วิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล



รูปที่ 2-1 แผนภาพ CRISP-DM

2.2 กระบวนการในการทำเหมืองข้อมูล

ในปี ค.ศ.1996 ได้มีการพัฒนาเครื่องมือมาตรฐานในการทำเหมืองข้อมูลขึ้นมา โดยความร่วมมือของ Daimler Chrysler (Daimler-Benz) SPSS(ISL) และ NCR CRISP-DM (Cross-Industry Standard Process for Data Mining) (CRISP-DM 1.0 step by step data mining guide, 1999) CRISP-DM คือ กระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล จากความร่วมมือกัน ของอุตสาหกรรมต่างๆ วัตถุประสงค์เพื่อให้การพัฒนาซอฟท์แวร์ในอุตสาหกรรมเป็นไปในทางเดียวกันที่เกี่ยวข้อง ผลลัพธ์ เทอมและการแบ่งชนิดของปัญหาการทำเหมืองข้อมูล (ชนวัฒน์ ศรีสอ้าน , 2550)

2.2.1 ขั้นตอนการทำเหมืองข้อมูล

ผู้วิจัยได้นำทฤษฎี CRISP-DM มาใช้ในการดำเนินการจัดทำการใช้วิธีทางระบบธุรกิจอัจฉริยะสำหรับการป้องกันและควบคุมโรคไข้เลือดออก มีขั้นตอนดังนี้

1) ทำความเข้าใจปัญหา (Business Understanding)

ขั้นตอนการทำความเข้าใจวัตถุประสงค์ของงานวิจัย และทำความเข้าใจเกี่ยวกับการระบาดของไข้เลือดออก ระบุปัญหา ระบุปัจจัย และสภาวะแวดล้อมที่ทำให้เกิดการระบาดของไข้เลือดออก

2) ทำความเข้าใจข้อมูล (Data Understanding)

ขั้นตอนความเข้าใจข้อมูล เป็นการรวบรวมข้อมูลโรคไข้เลือดและข้อมูลต่างๆ ที่เกี่ยวข้อง เข้าใจถึงลักษณะของข้อมูล เช่น เป็นจำนวนเต็ม ทศนิยม หรือเป็นตัวอักษร เพื่อใช้ในการวิเคราะห์ด้วยเทคนิคเหมืองข้อมูล การรวบรวมข้อมูลจะพิจารณาจากแหล่งข้อมูลที่ถูกต้องน่าเชื่อถือ มีปริมาณมากเพียงพอ และเหมาะสม มีรายละเอียดเพียงพอต่อการนำไปใช้ในการวิเคราะห์

3) เตรียมข้อมูล (Data preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานาน เนื่องจากแบบจำลองที่ได้จากการทำเหมืองข้อมูลจะให้ผลลัพธ์ที่ถูกต้องนั้น ขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ ถ้าข้อมูลที่ใช้ไม่ถูกต้อง มีความผิดพลาด จะทำให้ผลลัพธ์ที่ได้ขาดความน่าเชื่อถือ อาจทำให้ตีความผลลัพธ์คลาดเคลื่อนได้ สามารถแบ่งออกได้เป็น 3 ขั้นตอนย่อยดังนี้

3.1) การคัดเลือกข้อมูล (Data Selection) มีการกำหนดเป้าหมายว่าจะทำการวิเคราะห์สิ่งใดแล้วจึงเลือกใช้เฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่เราจะทำการวิเคราะห์

3.2) การทำความสะอาดข้อมูล (Data Cleansing) เมื่อพบข้อมูลที่ไม่ถูกต้อง ไม่สอดคล้องกับความเป็นจริง อันเนื่องมาจากปัญหาในระหว่างการจัดเก็บข้อมูล เช่น ข้อมูลไม่ครบ ข้อมูลซ้ำซ้อน ข้อมูลที่มาจากหลายแหล่งของข้อมูล ไม่ว่าจะเป็นข้อมูลที่มาจากสำนักงานระบาดวิทยา หรือสำนักงานสาธารณสุขจังหวัด สำนักงานควบคุมโรคติดต่อ

3.3) การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนการเตรียมข้อมูลให้ อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของการทำเหมืองข้อมูล

4) สร้างแบบจำลอง (Modeling)

ขั้นตอนการสร้างแบบจำลองเพื่อนำไปวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล ในบางครั้งพบว่ามีการนำเทคนิคเหมืองข้อมูล หลายเทคนิคมาใช้ในการวิเคราะห์ข้อมูลเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ดังนั้น เมื่อทำขั้นตอนนี้แล้วอาจมีการย้อนกลับไปที่ ขั้นตอนการเตรียมข้อมูล เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย นอกจากนี้ยังมีการประเมิน Model ที่ใช้วิเคราะห์ข้อมูล เพื่อเป็นตัวบ่งชี้ความน่าเชื่อถือของ Model ที่ได้

5) ประเมิน (Evaluate)

เป็นขั้นตอนการประเมินประสิทธิภาพของผลลัพธ์จากตัวแบบจำลองการวิเคราะห์ข้อมูล (Model) ว่าครอบคลุมและสามารถตอบโจทย์การพยากรณ์การระบาดของโรคไข้เลือดออก ในกรณีที่มีการสร้าง ตัวแบบจำลองการวิเคราะห์ข้อมูลหลายตัวแบบจำลอง เช่น Microsoft Time Series Algorithm, Microsoft Decision Trees, Microsoft Linear Regression และ Microsoft Neural Network จะทำการประเมินข้อดีและข้อด้อย และควรเลือกใช้ Model ใด

6) การนำไปใช้ (Deployment)

เป็นการนำผลลัพธ์หรือองค์ความรู้ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคข้อมูลไปใช้ประโยชน์ ตัวอย่างเช่น การรายงานพื้นที่เสี่ยงต่อการระบาดของไข้เลือดออกเพื่อเป็นการป้องกันและเฝ้าระวังการระบาดของโรคได้ทันทีและรวดเร็ว สามารถคำนวณงบประมาณ บุคลากรทางการแพทย์และเวชภัณฑ์ที่ใช้ในการรักษาผู้ป่วยโรคไข้เลือดออก เป็นต้น

2.2.2 เทคนิคการทำเหมืองข้อมูล

ผู้วิจัยได้เลือกเทคนิคการจำแนกข้อมูล (Classification) เป็นเทคนิคการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ ที่ได้มีการกำหนดไว้แล้วเทคนิคประเภทนี้เหมาะกับการสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูลในอนาคต ใช้ช่วยในการควบคุมการดำเนินการในปัจจุบันและในการวางแผนความต้องการในอนาคตคือ การพยากรณ์ (Forecasting) ซึ่งการพยากรณ์นั้นทำได้หลายวิธี แต่ละวิธีต่างมีเป้าหมายเดียวกันคือ ทำนายเหตุการณ์ในอนาคต ผู้วิจัยได้นำเทคนิคที่ใช้ค้นพบองค์ความรู้หรือสารสนเทศใหม่ ด้วยการเชื่อมโยงข้อมูลหรือกลุ่มข้อมูลที่เกิดขึ้นในเหตุการณ์เดียวกัน โดยใช้รูปแบบอนุกรมเวลา (Time series) ในการพยากรณ์ ข้อมูลอนุกรมเวลา (Time Series data) คือ ชุดของข้อมูลที่เก็บรวบรวมตามระยะเวลาเป็นช่วงๆ อย่างต่อเนื่องกัน เช่น ข้อมูลจำนวนผู้ป่วยโรคไข้เลือดออก ข้อมูลจำนวนผู้เสียชีวิตจากไข้เลือด เป็นต้น ข้อมูลอนุกรมเวลาอยู่ในลักษณะที่เป็นข้อมูลรายปี รายฤดู รายเดือน ทั้งนี้ขึ้นอยู่กับความเหมาะสมในการนำไปใช้ประโยชน์ เทคนิคการพยากรณ์โดยอาศัยข้อมูลอนุกรมเวลามีอยู่ด้วยกัน

1) ส่วนประกอบและรูปแบบของข้อมูลอนุกรมเวลา

ข้อมูลอนุกรมเวลาที่เก็บรวบรวมได้ในช่วงเวลาที่เท่ากัน จะมีค่าเปลี่ยนแปลงซึ่งจะเป็นตัวแปรตาม ส่วนระยะเวลาที่เปลี่ยนแปลงจะเป็นตัวแปรอิสระ สาเหตุที่ทำให้ข้อมูลอนุกรมเวลาเปลี่ยนแปลงจะเรียกว่า ซึ่งจะ ประกอบด้วยสาเหตุ 4 ชนิด ดังนี้

1.1) แนวโน้ม (Trend component: T) หมายถึง การเคลื่อนไหวของอนุกรมเวลาในระยะยาวว่าน่าจะมีแนวโน้มเพิ่มขึ้นหรือลดลง และลักษณะแนวโน้มนั้นอาจจะมีแนวโน้มเป็นเส้นตรงหรือเส้นโค้งก็ได้ ระยะเวลาที่ทำให้เห็นแนวโน้มส่วนใหญ่ไม่ควรต่ำกว่า 10 ช่วงเวลา ลักษณะเด่นของเส้นแนวโน้มคือจะต้องเรียบไม่มีการหักมุม ณ ที่ใดๆ

1.2) ฤดูกาล (Seasonal component: S) หมายถึง การเปลี่ยนแปลงของข้อมูลที่เกิดขึ้น เนื่องจากอิทธิพลของฤดูกาล ซึ่งจะเกิดขึ้นซ้ำๆ กันในช่วงเวลาเดียวกันของแต่ละปีโดยทั่วไป ช่วงเวลาของฤดูกาลหนึ่งๆ เช่น รายฤดู รายเดือน รายปี คำว่าฤดูกาลในที่นี้หมายถึง ช่วงเวลาที่

แตกต่างกันของแต่ละปี เช่น ไข้เลือดออกจะเกิดการระบาดมากในช่วงฤดูฝน และจำนวนลูกน้ำและยุงลายจะมีปริมาณมากในช่วงฤดูฝน

1.3) วัฏจักร (Cyclical component: C) หมายถึง การเคลื่อนไหวของข้อมูลที่มีลักษณะซ้ำๆ กัน คล้ายกับความผันแปรตามฤดูกาลต่างกันที่ระยะเวลาของการเคลื่อนไหวของข้อมูลจะมีระยะเวลานานกว่าหนึ่งปี เช่น 5 ปี หรือ 10 ปี โดยทั่วไปจะใช้ข้อมูลตั้งแต่ 5 ปีขึ้นไป

1.4) ผิดปกติ (Irregular component: I) หมายถึง การเคลื่อนไหวของข้อมูลที่ไม่มีรูปแบบที่แน่นอน ลักษณะของข้อมูลที่เกิดขึ้น ส่วนใหญ่จะเป็นลักษณะของเหตุการณ์ที่เราไม่สามารถทราบล่วงหน้าได้ เช่น ในปีที่เก็บข้อมูลปริมาณน้ำฝนจำนวนน้อย มีอุณหภูมิสูง และมีความชื้นสัมพัทธ์ต่ำ ทำให้มีผลต่อการวางไข่และการเจริญเติบโตของยุงลาย

2) เทคนิคที่ใช้ในการพยากรณ์โรคไข้เลือดออก

ในประเทศไทยมีนักวิจัยที่ได้ทำการวิจัยเกี่ยวกับโมเดลการพยากรณ์การระบาดของโรคไข้เลือดออกดังจะเห็นได้จากผลงานวิจัยของ นภา รัชตะ และคณะ (Rachata et al., 2008) ได้ทำการวิจัยระบบพยากรณ์ของโรคไข้เลือดออกโดยอัตโนมัติและความเสี่ยงการระบาดของโรคไข้เลือดออกโดยใช้โครงข่ายประสาทเทียม (Neural Network) พื้นที่ในการวิจัยจังหวัดเชียงราย โดยใช้ข้อมูลจากสำนักงานสถิติแห่งชาติ, จำนวนผู้ป่วยไข้เลือดออก, ปริมาณน้ำสูงสุด, ปริมาณน้ำต่ำสุด, ปริมาณฝน, ความชื้นสัมพัทธ์ โดยเก็บข้อมูลตั้งแต่ปี 1999-2007 ซึ่งเก็บข้อมูลจำนวนผู้ป่วยและลักษณะอากาศเป็นรายสัปดาห์ โดยผลการวิจัยสรุปได้ว่าโมเดลการพยากรณ์มีความถูกต้องแม่นยำถึง 85.92%

สำหรับงานวิจัยในต่างประเทศมีนักวิจัยหลายท่านได้ทำการพัฒนาโมเดลที่ใช้ในการพยากรณ์การระบาดของโรคไข้เลือดออก ตัวอย่างเช่น แอนนา แอล บัคซาส และคณะ (Buczak et al., 2008) ได้ทำการวิจัยวิธีการทำนายการระบาดของโรคไข้เลือดออกในประเทศเปรู โดยใช้ทฤษฎี Fuzzy logic ซึ่งงานวิจัยนี้ได้ผลสรุปคือได้โมเดลการพยากรณ์ที่มีประสิทธิภาพสูง ในประเทศมาเลเซีย นักวิจัยชื่อ อิบราฮิม และคณะ (Ibrahim et al., 2010) ได้ทำการวิจัยความเสี่ยงในผู้ป่วยโรคไข้เลือดออกโดยใช้การวิเคราะห์เครือข่ายประสาทเทียม โดยใช้ข้อมูลจากห้องปฏิบัติการทางการแพทย์ของผู้ป่วยที่ติดเชื้อ Den1, Den2, Den3, Den4, อุณหภูมิของร่างกาย และปริมาณน้ำในร่างกาย จากการวิจัยได้ผลสรุปคือมีความถูกต้องในทำนาย 96.86% จากผลการวิจัยนี้แสดงให้เห็นว่าเครือข่ายประสาทเป็นเทคนิคที่มีประสิทธิภาพในการวัดแนวโน้มและคาดการณ์ความเสี่ยงผู้ป่วยโรคไข้เลือดออก ในปีถัดมานักวิจัยชาวฝรั่งเศสชื่อ มายเรียม การ์บิ และคณะ (Gharbi et al., 2011) ได้ทำการวิจัยการวิเคราะห์อนุกรมเวลา (Time series analysis) ของอัตราการเกิดโรคไข้เลือดออกใน

ประเทศฝรั่งเศส ข้อมูลที่ใช้ในการวิจัยคือข้อมูลผู้ป่วยโรคไข้เลือดออก, อุณหภูมิ, ปริมาณฝน และลักษณะภูมิประเทศ ผลการวิจัยสรุปว่ารูปแบบการใช้ตัวแปรสภาพภูมิอากาศเป็นตัวพยากรณ์ในประเทศฝรั่งเศสที่น่าเชื่อถือได้มากที่สุดคือใช้ข้อมูลย้อนหลัง 3 เดือนในการพยากรณ์ของเดือนถัดไป ในปีเดียวกัน อารูล อาร์เนส (Earnest et al., 2011) ได้ทำการวิจัยเปรียบเทียบโมเดลทางสถิติเพื่อการพยากรณ์โรคไข้เลือดออกและการแจ้งเตือนการระบาดของโรคในประเทศสิงคโปร์ โดยใช้ข้อมูลจำนวนผู้ป่วยโรคไข้เลือดออก ระหว่างปี 2001-2008 ซึ่งผลการวิจัยนี้สรุปได้ว่าแบบจำลอง K-H model พยากรณ์และแจ้งเตือนได้รวดเร็วกว่าแบบจำลอง ARIMA time series ในเวลาต่อมา ซาลีม ฟาทิมา และคณะ (Fathima et al., 2011) ได้ทำการวิจัยเปรียบเทียบการระบาดของโรคไข้เลือดออกด้วยเทคนิค SVM และ Naives Bayes ในเมือง Chennai และ Tirunelveli ประเทศอินเดีย โดยใช้ข้อมูลของผู้ป่วยไข้เลือดออกจำนวน 5,000 คน ผลการวิจัยสรุปว่า SVM มีความสามารถในการทำนายการระบาดที่ถูกต้องมากกว่า Naives Bayes ในปี 2012 ที่ผ่านมานี้ นักวิจัยชื่อ นายแพทย์นาซมูล คาร์ริม และคณะ (Karim et al., 2012) ได้ทำการวิจัยปัจจัยที่มีอิทธิพลต่อการระบาดของไข้เลือดออกในประเทศบังคลาเทศ โดยใช้เทคนิคคือ Multiple regression analysis ผลสรุปของการวิจัยคือ การระบาดของโรคไข้เลือดออกตามอัตราแรงลมลมและขึ้นอยู่กับความหนาแน่นและการขยายตัวของระบบนิเวศซึ่งปัจจัยที่มีอิทธิพลต่อสภาพภูมิอากาศและความชุกของลูกน้ำยุงลายและโรคไข้เลือดออก

จากการทบทวนวรรณกรรมเพื่อศึกษาเทคนิคการพยากรณ์ต่างๆ สรุปได้ว่ามีเทคนิคต่างๆ ดังนี้คือ

ตารางที่ 2-1 การสรุปเทคนิคต่างๆ ที่นำมาใช้ในการการพยากรณ์โรคไข้เลือดออก

เทคนิค	ข้อดี	ข้อเสีย
Neural Network	1. มีความยืดหยุ่นสูงจนสามารถจำลองปัญหาต่างๆ ได้ 2. มีความสามารถในการจำชุดของคู่อินพุต - เอาต์พุตที่มีความซับซ้อนมากจนไม่สามารถจำลองแบบในเชิงความน่าจะเป็นได้ 3. มีความสามารถในการ	1. ใช้เวลา training นาน (หาค่าถ่วงน้ำหนักที่ดีที่สุดในการเรียนรู้) 2. ยากที่จะทำความเข้าใจเกี่ยวกับการเลือกพารามิเตอร์ที่เหมาะสม (best topology) 3. ยากที่จะทำความเข้าใจเกี่ยวกับ learned function

เทคนิค	ข้อดี	ข้อเสีย
	ปรับตัวเข้ากับการเปลี่ยนแปลงของสิ่งแวดล้อม 4. มีความสามารถในการตอบสนองต่อข้อมูลที่ไม่เคยเห็น 5. ความรู้จะกระจายอยู่ทั่วทั้งโครงข่ายประสาทเทียม ทำให้ดำเนินการกับสารสนเทศเชิงอรรถาธิบาย(Contextual Information) ได้อย่างเป็นธรรมชาติ	

ตารางที่ 2-1(ต่อ)

เทคนิค	ข้อดี	ข้อเสีย
Fuzzy logic	สามารถหาค่าจากข้อมูลที่มีความไม่แน่นอนได้ เช่น สูง, มาก, อร่อย,บริการดี และสามารถสร้างกฎเกณฑ์ในการทำงานได้	ไม่เหมาะกับกฎเกณฑ์จำนวนมาก
ARIMA time series	ค่าการพยากรณ์มีความถูกต้องสูงหากใช้ระยะเวลาสั้นในการพยากรณ์ เช่น พยากรณ์ล่วงหน้า 3 เดือน โดยใช้ข้อมูลย้อนหลังในการพยากรณ์ 3 เดือน	การพยากรณ์คลาดเคลื่อนสูงหากหากพยากรณ์ระยะเวลานาน
SVM (Support Vector Machine)	1. สามารถคัดแยกข้อมูลในลักษณะ Non-linear ได้ดี 2. รองรับจำนวนคุณลักษณะได้มาก และมีความถูกต้องสูง	ต้องเลือก Kernel Function ที่เหมาะสม
Multiple regression analysis	สามารถหาความสัมพันธ์เชิงเส้นระหว่างตัวแปรมาใช้ในการทำนายได้ โดยเมื่อทราบค่าตัวแปรหนึ่งก็สามารถทำนายอีกตัวแปรหนึ่งได้หรือหาความสัมพันธ์ระหว่างตัวแปรตาม (Y) หรือตัวแปรเกณฑ์ (Criterion Variable) จำนวน 1 ตัว กับตัว	ไม่สามารถหาความสัมพันธ์เชิงเส้นได้หากไม่ทราบค่าตัวแปร

เทคนิค	ข้อดี	ข้อเสีย
	แปรอิสระ (X) หรือตัวแปร พยากรณ์ หรือตัวแปรทำนาย (Predictor Variable) ตั้งแต่ 2 ตัวขึ้นไป	

ในงานวิจัยนี้จะลองศึกษาเปรียบเทียบเทคนิคในการพยากรณ์ต่างๆ เหล่านี้และเปรียบเทียบผลลัพธ์ที่ได้จากการทดลองเพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของแต่ละวิธี และเลือกวิธีที่ให้ผลการพยากรณ์ที่มีค่าใกล้เคียงกับข้อมูลจริงมากที่สุด

2.2.3 ประโยชน์ของเหมืองข้อมูล

1) ช่วยชี้แนวทางการตัดสินใจและช่วยคาดการณ์ผลลัพธ์ที่จะได้รับการจาการตัดสิน

2) เพิ่มความเร็วในการวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่

3) ค้นหาส่วนประกอบที่ซ่อนอยู่ภายในเอกสาร รวมถึงความสัมพันธ์ระหว่างส่วนประกอบต่างๆ ด้วย

4) เชื่อมโยงเอกสารที่มีความเกี่ยวข้องกันระหว่างหน่วยงานต่างๆ ภายในองค์กร เช่น การค้นพบว่าจำนวนผู้ป่วยโรคไข้เลือดออกไม่เท่ากันของสำนักงานสาธารณสุขจังหวัด สำนักงานสาธารณสุขอำเภอ สำนักงานควบคุมป้องกันโรค และสำนักงานระบาดวิทยา

5) การจัดกลุ่มเอกสารตามหัวข้อต่างๆ เช่น จัดกลุ่มลูกค้าทั้งหมดของบริษัทประกันภัยที่ประสบเหตุลักษณะเดียวกันเพื่อนดำเนินการต่างๆ ตามนโยบายของบริษัท

2.2.4 อัลกอริทึมในการทำเหมืองข้อมูลโดยใช้ SQL Server 2008

1) Microsoft Naive Bayes

อัลกอริทึมที่ใช้ค้นหาความน่าจะเป็นจากความสัมพันธ์ของคอลัมน์ข้อมูลนำเข้ากันคอลัมน์ ทำนายหรือพยากรณ์ข้อมูล โดยที่แต่คอลัมน์ข้อมูลนำเข้าเป็นอิสระต่อกัน สามารถสร้างแบบจำลอง ได้อย่างรวดเร็ว การนำไปใช้งานใน SQL Server 2008

- Analyze Key Influencers (Table Analysis Tools for Excel)

2) Microsoft Decision Tree Algorithm

อัลกอริทึมใช้สร้างทางเลือกการตัดสินใจแบบต้นไม้ โดยที่แต่ละโหนดหรือในจะแสดงคุณลักษณะ (attribute) แต่ละกิ่งแสดงผลการพิจารณาเงื่อนไขหรือทดสอบข้อมูล และลีฟโหนด (leaf node) คือกลุ่มของผลลัพธ์ที่กำหนดไว้ การนำไปใช้งานใน SQL Server 2008

- Classify (SQL Server 2008 Data Mining Add-ins)
- Estimate (SQL Server 2008 Data Mining Add-ins)

3) Microsoft Time Series Algorithm

อัลกอริทึม วิเคราะห์ความสัมพันธ์ของข้อมูลกับช่วงเวลา โดยใช้การตัดสินใจแบบต้นไม้แบบเชิงเส้น เพื่อพยากรณ์ข้อมูลในอนาคตในรูปแบบอนุกรมเวลา (Time Series) การนำไปใช้งาน ใน SQL Server 2008

- Forecast (Table Analysis Tools for Excel)
- Forecast (SQL Server 2008 Data Mining Add-ins)

4) Microsoft Cluster

อัลกอริทึมที่ใช้จัดกลุ่มข้อมูลเป็นกลุ่มย่อยจากคุณลักษณะของข้อมูลที่มีความคล้ายคลึงกัน การนำไปใช้งานใน SQL Server 2008

- Detect Categories (Table Analysis Tools for Excel)
- Cluster (SQL Server 2008 Data Mining Add-ins)

5) Microsoft Sequence Clustering

อัลกอริทึมที่ระบุกลุ่มของลำดับเหตุการณ์ เป็นการผสมผสานการใช้ การวิเคราะห์ลำดับ (Sequence Analysis) และอัลกอริทึมการจิตกลุ่ม (Clustering) การนำไปใช้งานใน SQL Server 2008

- Microsoft SQL Server Analysis Services
- Data Mining Add-ins for Office 2007

6) Microsoft Association Rule

อัลกอริทึมที่ใช้ในการสร้างกฎความสัมพันธ์ อธิบายรูปแบบความส้มพันธ์ที่เกิดขึ้นจากข้อมูลเวชภัณฑ์ที่ใช้พร้อมกับที่ใช้รักษาผู้ป่วยหนึ่งราย เช่น ให้ยาพาราต้องให้น้ำเกลือด้วย การนำไปใช้งานใน SQL. Server 2008

- Associate (SQL Server 2008 Data Mining Add-ins)
- Shopping Basket Analysis (Table Analysis Tools for Excel)

7) Poisson Regression

คืออัลกอริทึมที่ใช้ในการหาความเป็นไปได้ที่จะเกิดเหตุการณ์ใดๆ ขึ้นโดยที่ไม่รู้ความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม โดยตัวแปรตามจะเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรต้นจะเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพก็ได้ทำให้เหมาะสมกับการวิเคราะห์งานที่มีความหลากหลายของตัวแปรต้น ผู้ใช้งานสามารถใช้ความน่าจะเป็นที่ได้จาก Microsoft Neural Network algorithm มาใช้ในการจัดกลุ่ม โดยจะอยู่ในมุมมองของตารางของการจับคู่ตัวแปรต้นและตัวแปรตามแบบ โดยจะแสดงผลออกมาในรูปของความน่า 1 ต่อ 1 จะเป็น โดยผู้ใช้งานสามารถเลือกค่า ของตัวแปรต้นเพื่อดูความน่าจะเป็นที่จะเกิดตัวแปรตามแบบใดๆ ได้เช่น การจบคู่ระหว่างเวลาที่เกิดการระบาทของโรคไข้เลือดออกกับจำนวนประชากร เพศกับการระบาด ยุงตัวเมียกับจำนวนผู้ป่วย เป็นต้นหรือทำนายตัวแปรตามจากข้อมูลตัวแปรต้นที่ใส่เข้าไปได้

8) Microsoft Neural Network ประกอบด้วย

8.1) Input Layer : ชั้นของการระบุตัวแปรต้นที่จะถูกนำไปสร้างเป็นต้นแบบของการทำเหมือง ข้อมูล

8.2) Hidden layer: ชั้นของการประมวลผลค่านำหนักของตัวแปรต้นแต่ละตัวเพื่อนำไปสู่ความ น่าจะเป็นที่จะเกิดตัวแปรตามขึ้น

8.3) Output layer: ชั้นของการทำนายผลโดยจะแสดงผลออกมาเป็นความน่าจะเป็นที่จะ เกิดขึ้นของตัวแปรตาม