บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ ผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องนำมาเสนอดังนี้

2.1 นโยบายและทิศทางการพัฒนาระบบบริการสุขภาพของกระทรวงสาธารณสุข

2.2 ทฤษฎีการทำเหมืองข้อมูล

2.3 โปรแกรม Weka

2.4 งานวิจัยที่เกี่ยวข้อง

2.1 นโยบายและทิศทางการพัฒนาระบบบริการสุขภาพของกระทรวงสาธารณสุข

การพัฒนาประเทศไทยในระยะแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ ฉบับที่ 11 (พ.ศ.2555-2559) ยังคงยึดหลักปรัชญาของเศรษฐกิจพอเพียงที่มุ่งเน้นให้คนเป็นศูนย์กลางของการพัฒนา และสร้างสมดุลการพัฒนาในทุกมิติภายใต้วิสัยทัศน์ประเทศไทย ปี พ.ศ. 2570 “สังคมอยู่ร่วมกันอย่างมีความสุขด้วยความเสมอภาคเป็นธรรมและมีภูมิคุ้มกันต่อการเปลี่ยนแปลง” และทิศทางของแผนพัฒนาสุขภาพแห่งชาติ ฉบับที่ 11 (พ.ศ.2555-2559) มุ่งพัฒนาภายใต้หลักปรัชญาเศรษฐกิจพอเพียง เน้นการสร้างเอกภาพและธรรมาภิบาลในการอภิบาลระบบสุขภาพให้ความสำคัญกับการสร้างกระบวนการการมีส่วนร่วมของทุกภาคส่วนในสังคม มุ่งเน้นการสร้างหลักประกันและการจัดบริการที่ครอบคลุมเป็นธรรม โดยมีวิสัยทัศน์ “ประชาชนทุกคนมีสุขภาพดี ร่วมสร้างระบบสุขภาพพอเพียงเป็นธรรม นำสู่สังคมสุขภาวะ” ซึ่งมีความสอดคล้องกับรัฐธรรมนูญแห่งราชอาณาจักรไทย พ.ศ.2550 ได้บัญญัติไว้ว่า “ส่งเสริมสนับสนุนและพัฒนาระบบสุขภาพที่เน้นการสร้างเสริมสุขภาพ อันนำไปสู่สุขภาวะที่ยั่งยืนของประชาชน”

กระทรวงสาธารณสุขมีอำนาจหน้าที่เกี่ยวกับการสร้างเสริมสุขภาพอนามัย การป้องกันโรคและการรักษาโรค การฟื้นฟูสมรรถภาพของประชาชน เพื่อให้ประชาชนได้รับบริการอย่างครอบคลุม ทั่วถึงเป็นธรรม และสร้างระบบบริการสุขภาพให้มีมาตรฐานในทุกระดับเพื่อตอบสนองต่อปัญหาในกลุ่มเป้าหมายและพัฒนาระบบส่งต่อที่ไร้รอยต่อ ตลอดจนสร้างกลไกกลางระดับชาติในการดูแลระบบบริการสุขภาพและพัฒนาระบบบริหารจัดการทรัพยากรให้มีประสิทธิภาพ ลดค่าใช้จ่ายด้านสุขภาพของประชาชน เพื่อเป้าหมายประชาชนมีคุณภาพชีวิตที่ดี จึงได้นำแนวคิดแบบ

ใหม่ (New management) มาใช้ในการบริหารจัดการเพื่อแก้ไขปัญหาสุขภาพและเป็นไปตามทิศทางของแผนพัฒนาสุขภาพแห่งชาติฉบับที่ 11 โดยการจัดทำแผนแม่บทในการพัฒนาระบบบริการสุขภาพทุกระดับเป็นแผน 5 ปี (Service Plan 2555-2559) ภายใต้ปรัชญาการจัดบริการทั่วถึงและเป็นธรรมไม่ว่าจะอยู่ ณ ภูมิภาคไหนของประเทศไทย โดยมีกลยุทธ์หลักที่สำคัญไว้ 3 ประเด็น (Key Strategic areas) ดังนี้ 1) การพัฒนาโรงพยาบาลระดับต่างๆให้เป็นส่วนหนึ่งของระบบเครือข่ายบริการสุขภาพ (Health service network system) เพื่อให้โรงพยาบาลแต่ละแห่งมีการเติบโตอย่างมีทิศทาง มีภารกิจหน้าที่ชัดเจน และมีการเกื้อหนุนซึ่งกันและกันภายในเครือข่ายบริการ 2) การพัฒนาเป็นศูนย์ความเชี่ยวชาญระดับสูง 4 สาขาที่สำคัญคือ โรคหัวใจ โรคมะเร็ง อุบัติเหตุและทารกแรกเกิด 3) การพัฒนาระบบบริการปฐมภูมิ ซึ่งเป็นด่านหน้าใกล้ชิดประชาชนมากที่สุด การจัดบริการในระดับปฐมภูมิที่สำคัญคืองานบริการเวชปฏิบัติครอบครัวและงานสุขภาพชุมชน ดูแลในมิติของสุขภาพองค์รวม การสร้างความเข้มแข็งให้ครอบครัวและชุมชน การสร้างเสริมสุขภาพเชิงรุกโดยให้ชุมชนและประชาชนมีบทบาทสำคัญในการจัดการสุขภาพตนเองในระดับต้น ดังนั้นการคัดกรองสุขภาพประชาชนจึงเป็นเครื่องมือที่สำคัญที่สามารถป้องกันโรค หยุดยั้งการลุกลามเข้าสู่ระยะรุนแรงและแก้ไขได้ในเบื้องต้น รวมทั้งนำข้อมูลที่ได้มาใช้ในการสะท้อนพฤติกรรมสุขภาพและสภาวะสุขภาพ สถานการณ์ใหม่ๆที่เป็นปัจจุบัน เพื่อวางแผนในการกำหนดทิศทางนโยบายและการแก้ไขที่ถูกต้องต่อไป

สำนักงานหลักประกันสุขภาพแห่งชาติจึงได้ส่งเสริมสนับสนุนงบประมาณในการเฝ้าระวังโรคเรื้อรังทุกจังหวัด โดยการใช้แบบสำรวจสภาวะสุขภาพของประชาชนอายุตั้งแต่ 15 ปีขึ้นไป ซึ่งใช้ในการคัดกรองสุขภาพเบื้องต้นเป็นประจำทุกปี ทั้งในเขตเทศบาลและนอกเขตเทศบาล ศูนย์แพทย์ชุมชนโรงพยาบาลบุรีรัมย์ ได้เล็งเห็นความสำคัญของนโยบายดังกล่าว จึงได้จัดทำการสำรวจสภาวะสุขภาพของประชาชนอายุตั้งแต่ 15 ปีในเขตเทศบาล ซึ่งข้อมูลประกอบด้วย 1) ข้อมูลทั่วไป 2) ข้อมูลครอบครัว 3) ข้อมูลการเจ็บป่วยและพฤติกรรมเมื่อเจ็บป่วย 4) ข้อมูลคะแนนพฤติกรรมด้านสุขภาพ เพื่อนำข้อมูลที่ได้มาออกแบบกิจกรรมและงบประมาณให้สอดคล้องกับผลที่สำรวจได้ทุกปี ดังเช่น ผลการคัดกรองสุขภาพเบื้องต้นในชุมชนหนึ่งพบว่ามีคนที่อยู่ในกลุ่มเสี่ยงเป็นจำนวนมาก และเมื่อนำมาวิเคราะห์ทำให้ทราบว่าส่วนใหญ่เป็นคนอ้วนลงพุง จึงจัดให้มีการอบรมเพื่อปรับเปลี่ยนพฤติกรรมในกลุ่มเสี่ยงของชุมชนนั้นด้วยโครงการชุมชนลดหวาน มัน เค็ม ลดอ้วน ลดโรคใน ปี 2555 เป็นต้น

2.2 ทฤษฎีการทำเหมืองข้อมูล (Data Mining)

Data Mining คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น (Han, J., et al., 2006) ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม

การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูลจากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

2.2.1 วิวัฒนาการของ Data Mining

1. ปี ค.ศ 1960 Data Collection มีการนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือ เพื่อป้องกันการสูญหายได้เป็นอย่างดี

2. ปี 1980 Data Access มีการนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ระหว่างกัน เพื่อนำไปวิเคราะห์ และตัดสินใจอย่างมีประสิทธิภาพ

3. ปี 1990 Data Warehouse and Dicision Support มีการนำข้อมูลมาเก็บลงในฐานข้อมูลขนาดใหญ่ ครอบคลุมการใช้งานทั้งหมดขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ

4. ปี 2000 Data Mining นำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล โดยสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

2.2.2 วัตถุประสงค์ในการใช้ Data Mining

1. เพื่อการค้นพบองค์ความรู้ใหม่ในฐานข้อมูล (Knowledge Discovery in Databases)
2. เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่ (Knowledge Extraction)
3. เพื่อจัดการกับข้อมูลในอดีต (Data Archeology)
4. เพื่อสำรวจข้อมูล (Data Exploration)
5. เพื่อค้นหา Pattern ของข้อมูลที่ซ่อนอยู่ (Data Pattern Processing)
6. เพื่อใช้ขุดเจาะข้อมูล (Data Dredging)
7. เพื่อให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

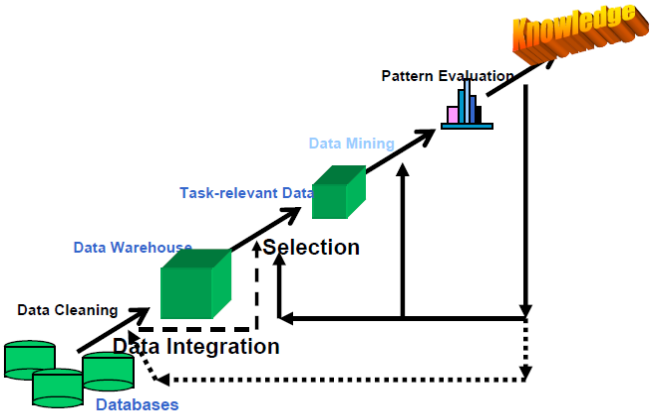
2.2.3 เป้าหมายหลักของ Data Mining

คุณลักษณะและเป้าหมายหลักของ Data Mining คือ ใช้ค้นหา Pattern ของข้อมูลที่ฝังลึก และซ่อนเร้นอยู่ภายในฐานข้อมูลขนาดใหญ่ โดยใช้สถาปัตยกรรม Client-Server ใช้เครื่องมือสมัยใหม่ที่สามารถแสดงผลแบบกราฟฟิก ผู้ใช้สามารถดูข้อมูลแบบเจาะลึก และสามารถใช้เครื่องมือในการสอบถามข้อมูลได้อย่างง่ายดาย โดยไม่ต้องอาศัยความชำนาญของ

ผู้พัฒนาโปรแกรม เพราะเครื่องมือถูกออกแบบมาให้ใช้งานได้ง่าย บ่อยครั้งอาจค้นพบผลลัพธ์ที่ไม่คาดหวังมาก่อน

2.2.4 กระบวนการของการทำเหมืองข้อมูล

เป็นกระบวนการในการค้นหาลักษณะแฝงของข้อมูล (Pattern) ที่ซ่อนอยู่ในฐานข้อมูล โดยมีขั้นตอนดังภาพที่ 2.1

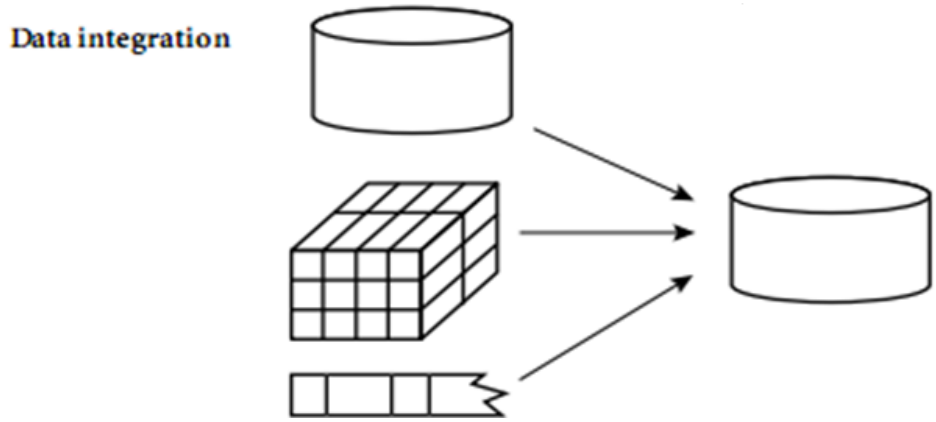


ภาพที่ 2.1 กระบวนการของ Data Mining

ขั้นตอนของการทำเหมืองข้อมูล มี 5 ขั้นตอนดังนี้

1. Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป โดยทั่วไปข้อมูลที่ถูกจัดเก็บอาจมีความผิดปกติต่างๆได้ เช่น ข้อมูลบางแอทริบิวต์ขาดหายไป (missing value) ขาดแอทริบิวต์ที่น่าสนใจ หรือขาดรายละเอียดของข้อมูล เป็นข้อมูลรบกวน (noisy data) เช่น ข้อมูลมีค่าผิดพลาด (error) หรือมีค่าผิดปกติ (Outliers)

2. Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน เช่น มีข้อมูลในคลังข้อมูล (Data Warehouse) ในรูปแบบของดาต้าคิวบ์ (Data Cube) และมีข้อมูลในรูปแบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database) จำเป็นต้องทำการรวมข้อมูลให้เป็นข้อมูลชุดเดียวกันดังภาพที่ 2.2



ภาพที่ 2.2 กระบวนการของ Data Integration

3. Data Selection เป็นขั้นตอนระบุถึงแหล่งข้อมูลที่จะนำมาทำ mining รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น

4. Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่ โดยใช้เทคนิคและอัลกอริทึมที่เหมาะสมกับข้อมูลที่มีอยู่

5. Pattern Evaluation เป็นขั้นตอนประเมินผล Pattern และนำเสนอองค์ความรู้ ในขั้นตอนนี้จะเป็นการวิเคราะห์ผลลัพธ์ที่ได้ แปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่และนำเสนอ

2.2.5 ประเภทของข้อมูลที่ใช้ในการทำเหมืองข้อมูล

1. ข้อมูลที่มาจากฐานข้อมูลเชิงสัมพันธ์ (Relational databases)
2. ข้อมูลจากคลังข้อมูล (Data warehouses)
3. ข้อมูลจากฐานข้อมูลรายการปรับปรุง (Transactional databases)
4. จากฐานข้อมูลพิเศษหรือที่เก็บข่าวสารพิเศษ ซึ่งได้แก่
   - ฐานข้อมูลเชิงวัตถุ
   - ฐานข้อมูลเกี่ยวกับเวลา
   - ฐานข้อมูลข้อความ (Text databases) และฐานข้อมูลมัลติมีเดีย
   - ฐานข้อมูลแบบเก่าในอดีตหรือข้อมูลที่มาจากต่างฐานข้อมูลกัน -
   - ข้อมูลจากแหล่ง WWW

2.2.6 ชนิดขององค์ความรู้ที่ค้นพบ มีดังนี้

1. องค์ความรู้เกี่ยวกับคุณลักษณะของข้อมูล (Characterization) เช่น รู้ว่าคนที่สามารถเรียนต่อในระดับปริญญาเอกได้จะพิจาณาได้จากคุณลักษณะใด

2. องค์ความรู้เกี่ยวกับการจำแนกข้อมูล (Discrimination)

3. องค์ความรู้เกี่ยวกับความสัมพันธ์ของข้อมูล (Association) เช่น มีความสัมพันธ์ของการซื้อสินค้าพบว่า ถ้าลูกค้าซื้อป๊อบคอร์น จะต้องซื้อเป๊ปซี่ตามมา

4. องค์ความรู้เกี่ยวกับการแยกประเภทข้อมูลและการพยากรณ์ (Classification/prediction)

5. องค์ความรู้เกี่ยวกับการจัดกลุ่มข้อมูล (Clustering)

6. องค์ความรู้เกี่ยวกับการวิเคราะห์ข้อมูลที่ผิดปกติ (Outlier analysis)

7. องค์ความรู้เกี่ยวกับข้อมูลอื่นๆ ในงานที่ค้นพบ (Other data mining tasks)

2.2.7 งานของการทำเหมืองข้อมูล

การทำเหมืองข้อมูลสามารถทำงานในการขุดค้นข้อมูล ดังนี้

1. การวิเคราะห์คุณสมบัติและการแยกแยะข้อมูล (Characterization and Discrimination)

การวิเคราะห์คุณสมบัติ เช่น การพิจารณารับสมัครพนักงานของฝ่ายทรัพยาการบุคคลต้องวิเคราะห์คุณสมบัติจากใบสมัครและการสัมภาษณ์

การแยกแยะข้อมูล เช่น การสูญเสียการได้ยินเป็นปัญหาสำคัญของคนงานที่ทำงานในโรงงานที่มีเสียงดัง ซึ่งการทราบถึงสมรรถภาพการได้ยินของตนเองจะช่วยให้คนงานสามารถป้องกันตนเองจากการสูญเสียการได้ยินได้

2. การหาความสัมพันธ์ของข้อมูล (Association) นิยมใช้ในการวิเคราะห์ตะกร้าสินค้า เพราะสามารถช่วยให้ค้นพบความสัมพันธ์ที่เป็นไปได้ทั้งหมดของการผสมกันของสินค้าที่น่าสนใจ เพื่อนำผลที่ได้ไปใช้ทำการตลาด

3. การจัดหมวดหมู่และการวิเคราะห์การถดถอย (Classification/ Regression)

การจัดหมวดหมู่ (Classification) ตัวอย่างของการจัดหมวดหมู่ ที่นำมาใช้กับงานด้านธุรกิจ เช่น การวิเคราะห์ลูกค้าที่ซื่อสัตย์จงรักภักดีต่อยี่ห้อสินค้า (Brand Loyalty) ขององค์กร

เทคนิคของ Data Mining ที่ใช้ ในการแก้ปัญหาแบบ Classification ได้แก่ Decision Tree, Neural Networks, Naïve-Bayes และ K-nearest neighbor (K-NN)

ปัญหาแบบ Regression จะเหมือนกับแบบ Classification ต่างกันตรงที่ผลลัพธ์ที่ได้จาก Regression เป็นค่าแน่นอน ที่ไม่จำกัด จะเป็นค่าอะไรก็ได้ เช่น แบบจำลองทำนายว่า นาย B จะตอบรับข้อเสนอของบริษัท ถ้านาย B ได้รับผลกำไร 1,000 บาท (1,000 เป็นคำตอบเฉพาะที่แน่นอน แต่ไม่จำกัด ซึ่งตัวเลขอาจจะเป็นค่าอื่นไปได้เรื่อย ๆ ต่างจากคำตอบแบบ Yes, No)

4. การวิเคราะห์การรวมกลุ่ม หรือ การแบ่งแยกข้อมูล (Cluster analysis/ Segmentation)

การวิเคราะห์การรวมกลุ่ม (Clustering) เป็นการรวมกลุ่มข้อมูลที่มีลักษณะเหมือนกัน รูปแบบหรือแนวโน้มที่จะเหมือนกัน การใช้เทคนิค Clustering จะไม่มีผลลัพธ์ (Output) ไม่มีตัวแปรอิสระ (Independent Variable) ไม่มีการจัดโครงร่างของวัตถุ เราจะเรียกเทคนิคของ Clustering ว่าเป็นแบบเรียนรู้ข้อมูลโดยไม่ต้องอาศัยครูสอน(Unsupervied Learning) การทำ Clustering จะทำบนพื้นฐานของข้อมูลในอดีต

ตัวอย่างเช่น องค์กรต้องการทราบความเหมือนที่มีในกลุ่มของลูกค้าของตน เพื่อที่จะให้เข้าใจลักษณะเฉพาะของลูกค้ากลุ่มเป้าหมาย และสร้างกลุ่มของลูกค้าเพื่อที่องค์กรจะได้สามารถขายสินค้าได้ในอนาคต องค์กรจะทำการแยกกลุ่มของข้อมูลลูกค้าออกเป็นกลุ่ม ๆ (หาส่วนที่เป็น Intersection และ Union)

เทคนิคของ Data Mining เพื่อแก้ปัญหาแบบ Clustering คือวิธี Demographic Clustering กับ Neural Clustering

5. การประเมินและการพยากรณ์ (Estimation/Prediction)

การประเมิน (Estimation) เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติที่ชัดเจนได้ ใช้จัดการกับค่าที่มีผลแบบต่อเนื่อง เช่น ใช้ประเมินรายได้ของครอบครัว ประเมินความสูงของบุคคลในครอบครัว ประเมินจำนวนเด็กๆ ในครอบครัว

การพยากรณ์ (Prediction) จะเหมือนกับ Classification และ Estimation ต่างกันตรงที่ Record ถูกแยกจัดลำดับในการทำนายค่าในอนาคต และนำข้อมูลในอดีตมาสร้างเป็นแบบจำลอง ใช้ทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น การทำนายว่าลูกค้ากลุ่มใด ที่องค์กรจะสูญเสียไปในอีก 6 เดือนข้างหน้า หรือ การทำนายยอดซื้อของลูกค้าจะเป็นเท่าใด ถ้าบริษัทลดราคาสินค้าลง 10%

6. การบรรยายและการแสดงภาพของข้อมูล (Description / Visualization)

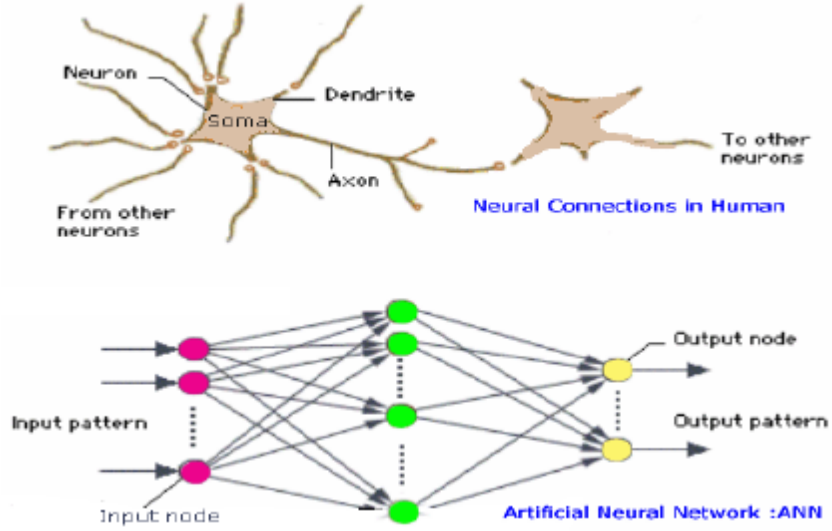
การบรรยาย (Description) เป็นการหาคำอธิบายถึงสิ่งที่จะเกิดขึ้น โดยอาศัยข้อมูลจากฐานข้อมูล เช่น กลุ่มคนที่มีการศึกษาหรือรายได้น้อย จะเลือกนักการเมืองที่มีนโยบายทุนนิยมมากกว่ากลุ่มคนชั้นกลาง

การแสดงภาพของข้อมูล (Visualization) เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิก หรืออาจนำเสนอในแบบ 2 มิติ สร้างรายละเอียดในการนำเสนอให้เข้าใจมากยิ่งขึ้น เช่น องค์กรต้องการหาสถานที่ในขยายสาขาใหม่ที่อยู่ในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้แผนที่ Plot ที่ตั้งขององค์กรคู่แข่งที่มีสาขาอยู่ในเขตนั้น เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด

2.2.8 เทคนิคและอัลกอริทึมในการทำเหมืองข้อมูล

1. Neural Network เป็นแนวคิดให้คอมพิวเตอร์ทำงานเสมือนสมองของมนุษย์ เปลี่ยนตัวเองจากการประมวลผลตามลำดับ (Sequential Processing) ให้เป็นการประมวลผลแบบคู่ขนานได้ (Parallel Processing) มีลักษณะการทำงานโดยแต่ละ Process จะรับ Input เข้าไปคำนวณ และสร้าง Output ออกมาในลักษณะที่ไม่ใช่การทำงานแบบเชิงเส้นตรง เพราะ Input แต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงกันนี้ จะถูกนำมาเปรียบเทียบกับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็จะนำไปสู่การปรับค่าหรือน้ำหนัก (weight) ของค่าที่ใส่ไว้ให้แต่ละ Input

Neural Network เป็นการสร้างแบบจำลอง ที่เลียนแบบการทำงานของสมองมนุษย์ มีโครงสร้างเป็นกลุ่มของ Node ที่เชื่อมโยงถึงกันในแต่ละ Layer คือ Input layer, Hidden layer, output layer



ภาพที่ 2.3 ตัวอย่างของ Neural Network

2. Decision Trees เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ (Decision Trees) ซึ่ง Decision Trees จะมีการทำงานแบบ Supervised Learning (คือการเรียนรู้แบบมีครูสอน) สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ก่อนล่วงหน้าเรียกว่า Training set ได้อัตโนมัติ และพยากรกลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ ได้ด้วยรูปแบบของ Tree โครงสร้างประกอบด้วย Root Node, Child และ Leaf Node

3. Memory Based Reasoning (MBR) เปรียบเหมือนกับประสบการณ์ การเรียนรู้ของมนุษย์ ซึ่งอาศัยการสังเกตที่เกิดขึ้น แล้วสร้างรูปแบบของสิ่งนั้นขึ้นมา เราใช้ MBR เพื่อวิเคราะห์ฐานข้อมูลที่มีอยู่ และกำหนดลักษณะพิเศษของข้อมูลที่อยู่ในนั้น ซึ่งข้อมูลจะต้องมีลักษณะที่สมบูรณ์ การสังเกตจึงจะสมบูรณ์และทำนายผลได้แม่นยำยิ่งขึ้น แบบจำลองจะถูกบอกคำตอบที่ถูกต้อง มีการเก็บคำตอบสำหรับแก้ปัญหาไว้ก่อนล่วงหน้าแล้ว (Supervised Learning)

4. Cluster Detection คือจะแบ่งฐานข้อมูลออกเป็นส่วน ๆ เรียกว่า Segment (กลุ่ม Record ที่มีลักษณะคล้ายกัน) ส่วน Record ที่ต่างกันก็จะอยู่นอก Segment, Cluster Detection ถูกใช้เพื่อค้นหากลุ่มย่อย (Sub Group) ที่เหมือน ๆ กันในฐานข้อมูล เพื่อที่จะเพิ่มความถูกต้องในการวิเคราะห์และสามารถมุ่งไปยังกลุ่มเป้าหมายได้ถูกต้อง

5. Link Analysis มุ่งเน้นทำงานบน Record ที่มีความสัมพันธ์กัน หรือเรียกว่า Association เทคนิคนี้จะมุ่งไปที่รูปแบบการซื้อหรือเหตุการณ์ที่เกิดขึ้นเป็นลำดับ มีอยู่ 3 เทคนิค คือ

5.1 Association Discovery ใช้วิเคราะห์การซื้อขายสินค้าในรายการเดียวกัน ศึกษาความสัมพันธ์อย่างใกล้ชิดที่ถูกปิ ดซ่อนอยู่ของสินค้า ซึ่งสินค้าเหล่านั้นอาจมีแนว โน้มที่จะถูกซื้อควบคู่กันไปการวิเคราะห์แบบนี้เรียกว่า Market Basket Analysis คือ รายการทั้งหมดที่ลูกค้าซื้อต่อครั้งที่ Super market การวิเคราะห์นี้สามารถนำมาใช้ประโยชน์ในการตัดสินใจ เช่น การเตรียมสินค้าคงเหลือ การวางแผนจัดชั้นวางสินค้า การทำ Mailing list สำหรับ Direct Mail การวางแผนเพื่อจัด Promotion สนับสนุนการขาย ตัวอย่างของ Association เช่น 75% ของผู้ซื้อน้ำอัดลมจะซื้อข้าวโพดคั่วด้วย

5.2 Sequential Pattern Discovery ถูกใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าของลูกค้า มีจุดหมายที่จะเข้าใจพฤติกรรมการซื้อสินค้าของลูกค้าในลักษณะ logn term เช่น ผู้ขายอาจพบว่าลูกค้าที่ซื้อ TV มีแนวโน้มที่จะซื้อ VDO ในเวลาต่อมา

5.3 Similar Time Sequence Discovery ค้นหาความเกี่ยวเนื่องกันระหว่างข้อมูล 2 กลุ่ม ซึ่งขึ้นต่อกันทางด้านเวลา โดยมีรูปแบบการเคลื่อนที่เหมือนกัน ผู้ขายสินค้ามักใช้เพื่อดูแนวโน้มเพื่อเตรียม Stock เช่น เมื่อไรก็ตามที่ยอดขายสินค้าน้ำอัดลมสูงขึ้น ยอดขายมันฝรั่งจะสูงขึ้นตาม

6. Genetic Algorithm (GA) เปรียบเสมือนเป็นการสร้างพันธุกรรมที่ดีสุด บนขั้นตอนของวิวัฒนาการทางชีวภาพ แนวคิดหลักคือ เมื่อเวลาผ่านไป วิวัฒนาการของเซลล์ชีวิตจะเลือกสายพันธุ์ที่ดีที่สุด “Fittest Species” GA มีความสามารถในการทำงานแบบรวมกลุ่มเข้าด้วยกัน เช่น มีการแบ่งกลุ่มและจัดรวมกลุ่มข้อมูลเป็น 3 ชุด ขั้นตอนการทำงานของ GA เริ่มจาก

จับกลุ่มข้อมูลเป็นกลุ่ม ๆ ด้วยการสุ่มเดา เปรียบเหมือนกลุ่ม 3 กลุ่มนี้เป็นเซลล์ของสิ่งมีชีวิต GA จะมี Fittest Function ที่จะบอกว่ากลุ่มข้อมูลใดเหมาะกับกลุ่ม ๆ ใด โดย Fittest Function จะเป็นตัวบ่งชี้ว่าข้อมูลเหมาะกับกลุ่มมากกว่าข้อมูลอื่น ๆ

GA จะมี Operator ซึ่งยอมให้มีการเลียนแบบและแก้ไขลักษณะของกลุ่มข้อมูล Operator จะจำลองหน้าที่ของชีวิตที่ถูกพบในธรรมชาติ คือ มีการแพร่พันธุ์ จับคู่ผสมพันธ์ และเปลี่ยนรูปร่างตามต้นแบบของพันธุกรรม เปรียบกับข้อมูลถ้ามีข้อมูลใดในกลุ่มถูกพบว่าตรงกับคุณสมบัติของ Fittest function แล้ว มันจะคงอยู่และถูกถ่ายเข้าไปในกลุ่มนั้น แต่ถ้าไม่ตรงกับคุณสมบัติ ก็ยังมีโอกาสที่จะถ่ายข้ามไปยังกลุ่มอื่นได้

7. Rule Induction ดึงเอาชุดกฎเกณฑ์ต่าง ๆ มาสร้างเป็นเงื่อนไขหรือกรณี วิธีการของ Rule Induction จะสร้างชุดของกฎที่เป็นอิสระ ซึ่งไม่จำเป็นต้องอยู่ในรูปแบบของโครงสร้างต้นไม้

8. K-nearest neighbor (K-NN) จะใช้วิธีในการจัดแบ่งคลาส โดยจะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวนของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไขหรือกรณีต่าง ๆ สำหรับแต่ละคลาสและกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกับมันมากที่สุด

K-NN ค่อนข้างใช้ปริมาณงานในการคำนวณสูงมากบนคอมพิวเตอร์ เพราะเวลาสำหรับการคำนวณจะเพิ่มขึ้นแบบแฟคทอเรียล ตามจำนวนจุดทั้งหมด เทคนิคของ K-NN จะมีการคำนวณเกิดขึ้นทุกครั้งที่มีกรณีใหม่ๆ เกิดขึ้น ดังนั้นถ้าจะให้เทคนิคแบบ K-NN ทำงานได้เร็ว ข้อมูลที่ใช้บ่อยควรเก็บอยู่ใน MBR (Memory-Based Reasoning)

9. Logic Regression เป็นการวิเคราะห์ความถดถอยแบบเส้นตรงทั่วๆ ไป ใช้ในการพยากรณ์ผลลัพธ์ของ 2 ตัวแปร เช่น Yes/No , 0/1 แต่เนื่องจากตัวแปรตาม (Dependent Variable) มีค่าเพียง 2 อย่างเท่านั้น จึงไม่สามารถสร้างแบบจำลอง (Model) ได้สำหรับการวิเคราะห์แบบ Logic Regression

ดังนั้นแทนที่จะทำการพยากรณ์โดยอาศัยเพียงค่าของตัวแปรตามที่ได้ เราจะสร้าง Model โดยอาศัย Algorithm ของความน่าจะเป็นของการเกิดเหตุการณ์ เราเรียก Algorithm นี้ว่า Log Odds หรือ Logic Transtromation

อัตราส่วนความน่าจะเป็น : $\frac{\text{ความน่าจะเป็นที่จะเกิดเหตุการณ์}}{\text{ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์}}$

2.2.9 การประยุกต์ใช้ Data Mining กับงานด้านธุรกิจ

สามารถนำเทคนิคของ Data Mining ไปวิเคราะห์ข้อมูลในฐานข้อมูล เพื่อนำข้อมูลที่ได้ไปใช้ประโยชน์ในงานด้านต่าง ๆ ดังต่อไปนี้

1. งานด้านการตลาด (Marketing) เช่น การทำ Promotion ส่งเสริมการขาย

2. งานด้านธนาคารและการเงิน (Banking / Financial Analysis) เช่น ใช้ในการวิเคราะห์การให้สินเชื่อแก่ลูกค้า การจัดทำ Package ในการกู้ยืม การทำนายอัตราการจ่ายเงินกู้ การแบ่งกลุ่มลูกค้าเพื่อหาเป้ าหมายทางการตลาด (ลูกค้าชั้นดี)

3. งานด้านการขายปลีก (Retailing and sales) เป็นงานที่มีการเก็บข้อมูลจำนวนมาก ประยุกต์ใช้เพื่อหากลยุทธ์ ทำให้เกิดการได้เปรียบคู่แข่งทางการค้าในการหาลักษณะการซื้อของลูกค้า ความสัมพันธ์ของการซื้อกับช่วงเวลา ความสัมพันธ์ระหว่างตัวสินค้า และการวิเคราะห์ประสิทธิภาพของการโฆษณา เป็นต้น ช่วยให้สามารถหาวิธีการตอบสนองความต้องการของลูกค้าได้มากที่สุด และอาจหมายถึงส่วนแบ่งทางการตลาดที่เพิ่มขึ้นนั่นเอง

4. งานด้านการวางแผนในการผลิตสินค้า (Manufacturing and production) เช่น การพยากรณ์ยอดจำนวนการผลิตสินค้าเพื่อให้ได้กำไรมากสุด

5. งานด้านนายหน้าและความปลอดภัยด้านการค้า (Brokerage and securities trading) เช่น การพัฒนาวิธีการเพื่อสร้างความเชื่อมั่นในเรื่องความปลอดภัยของข้อมูล ในขณะที่มีการพัฒนาวิธีการเข้าถึงข้อมูล และการ Mining ให้สะดวกต่อการใช้งานมากขึ้น

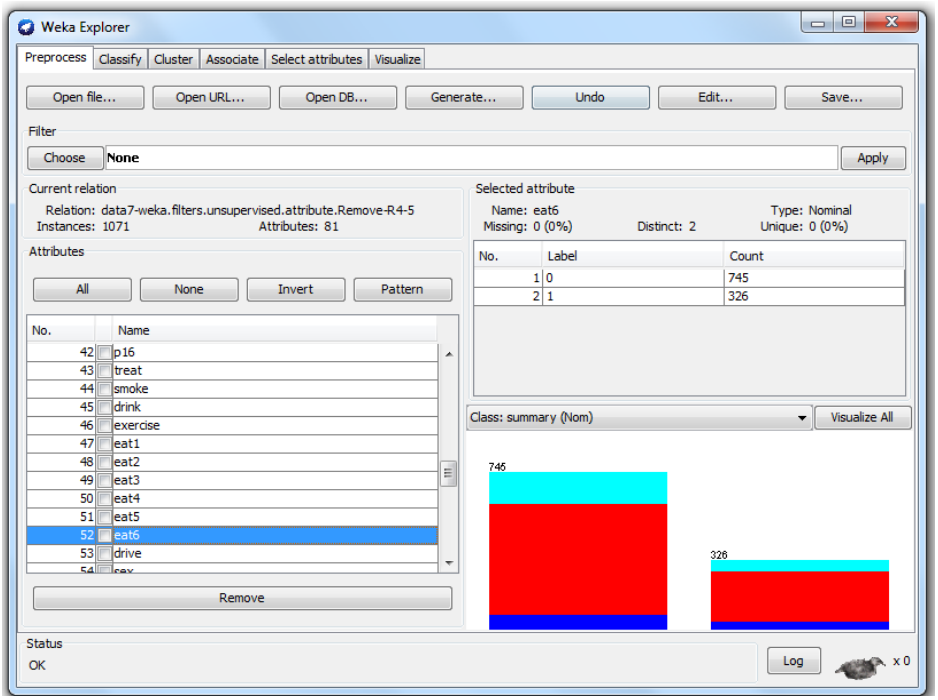
6. งานด้านชีวการแพทย์และวิเคราะห์ DNA (Biomedical an DNA Analysis) เช่น การวิเคราะห์รูปแบบการจัดเรียงตัวของหน่วยพันธุกรรม เพื่อหาสาเหตุความผิดปกติที่ทำให้เกิดโรค รวมไปถึงด้านการวินิจฉัยโรค การป้องกัน และการรักษา

นอกจากที่กล่าวมา ยังนำไปประยุกต์ใช้กับธุรกิจทางด้านประกันภัย (Insurance), Computer hardware และ software, หน่วยงานรัฐบาลและกระทรวงกลาโหม (Government and defense), สายการบิน (Airlines), งานด้านสุขภาพ (Health care), งานด้านการข่าว (Broadcasting) และงานด้านกฎหมาย (Law enforcement) ได้อีกด้วย

2.3 โปรแกรม Weka

Weka ย่อมาจาก Waikato Environment for Knowledge Analysis เป็นซอฟต์แวร์ฟรีที่แจกจ่ายภายใต้ GPL license โดยทั้งหมดนี้สร้างด้วยภาษาจาวาซึ่งประกอบไปด้วยอุปกรณ์การวิเคราะห์ข้อมูลและเพื่อนำไปใช้ทำเหมืองข้อมูลได้ เน้นการเรียนรู้ด้วยเครื่อง (Machine Learning)

กับการทำเหมืองข้อมูล (Data Mining) มีโมดูลย่อยสำหรับจัดการข้อมูล และตัวแสดงผลใช้ GUI และคำสั่ง ในการสั่งให้ซอฟต์แวร์ประมวลผล



ภาพที่ 2.4 แสดงหน้าจอการทำงานโปรแกรม Weka

โปรแกรมหลักของซอฟต์แวร์ Weka

- Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งการทำงานผ่านการพิมพ์
- Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
- Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
- Knowledge Flow เป็นโปรแกรมออกแบบผังการไหลของความรู้
- Arff Viewer เป็นโปรแกรมที่ใช้สำหรับแก้ไขแฟ้มประเภท Arff
- Log เป็นโปรแกรมที่ใช้อ่านข้อความบันทึกเก็บระหว่างการทำงาน

ในการวิจัยนี้จะใช้เฉพาะโปรแกรม Explorer ในการศึกษาวิเคราะห์และทำเหมืองข้อมูล ซึ่งมีหน้าจอการทำงานดังรูปที่ 2.4

เมนูหลักของ Explorer

- Preprocess การเตรียมข้อมูล

- Classify รวมโมดูลการทำเหมืองข้อมูลแบบจำแนกกลุ่ม
- Cluster รวมโมดูลการทำเหมืองข้อมูลแบบการจัดกลุ่ม
- Associate รวมโมดูลการทำเหมืองข้อมูลแบบกฎเชื่อมโยง
- Select attributes รวมโมดูลสำหรับการวิเคราะห์ความเกี่ยวพันของลักษณะประจำ
- Visualize นำเสนอข้อมูลด้วยภาพนามธรรมสองมิติ

ประเภทของแฟ้มข้อมูลที่นำเข้าข้อมูล ได้แก่

- แฟ้มข้อมูลในรูปแบบ arff, csv, C45
- ในกรณีแฟ้มข้อมูลอยู่ในเครือข่ายผู้ใช้สามารถเรียกใช้โดยอาศัย URL
- หรืออาจใช้ข้อมูลที่อยู่ในฐานข้อมูลที่เชื่อมโยงผ่าน JDBC

Weka เป็นโปรแกรที่ได้พัฒนาขึ้นบนพื้นฐานของภาษาจาวา (Java) สามารถรัน (run) ได้หลายระบบปฏิบัติการ และสามารถพัฒนาต่อยอดโปรแกรมได้ เป็นเครื่องมือที่ใช้ทำงานในด้านการทำเหมืองข้อมูล ที่รวบรวมแนวคิดอัลกอริทึมมากมาย ซึ่งอัลกอริทึมสามารถเลือกใช้งานโดยตรงได้ 2 ทางคือ จากชุดเครื่องมือที่มีอัลกอริทึมมาให้ หรือเลือกใช้จากอัลกอริทึมที่ได้เขียนเป็นโปรแกรมลงไปเป็นชุดเครื่องมือเพิ่มเติม และชุดเครื่องมือมีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูล ได้แก่ Pre-Processing, Classification, Regression, Clustering, Association rules, Selection และ Visualization ข้อดีของโปรแกรม Weka คือ มีอัลกอริทึมที่รู้จักกันดีของการทำเหมืองข้อมูลให้เลือกใช้อย่างครบถ้วน และสามารถเขียนฟังก์ชันเพิ่มเข้าไปในโปรแกรมได้

2.4 งานวิจัยที่เกี่ยวข้อง

ศุกร์ใจ วุฒิกิจโกศล (2551,น.77) ใช้เทคนิคการทำเหมืองข้อมูลในผู้ป่วยข้อไหล่ติด โรงพยาบาลพระนั่งเกล้าเพื่อหาปัจจัยที่มีผลต่อการรักษาและสร้างกฎความสัมพันธ์ที่น่าสนใจเพื่อเป็นแนวทางช่วยสนับสนุนการตัดสินใจเลือกเทคนิคการรักษาสำหรับนักกายภาพบำบัด โรงพยาบาลพระนั่งเกล้า โดยใช้ข้อมูลผู้ป่วยโรคข้อไหล่ติดที่ไม่มีโรคประจำตัวและไม่ได้เกิดจากอุบัติเหตุมารับการรักษาทางกายภาพบำบัดโรงพยาบาลพระนั่งเกล้าที่รักษาหายในช่วงปี พ.ศ. 2548-2550 จำนวน 550 ราย มีการเตรียมข้อมูลเพื่อให้ได้ข้อมูลที่ถูกต้อง และหาปัจจัยที่มีผลต่อการรักษา จากนั้นวิเคราะห์โดยการทำเหมืองข้อมูลด้วยเทคนิค Clustering และ Association Rule เลือกกฎความสัมพันธ์ที่น่าสนใจด้วยค่าสนับสนุนของกฎไม่ต่ำกว่าร้อยละ 20 และค่าความเชื่อมั่นไม่ต่ำกว่าร้อยละ 90 ผลการศึกษาพบว่ามีปัจจัย อายุ ระยะเวลาที่เป็น องศาการยกแขน การไขว้หลัง ระดับความเจ็บปวด และเทคนิคการรักษาที่มีผลต่อการรักษา ตัวอย่างเช่น เมื่อใช้เทคนิค Association Rule หากฎความสัมพันธ์ช่วงองศาการยกแขน 90-120 เลือกเทคนิคการรักษาด้วยแผ่นประคบความร้อน

บริเวณคอ ความร้อนคลื่นเหนือเสียงบริเวณไหล่ด้านหน้า ความร้อนคลื่นเหนือเสียงบริเวณกล้ามเนื้อ Infraspinatus และกระตุ้นไฟฟ้าแบบ Surge บริเวณไหล่ด้านหน้า-หลังจะใช้จำนวนครั้งในการรักษาช่วง 19-24 ครั้งด้วยค่าสนับสนุนของกฎ 43.48 เปอร์เซ็นต์และค่าความเชื่อมั่นที่ 100%

รักถิ่น เหลาหา (2553,น.45) ศึกษาการพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด โดยใช้ทฤษฎีของการทำเหมืองข้อมูล สร้างระบบที่จัดกลุ่มของผู้ป่วยและพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด ทั้งนี้การจัดกลุ่มจะให้ทราบถึงค่าความเสี่ยงของปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งปอด จากนั้นนำค่าปัจจัยเสี่ยงมาวิเคราะห์และพยากรณ์ผู้ป่วยด้วยต้นไม้ตัดสินใจแบบ C4.5 โดยใช้กลุ่มข้อมูลผู้ป่วยโรงพยาบาลมหาสารคาม ในเดือนสิงหาคม – ธันวาคม 2552 จำนวน 2,215 คน ผู้ป่วยมีสองกลุ่ม คือ ผู้ป่วยที่เป็นโรคมะเร็งปอดจำนวน 118 คน ไม่เป็นมะเร็งปอด 2,097 คนและได้วัดค่าประสิทธิภาพการพยากรณ์จากค่าความแม่นยำและค่าความระลึก จากผลการวิจัยพบว่าปัจจัยที่ส่งผลต่อค่าความเสี่ยงที่ทำให้เป็นมะเร็งปอดมากที่สุดคือ ปัจจัยประวัติด้านพันธุกรรม โดยมีค่าความเสี่ยงเป็น 34.59 เท่าของคนที่ไม่มีประวัติด้านพันธุกรรม รองลงมาคือ ปัจจัยเสี่ยง ด้านประวัติการสูบบุหรี่ ประวัติการดื่มสุรา และอายุ ตามลำดับ ผลการพยากรณ์ให้ค่าความแม่นยำ (accuracy) ที่ 0.796 ค่าความระลึก (recall) 0.5 และค่าความถูกต้อง (precision) ของการจำแนกประเภทเฉลี่ยมีค่าร้อยละ 96.8387

มาลิวรรณ บุญพลอย (2551,น.73) ได้ศึกษาเรื่องเหมืองข้อมูลและการค้นหาองค์ความรู้สำหรับโรคกล้ามเนื้ออ่อนแรง นำเสนอวิธีการใช้ขั้นตอนวิธีต้นไม้ช่วยตัดสินใจสองชั้นเพื่อจำแนกประเภทข้อมูลโรคกล้ามเนื้ออ่อนแรงเรื้อรัง ซึ่งกระบวนการทำงานของขั้นตอนวิธีนี้สามารถแบ่งออกเป็นสองขั้นตอนใหญ่ๆคือ ขั้นตอนการจำแนกกลุ่มผู้ป่วยปกติ (Non-fatigued - NF) และขั้นตอนการจำแนกกลุ่มผู้ป่วยที่มีอาการผิดปกติแต่ไม่ใช่เป็นโรคกล้ามเนื้ออ่อนแรงเรื้อรัง (Insufficient number of symptoms or fatigue severity - ISF) และ กลุ่มผู้ป่วยที่เป็นโรคกล้ามเนื้ออ่อนแรงเรื้อรัง (Chronic Fatigue Syndrome - CFS) วัตถุประสงค์ของงานวิจัยนี้เพื่อทำการพัฒนาอัลกอริทึมที่สามารถจำแนกประเภทข้อมูลของโรคกล้ามเนื้ออ่อนแรงเรื้อรังให้มีประสิทธิภาพมากขึ้นและทำการค้นหาองค์ความรู้ซึ่งจะทำการเปรียบเทียบประสิทธิภาพของวิธีต้นไม้ช่วยตัดสินใจสองชั้น (Double Decision Trees) วิธีกลุ่มของต้นไม้ (Decision Tree Forests) และวิธีต้นไม้ช่วยตัดสินใจ (Decision Trees) ข้อมูลที่นำมาใช้ในการทดลอง ประกอบด้วยข้อมูลของผู้ป่วยที่เป็นโรคกล้าเนื้ออ่อนแรงเรื้อรัง ซึ่งข้อมูลของผู้ป่วยที่เป็นโรกล้ามเนื้ออ่อนแรงนั้นประกอบไปด้วยข้อมูลผลเลือด และข้อมูลอาการป่วย ผลการวิจัยพบว่าประสิทธิภาพในการจำแนกประเภทข้อมูลโรคกล้ามเนื้ออ่อนแรงที่เรื้อรังโดยใช้ขั้นตอนวิธีต้นไม้ช่วยตัดสินใจสองชั้น นั้นสามารถจำแนกประเภทของผู้ป่วยที่เป็นโรคกล้ามเนื้ออ่อนแรงเรื้อรังได้แม่นยำกว่าวิธีกลุ่มของต้นไม้และวิธีต้นไม้ช่วย

ตัดสินใจ โดยวิธีต้นไม้ช่วยตัดสินใจสองชั้นสามารถจำแนกข้อมูลกลุ่ม NF, CFS และ ISF มีค่าความแม่นยำเท่ากับ 83.60%, 83.30% และ 83.34% ตามลำดับ วิธีกลุ่มของต้นไม้สามารถจำแนกข้อมูลกลุ่ม NF, CFS และ ISF มีค่าความแม่นยำเท่ากับ 81.94%, 82.20% และ 81.79% ตามลำดับ วิธีต้นไม้ช่วยตัดสินใจสามารถจำแนกข้อมูลกลุ่ม NF, CFS และ ISF มีค่าความแม่นยำเท่ากับ 67.00%, 66.26% และ 67.00% ตามลำดับ เนื่องจากวิธีต้นไม้ช่วยตัดสินใจสองชั้นมีการแบ่งการจำแนกประเภทออกเป็น 2 ขั้นตอนโดยอาศัยหลักการคัดกรองข้อมูลผู้ป่วย ซึ่งทำให้ปัญหาการจำแนกประเภทข้อมูลลดความซับซ้อนลงจากเดิมเนื่องจากปัญหาในการจำแนกข้อมูลเป็นแบบหลายคลาส (Multi-class Problem) ได้ถูกลดรูปเป็นปัญหาแบบสองกลุ่ม (Two-class Problem) ซึ่งเป็นปัญหาที่ง่ายทำให้ขั้นตอนวิธีต้นไม้ช่วยตัดสินใจสองชั้นสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพมากกว่า

อังคณา พิจารโชติ (2552,น.73) ได้พัฒนาระบบสนับสนุนการตัดสินใจสำหรับวิเคราะห์ปัจจัยเสี่ยงการเป็นโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล เพื่อช่วยในการวิเคราะห์หาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆที่ก่อให้เกิดโรคเบาหวาน เช่น พ่อแม่มีประวัติเป็นโรคเบาหวาน มีประวัติการปัสสาวะบ่อย เป็นต้น ข้อมูลความสัมพันธ์ของปัจจัยเสี่ยงที่ได้จากการวิเคราะห์จะเป็นสารสนเทศสำคัญที่เป็นตัวช่วยให้หน่วยงานด้านการแพทย์นำไปใช้สำหรับวางแผนป้องกันโรค โดยระบบที่พัฒนานี้จะประกอบด้วยสามส่วนคือ ส่วนของคลังข้อมูลใช้สำหรับเก็บข้อมูลปัจจัยเสี่ยงที่ได้จากการคัดกรองกลุ่มเสี่ยง ส่วนของเหมืองข้อมูลเป็นตัวค้นหาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆ และส่วนการแสดงข้อมูลรายงานซึ่งจะจัดทำในรูปแบบของเว็บแอพพลิเคชั่นเพื่อให้สะดวกต่อการเรียกใช้งาน ผลการค้นหาความสัมพันธ์ด้วยค่าสนับสนุนที่ 0.5 และค่าความเชื่อมั่นที่ 0.3 พบว่าประวัติครอบครัวเป็นปัจจัยเสี่ยงที่มีอัตราการเสี่ยงมากที่สุดซึ่งพบคู่กับค่าดัชนีมวลกาย มีค่าความเชื่อมั่นที่ 0.94

งานวิจัยนี้ใช้เทคนิค Association Rules เช่นเดียวกับงานวิจัยของ ศุกร์ใจ วุฒิกิจโกศล (2551,น.77) และ อังคณา พิจารโชติ (2552,น.73) สิ่งที่แตกต่างกันคือ งานวิจัยผู้ป่วยข้อไหล่ติดหาความสัมพันธ์ของเทคนิคการรักษากับผลการรักษา ส่วนงานวิจัยวิเคราะห์ปัจจัยเสี่ยงเบาหวานหาความสัมพันธ์ระหว่างปัจจัยเสี่ยงกับการเป็นโรค แต่งานวิจัยนี้หาความสัมพันธ์ของการเป็นโรคต่างๆ เทคนิค Classification ใช้ในงานวิจัยของ รักถิ่นเหลาหา (2553,น.45) ใช้อัลกอริทึม C4.5 ในการพยากรณ์ผู้ป่วยโรคมะเร็ง แต่งานวิจัยนี้ได้ทดลองอัลกอริทึม 2 ตัวคือ C4.5 และ Partial Rules แล้วจึงนำผลลัพธ์ที่ได้จากอัลกอริทึมที่มีค่า F-measure สูงกว่านั่นคือ Partial Rules นำไปพัฒนาระบบคัดกรองสุขภาพเบื้องต้น