การจำแนกโรคหลอดเลือดสมองโดยหาความสัมพันธ์ของปัจจัยร่วมกับการเกิดโรค

วิทยานิพนธ์

ของ

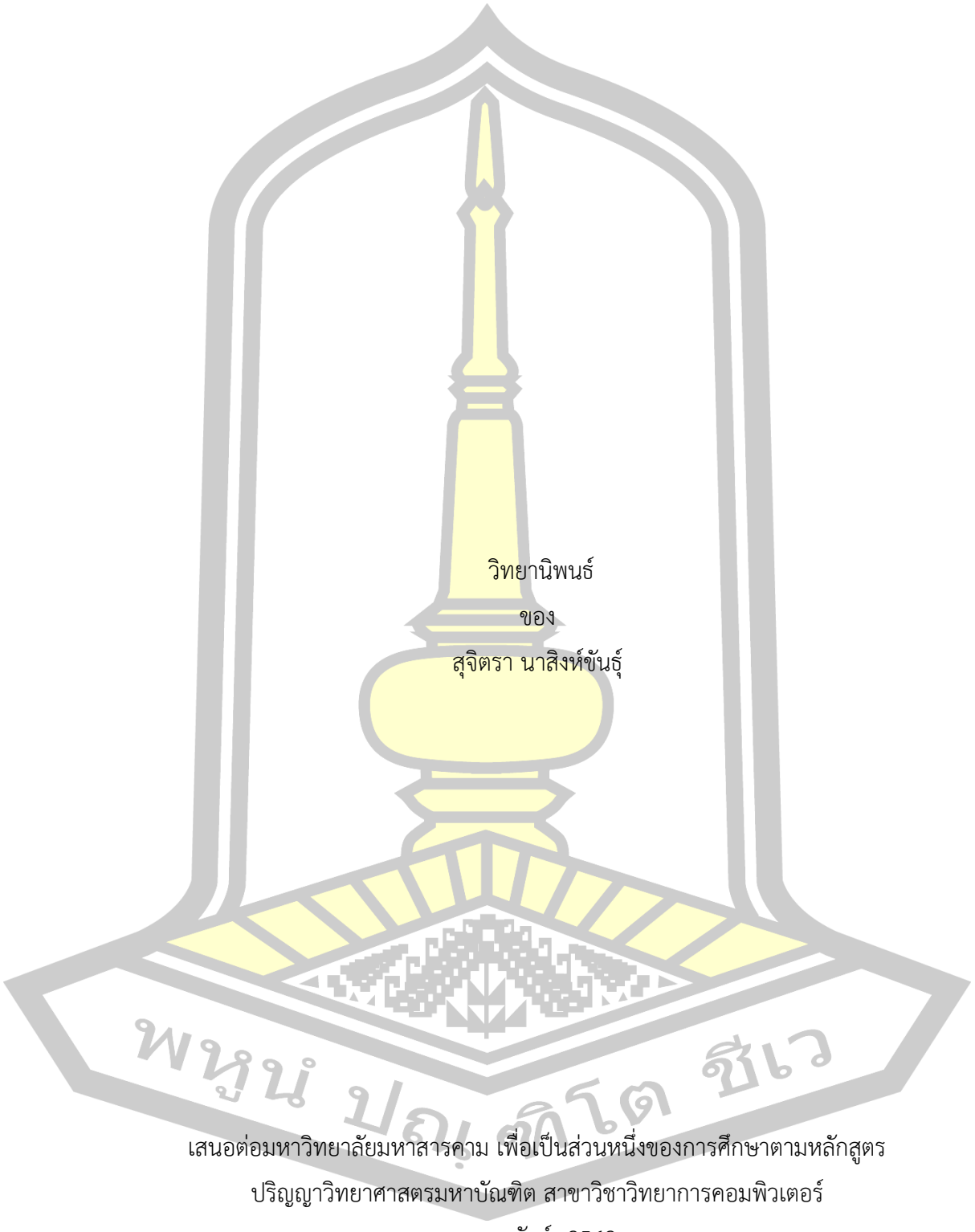
สุจิตรา นาสิงห์ขันธุ์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

กุมภาพันธ์ 2562

การจำแนกโรคหลอดเลือดสมองโดยหาความสัมพันธ์ของปัจจัยร่วมกับการเกิดโรค



วิทยานิพนธ์

ของ

สุจิตรา นาสิงห์ขันธุ์

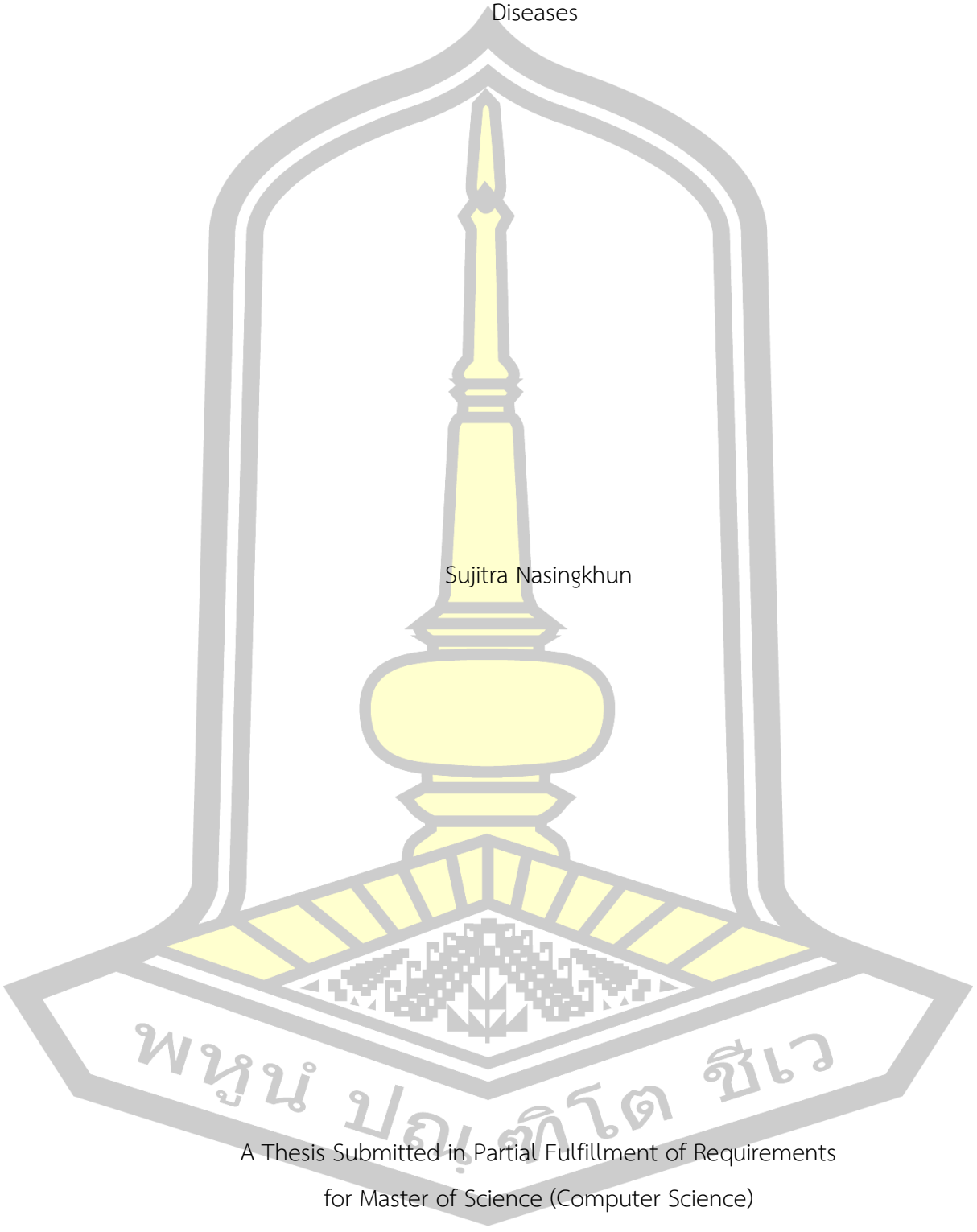
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

กุมภาพันธ์ 2562

Stroke Classification Based on the Relationships between Associated Factors and the Diseases

Sujitra Nasingkhun



A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Computer Science)

February 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางสาวสุจิตรา นาสิงห์ขันธุ์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ.ดร พนิดา ทรงรัมย์)

.....กรรมการ
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

.....กรรมการ
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....

(ผศ. ศศิธร แก้วมั่น) (ผศ. ดร. กริสน์ ชัยมูล)
คณบดีคณะวิทยาการสารสนเทศ คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การจำแนกโรคหลอดเลือดสมองโดยหาความสัมพันธ์ของปัจจัยร่วมกับการเกิดโรค

ผู้วิจัย สุจิตรา นาสิงห์ขันธุ์

อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ พนิดา ทรงรัมย์

ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

โรคหลอดเลือดสมองเป็นภาวะฉุกเฉินทางการแพทย์ที่ต้องการรักษาอย่างเร่งด่วน ซึ่งเป็นสาเหตุการเสียชีวิตอันดับ 3 ของโลกและเป็นอันดับ 1 ของประเทศไทยในเพศหญิงวัยสูงอายุ ดังนั้นจำเป็นจะต้องมีการคาดการณ์การเกิดโรคหลอดเลือดสมองเพื่อหาทางป้องกันและเตรียมการรักษาที่เหมาะสมสำหรับผู้ป่วยงานวิจัยที่ผ่านมาทำการศึกษาปัจจัยต่างๆเพื่อทำนายการเกิดโรคหลอดเลือดสมอง เช่น ความดันโลหิต การสูบบุหรี่ และคอเลสเตอรอล เป็นต้น ซึ่งแตกต่างจากงานวิจัยนี้ที่นำความสัมพันธ์ของลำดับการเกิดโรคและปัจจัยมาใช้ในการทำนายโรคหลอดเลือดสมอง โดยความสัมพันธ์แสดงอยู่ในรูปของกฎลำดับเหตุการณ์ที่มีคลาส ซึ่งแสดงให้เห็นความสัมพันธ์ของลำดับการเกิดโรคและปัจจัยที่นำไปสู่โรคหลอดเลือดสมอง (คลาส) โดยกฎถูกสร้างขึ้นจากวิธีการสืบค้นลำดับเหตุกาณ์และการจำแนกเชิงความสัมพันธ์ร่วมกัน จากผลการทดลองแสดงให้เห็นว่าเทคนิคที่นำเสนอให้ประสิทธิภาพสูงในการทำนาย และงานวิจัยนี้ยังแสดงกฎที่มีความสำคัญ 10 อันดับแรกที่นำไปสู่โรคหลอดเลือดสมอง

คำสำคัญ : การสืบค้นโรคหลอดเลือดสมอง, ความสัมพันธ์ลำดับการเกิดโรค, เหมืองข้อมูลแบบลำดับกฎ, การจำแนกแบบลำดับ

TITLE Stroke Classification Based on the Relationships between Associated Factors and the Diseases
AUTHOR Sujitra Nasingkhun
ADVISORS Assistant Professor Panida Songram , Ph.D.
DEGREE Master of Science **MAJOR** Computer Science
UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

Stroke is a medical emergency that need immediate medical attention. It is the third cause of death in the world and it is the first cause of death of elderly woman in Thailand. Stroke need to be predicted for prevention and early treatment. Many works tried to study factors, such as blood pressure, smoking, and cholesterol, for predicting stroke. Unlike the previous works, the disease sequence association is used for predicting stroke in this paper. The association is represented in form class sequential rule which shows association of diseases sequence leading to stroke. The combination of sequential pattern mining and associative classification is proposed as a method for generate class sequential rules. From the method, it give high performance for prediction. In addition, this paper shows top ten association of disease and factors leading to stroke.

Keyword : stroke detection, disease sequence association, sequential rule mining, sequence classification

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาจากผู้ช่วยศาสราจารย์ ดร.พนิดา ทรงรัมย์ ที่ปรึกษาวิทยานิพนธ์ที่ให้คำแนะนำ แนวคิด ตลอดจนแก้ไขข้อบกพร่องต่างๆ มาโดยตลอดจนวิทยานิพนธ์นี้เสร็จสมบูรณ์ จึงขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณผู้ช่วยศาสตราจารย์ ดร.วรารัตน์ สงฆ์แป้น ประธานกรรมการสอบ ศาสตราจารย์ ดร.พัฒนพงษ์ ชมพูวิเศษ กรรมการสอบและ ผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล กรรมกรสอบที่เป็นผู้ให้การสนับสนุนในการทำวิทยานิพนธ์ครั้งนี้

ขอขอบพระคุณคณะอาจารย์ เจ้าหน้าที่ ภาควิชาวิทยาการคอมพิวเตอร์และภาควิชาเทคโนโลยีสารสนเทศ ที่มอบความรู้ในการดำเนินงานวิจัยให้เป็นไปอย่างราบรื่น

ขอขอบคุณเพื่อนๆ พี่ น้องทุกคนที่มีส่วนร่วมในการช่วยเหลือและสนับสนุนตลอดระยะการดำเนินงานในการศึกษา จนผ่านลุล่วงไปได้ด้วยดี

ขอบพระคุณครอบครัวและคนรอบข้างที่ให้กำลังใจและสนับสนุนด้วยดีเสมอมา

สุจิตรา นาสิงห์ขันธุ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพประกอบ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ปัจจุบันความเจริญก้าวหน้าทางการแพทย์และสาธารณสุขพัฒนาอย่างรวดเร็วส่งผลให้ประชากรมีอายุยืนยาวและมีจำนวนประชากรผู้สูงอายุเพิ่มขึ้นทั่วโลก สำหรับประเทศไทยจากการสำรวจ “ ประชากรสูงอายุในประเทศไทยปี 2557 ” ของสำนักงานสถิติแห่งชาติ คาดการณ์ว่าในปี 2568 ประเทศไทยจะมีผู้สูงอายุเกินกว่าร้อยละ 20 หรือ 14.4 ล้านคนและประเทศไทยจะเป็นสังคมผู้สูงอายุโดยสมบูรณ์ ผลการศึกษาภาวะโรคและการบาดเจ็บของประชากรไทย พ.ศ.2557 โดยกระทรวงสาธารณสุขพบสาเหตุที่ทำให้ผู้สูงอายุสูญเสียปีสุขภาวะสูงที่สุดคือโรคหลอดเลือดสมองในเพศหญิงคิดเป็นร้อยละ 12.0 และร้อยละ 10.6 ในเพศชาย [1]

โรคหลอดเลือดสมอง (Cerebrovascular Disease, Stroke) เป็นสาเหตุที่ทำให้เกิดความพิการและเสียชีวิตที่สำคัญ โดยมีอัตราการเสียชีวิตทั่วโลก ข้อมูลจากองค์กรอนามัยโลก (World Health Organization: WHO) ปี 2008 [2] พบว่าเกิดจากการอุดตันของเส้นเลือดที่ไปเลี้ยงสมองและเส้นเลือดแดงตีบตันหรือเกิดจากก้อนเลือดไปอุดตัน ซึ่งเป็นปัญหาสาธารณะสุขที่สำคัญของประเทศไทย กล่าวคือเป็นสาเหตุการเสียชีวิตอันดับ 1 และอันดับ 3 ในหญิงและชายตามลำดับ จากสถิติการเสียชีวิตร้อยละ 10.0 อัตราการพิการร้อยละ 50.0 และเชื่อว่าในอนาคตแนวโน้มการเกิดโรคหลอดเลือดสมองมีมากขึ้นเรื่อยๆ กระทรวงสาธารณะสุขจึงประกาศให้โรคหลอดเลือดสมองเป็นปัญหาสุขภาพสำคัญของคนไทย จึงมีความจำเป็นอย่างยิ่งที่จะต้องเรียนรู้และเข้าใจแนวทางป้องกันการเกิดโรคหลอดเลือดสมองเนื่องจาก 80% ของโรคหลอดเลือดสมองสามารถป้องกันได้

ปัจจุบันได้มีการประยุกต์ใช้เหมืองข้อมูลในทางการแพทย์ เช่น ศุกร์ใจ วุฒิกิจโกศล[3] ใช้เทคนิคการทำเหมืองข้อมูลเพื่อหาปัจจัยที่มีผลต่อการรักษาในผู้ป่วยข้อไหล่ติดและสร้างกฎความสัมพันธ์ที่น่าสนใจช่วยสนับสนุนการตัดสินใจเลือกเทคนิคการรักษาให้แก่นักกายภาพบำบัด รักถิ่น เหลาหา[4] ใช้เทคนิคต้นไม้ตัดสินใจเพื่อจัดกลุ่มผู้ป่วยและพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอดโดยใช้ปัจจัย และอังคณา พิจารโชติ[5] วิเคราะห์ปัจจัยเสี่ยงการเป็นโรคเบาหวานเพื่อหาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆ เป็นต้น ซึ่งงานวิจัยส่วนใหญ่ใช้ปัจจัยเพียงอย่างเดียวในการวิเคราะห์โดยไม่ได้พิจารณาลำดับการเกิดโรคเข้ามาร่วม ซึ่งการเกิดโรคเป็นการเกิดอย่างต่อเนื่องและโรคหนึ่งอาจนำไปสู่อีกโรคหนึ่ง ดังนั้นงานวิจัยนี้จึงนำลำดับการเกิดโรคมาพิจารณาร่วมกับปัจจัยที่เกี่ยวข้องเพื่อจำแนกการเกิดโรค โดยนำเสนอวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับ

เหตุการณ์ในโรคหลอดเลือดสมองและหาความสัมพันธ์ของปัจจัยกับลำดับการเกิดโรคที่นำไปสู่โรคหลอดเลือดสมอง

1.2 วัตถุประสงค์ของการวิจัย

เพื่อนำเสนอวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมอง

1.3 ความสำคัญของการวิจัย

งานวิจัยนี้เสนอวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมอง ซึ่งวิธีการที่นำเสนอสามารถจำแนกโรคหลอดเลือดสมองโดยใช้ปัจจัยและลำดับการเกิดโรคได้อย่างมีประสิทธิภาพ นอกจากนี้งานวิจัยนี้ยังได้แสดงความสัมพันธ์ของปัจจัยและการเกิดโรคที่นำไปสู่การเกิดโรคหลอดเลือดสมอง ผลที่ได้สามารถนำไปใช้ในการพัฒนาระบบสาธารณะสุขในการดูแลรักษาผู้ป่วย ซึ่งมีความสำคัญต่อบุคลากรทางการแพทย์ สาธารณสุขและประชาชนผู้มารับบริการ

1.4 ขอบเขตของการวิจัย

1. สร้างวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมอง

2. หาความสัมพันธ์ของปัจจัยกับลำดับเกิดโรคที่นำไปสู่โรคหลอดเลือดสมอง

3. ข้อมูลที่ใช้สำหรับวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์จะต้องเป็นข้อมูลที่ประกอบด้วยข้อมูลที่ไม่พิจารณาลำดับการเกิดของข้อมูล คือ ปัจจัยเสี่ยง และข้อมูลที่พิจารณาลำดับการเกิดของข้อมูลเป็นสำคัญ คือ ลำดับการเกิดโรค พร้อมกับระบุคลาสที่จะจำแนกว่าเป็นหรือไม่เป็นโรคหลอดเลือดสมอง

4. ข้อมูลที่ใช้ในการทดลองในงานวิจัยนี้ คือ ข้อมูลปัจจัยเสี่ยงที่เกี่ยวข้องกับโรคหลอดเลือดสมองและลำดับการเกิดโรค ซึ่งได้จากผลการรักษาที่มารับบริการในโรงพยาบาลมหาสารคามจำนวน 1,000 คน เป็นผู้ป่วยที่อายุ 60 ขึ้นไปและมารับบริการในโรงพยาบาลตั้งแต่วันที่ 1 มกราคม พ.ศ. 2555 ถึง 31 ธันวาคม พ.ศ. 2559

5. งานวิจัยนี้วัดประสิทธิภาพในการจำแนก โดยใช้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าระลึก (Recal) และ ค่าอัตราการรู้จำ (F-measure)

1.5 นิยามศัพท์เฉพาะ

1. การจำแนก หมายถึง การสร้างตัวแบบเพื่อแยกข้อมูลออกจากกันโดยชัดเจน และนำตัวแบบที่ได้มาใช้ในการทำนายข้อมูล เพื่อระบุคลาสจากคุณสมบัติ เช่น หาความสัมพันธ์ระหว่างผลการตรวจร่างกายกับการเกิดโรค โดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยโดยแพทย์เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วยหรือการวิจัยทางการแพทย์

2. ลำดับการเกิดโรค หมายถึง ลำดับการเกิดอาการเจ็บป่วยที่เกิดจากพฤติกรรมหรือความไม่สมดุลของร่างกาย เกิดจากสภาวะเสื่อมของร่างกายหรือเกิดจากนิสัยหรือพฤติกรรมการดำเนินชีวิตซึ่งดำเนินอย่างช้าๆ การเกิดโรคหนึ่งอาจนำไปสู่อาการของโรคอื่นตามมา เช่น โรคที่ได้รับการวินิจฉัยว่าผู้ป่วยเป็นโรคนั้นและในเวลาต่อมาส่งผลให้เกิดโรคที่เป็นสาเหตุสำคัญทำให้ผู้ป่วยต้องมารับการรักษา เป็นต้น

3. ปัจจัยเสี่ยง หมายถึง ปัจจัยที่มีผลต่อสุขภาพ พฤติกรรมของบุคคลที่อาจจะส่งผลให้เกิดอันตรายต่อสุขภาพและชีวิต

4. ผลการรักษา หมายถึง ผลของการตรวจประเมินสภาพ การตรวจวินิจฉัยโรค การดูแลรักษาโรคผู้ป่วยโดยบุคลากรทางการแพทย์

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การศึกษาวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับโรคหลอดเลือดสมองประกอบไปด้วยแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ดังหัวข้อต่อไปนี้

2.1 โรคหลอดเลือดสมอง

โรคหลอดเลือดสมอง (Stroke) [6] เป็นภาวะสมองขาดเลือดหรืออีกชื่อหนึ่งก็คือโรคลมปัจจุบันทางการแพทย์เรียกว่า CardioVascular Disease (CVD) เกิดจากความผิดปกติของระบบประสาทเนื่องจากมีการเปลี่ยนแปลงของการไหลเวียนของเลือดภายในสมองหรือภาวะที่มีความบกพร่องของการทำงานของสมอง เนื่องจากหลอดเลือดสมองตีบหรืออุดตัน (Ischemic stroke) หรือหลอดเลือดสมองแตก (Hemorrhagic stroke) สมองจึงขาดเลือดไปเลี้ยง ทำให้เนื้อเยื่อในสมองถูกทำลาย การทำงานของสมองหยุดชะงัก โรคหลอดเลือดสมองเป็นโรคที่คุกคามต่อชีวิตและความเป็นอยู่ของคนทั่วโลก

โดยทั่วไปโรคหลอดเลือดสมองสมองอาจขาดเลือดทันทีภายในระยะเวลาเป็นนาทีหรือชั่วโมงแต่ไม่ใช่แบบค่อยเป็นค่อยไป โดยมีอาการที่เห็นได้ชัด คือ อ่อนแรงครึ่งซีก ชาครึ่งซีก เดินเซ พูดไม่ชัดหรือมองเห็นภาพซ้อนร่วมกับอาการต่างๆ ขึ้นอยู่กับบริเวณของสมองที่ขาดเลือด ความผิดปกติของหลอดเลือดที่ทำให้สมองขาดเลือด มีสาเหตุสำคัญที่ควรคำนึงอยู่ 2 ประการ คือ หลอดเลือดสมองอุดตันและหลอดเลือดสมองแตก

1. หลอดเลือดสมองอุดตัน (Ischemic stroke) พบได้ 70% ของโรคหลอดเลือดสมอง เกิดจากการที่เลือดไปเลี้ยงสมองไม่เพียงพอ ซึ่งเกิดจากสาเหตุสำคัญ 3 ประการ คือ

1.1 การอุดตันของหลอดเลือดจากการเสื่อมหรือการแข็งตัวของหลอดเลือด (Atherosclerosis) เป็นสาเหตุของหลอดเลือดอุดตันที่พบบ่อยที่สุด เกิดจากการที่ผู้ป่วยมีปัจจัยเสี่ยง เช่น สูงอายุ ความดันโลหิตสูง เบาหวาน สูบบุหรี่หรือไขมันในเลือดสูง เป็นต้น หลอดเลือดของผู้ป่วยจะค่อยๆ แข็งตัวและตีบลงเรื่อยๆ จากการที่มีไขมันใยไฟบรินและแคลเซียมมาสะสมที่ผนังหลอดเลือดที่เรียกว่า พลัค (plaque) เมื่อพลัคมีขนาดใหญ่ขึ้นจนเหลือช่องในหลอดเลือดเล็กลงเกิดการอุดตันทำให้ขาดเลือดไปเลี้ยงสมอง สมองหยุดทำงานและเกิดอาการของโรคหลอดเลือดสมองขึ้น

1.2 ก้อนเลือดจากหัวใจหรือตะกอนเลือดจากผนังหลอดเลือดแดงที่คอด้านหน้าหลุดเข้าไปอุดตันหลอดเลือดในสมอง มักเกิดในคนที่มีการเต้นหัวใจไม่สม่ำเสมอชนิดหัวใจห้องซ้ายบนเต้นพลิ้ว (Atrial Fibrillation หรือ AF) การเต้นของหัวใจที่บีบตัวไม่พร้อมกันทั้งห้อง ทำให้เลือดค้างในห้อง

หัวใจ เลือดจะเกิดการแข็งตัวเป็นก้อนเลือดหลุดเข้าไปในสมอง นอกจากนี้ตะกอนเลือดที่อยู่ที่ผิวของ plaque ในผนังหลอดเลือดใหญ่ที่คอสามารถหลุดเข้าไปติดในหลอดเลือดสมองทำให้เกิดการอุดตันของหลอดเลือดสมองได้เช่นกัน

1.3 ความดันเลือดลดลงมากจนไปเลี้ยงสมองไม่พอเป็นสาเหตุที่พบน้อยมาก

2. หลอดเลือดสมองแตก (Hemorrhagic stroke) พบได้ประมาณ 30% เกิดจากหลอดเลือดมีความเปราะบางร่วมกับภาวะความดันโลหิตสูง ทำให้หลอดเลือดบริเวณที่เปราะบางโป่งพองและแตกออกหรือสูญเสียความยืดหยุ่นจากการสะสมของไขมันในหลอดเลือดทำให้หลอดเลือดบริเวณนั้นปริแตกได้ง่าย ส่งผลให้ปริมาณเลือดที่ไปเลี้ยงสมองลดลงในทันทีและเกิดเลือดออกในสมอง เป็นสาเหตุให้ผู้ป่วยเสียชีวิตในเวลาอันรวดเร็วได้ พบได้บ่อยในผู้ป่วยโรคความดันโลหิตสูงและโรคหลอดเลือดสมองโป่งพอง (Aneurysm) นอกจากนี้ยังเกิดจากความเครียด การดื่มแอลกอฮอล์รวมทั้งยาบางชนิด

ปัจจุบันประชากรทั่วโลกป่วยเป็นโรคหลอดเลือดสมองจำนวน 17 ล้านคน [7] เสียชีวิตจากโรคหลอดเลือดสมองจำนวน 6.5 ล้านคนและมีชีวิตรอดจากโรคหลอดเลือดสมองจำนวน 26 ล้านคน ซึ่งผู้ที่มีชีวิตรอดจากโรคหลอดเลือดสมองจำนวนมากได้รับผลกระทบทั้งทางด้านสภาพร่างกาย จิตใจ สังคม รวมถึงความสูญเสียทางด้านเศรษฐกิจ

สำหรับในประเทศไทย พบว่าอัตราการตายด้วยโรคหลอดเลือดสมองต่อประชากรแสนคนในภาพรวมของประเทศในปี 2556-2558 เท่ากับ 36.13, 38.66 และ 42.62 ตามลำดับ จะเห็นได้ว่าอัตราการตายด้วยโรคหลอดเลือดสมองเพิ่มขึ้นทุกปี นอกจากนี้อัตราผู้ป่วยในด้วยโรคหลอดเลือดสมองต่อประชากรแสนคน ในภาพรวมของประเทศในปี 2557 เท่ากับ 352.30 ข้อมูลจากรายงานของสำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข

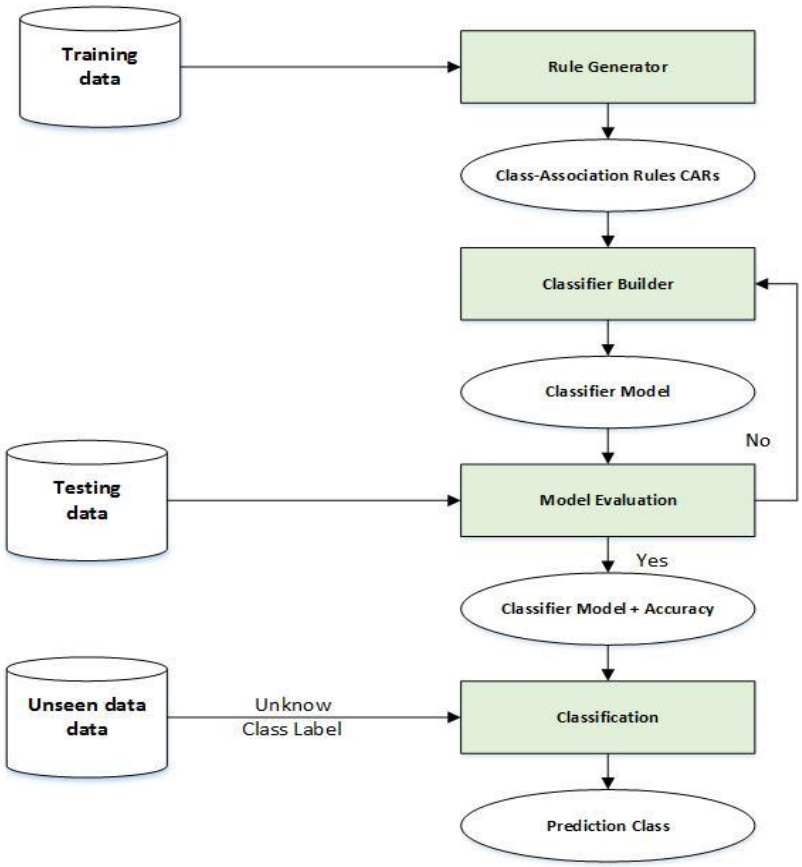
จากสถานการณ์ของโรคหลอดเลือดสมองของโลกและของประเทศไทย แสดงให้เห็นว่าโรคหลอดเลือดสมองเป็นภัยที่กำลังคุกคามประชากรทั่วโลกซึ่งในปีพ.ศ. 2559 องค์การอัมพาตโลก (World Stroke Organization: WSO) ได้ให้ความสำคัญและมุ่งเน้นการป้องกันควบคุมโรคหลอดเลือดสมองใน 3 ประเด็นหลัก คือ สร้างความตระหนักต่อโรค (Awareness) ส่งเสริมการเข้าถึงบริการสาธารณสุข (Access) และการลงมือปฏิบัติเพื่อควบคุมป้องกันโรคหลอดเลือดสมอง (Action)

2.2 การจำแนกเชิงความสัมพันธ์

การจำแนกข้อมูลเป็นการจำแนกประเภทข้อมูลโดยเรียนรู้จากชุดข้อมูลเรียนรู้ (Training set) ซึ่งมีการระบุคลาสปลายทาง (Class label) การจำแนกข้อมูลจะแบ่งข้อมูลออกเป็น 2 ส่วน คือ ส่วนของชุดข้อมูลเรียนรู้ใช้สำหรับสร้างโมเดลเพื่อใช้ในการทำนายคลาส ส่วนชุดข้อมูลสำหรับทดสอบ (Testing set) ใช้สำหรับทดสอบประสิทธิภาพในการจำแนกของโมเดลที่ถูกสร้างขึ้น โดยการสร้าง

โมเดลเพื่อใช้ในการทำนายข้อมูลนั้น ต้องผ่านขั้นตอนดังต่อไปนี้ เริ่มจากนำข้อมูลเรียนรู้เข้าสู่กระบวนการสร้างโมเดลจำแนกประเภทข้อมูล (Model construction) เพื่อให้โมเดลจำแนกประเภทข้อมูล (Classifier model) ออกมา หลังจากได้โมเดลจำแนกประเภทข้อมูลแล้ว ต้องทำการทดสอบว่าโมเดลที่ถูกสร้างขึ้นมีประสิทธิภาพในการจำแนกเพียงพอที่จะนำไปใช้หรือไม่ ดังนั้นจึงต้องใช้ข้อมูลทดสอบเพื่อทดสอบประสิทธิภาพในการจำแนกของโมเดลที่ถูกสร้างขึ้นมา (Model evaluation) ถ้าประสิทธิภาพของโมเดลเป็นที่น่าพึงพอใจ ก็จะนำโมเดลดังกล่าวไปประยุกต์ใช้เพื่อทำนายประเภทข้อมูลใหม่ (Unseen data) ที่ไม่ทราบประเภทข้อมูล (Unknown class label) ปัจจุบันนักวิจัยได้พัฒนาและนำเสนอวิธีการที่หลากหลายเพื่อสร้างตัวจำแนกที่มีประสิทธิภาพ เช่น Naive Bayes, K-Nearest Neighbor (KNN)[8] เป็นต้น

การจำแนกเชิงความสัมพันธ์ (Associative Classification) เป็นวิธีการจำแนกที่ได้รับความนิยมอย่างมาก เนื่องจากทำการจำแนกโดยใช้กฎที่ง่ายต่อความเข้าใจและมีความถูกต้องสูง การจำแนกแบบเชิงความสัมพันธ์เป็นเทคนิคที่เกิดจากการรวมกันระหว่าง 2 เทคนิค คือ 1) การจำแนกประเภทข้อมูล (Data classification)[13] และ 2) การสืบค้นกฎความสัมพันธ์ (Association rule discovery) การจำแนกเชิงความสัมพันธ์เป็นการค้นหากฎความสัมพันธ์ที่ใช้สำหรับการจำแนกจากฐานข้อมูลขนาดใหญ่ โดยที่กฎเหล่านั้นจะต้องผ่านค่าสนับสนุนขั้นต่ำ (Minimum Support) และค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) ขั้นตอนการจำแนกเชิงความสัมพันธ์สามารถแบ่งออกเป็น 2 ส่วนหลัก คือ วิธีการสร้างกฎ (Rule generator) และการสร้างโมเดลในการทำนาย (Classifier builder) โดยภาพรวมขั้นตอนทั้งหมดของเทคนิคการจำแนกประเภทข้อมูลอธิบายได้ดังภาพประกอบที่ 1



ภาพประกอบที่ 1 ขั้นตอนการจำแนกเชิงความสัมพันธ์

2.2.1 การสร้างกฎความสัมพันธ์ (Rule Generator Phase)

การสร้างกฎความสัมพันธ์ใช้หลักการหรือวิธีการเดียวกันกับเทคนิค Association rule discovery เกือบทั้งหมด การสร้างกฎที่ใช้ในการจำแนกสามารถสร้างได้จากเซตรายการความถี่ (Frequent Itemset Mining) กฎที่ถูกสร้างจากกระบวนการสร้างกฎความสัมพันธ์นั้นจะต้องเป็นกฎเฉพาะที่เรียกว่า กฎความสัมพันธ์แบบมีคลาสหรือ CARs (Class-Association Rules) นั่นคือกฎความสัมพันธ์ที่ทางด้านขวามือของกฎจะต้องเป็นคลาสเท่านั้น ซึ่งกฎที่ใช้ในการจำแนกอยู่ในรูปแบบของ r: $X \rightarrow c$ โดยที่ X คือ เซตรายการและ c คือ คลาส กฎถูกนำไปใช้ในการจำแนกเมื่อค่าสนับสนุน (Support) ของกฎมีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum Support Threshold: *min_supp*) และค่าความเชื่อมั่นของกฎมีค่ามากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence: *min_conf*) นิยามที่เกี่ยวข้องกับการจำแนกเชิงความสัมพันธ์มีดังต่อไปนี้

กำหนดให้ D เป็นเซตของทรานแซคชันและ $I = \{i_1, i_2, ..., i_n\}$ เป็นเซตของรายการที่ปรากฏใน D และ C คือเซตของคลาสที่ปรากฏใน D แต่ละทรานแซคชัน $d \in D$ อยู่ในรูปแบบ $(i_1, i_2, ..., i_k, c)$ โดยที่ $i \in I$ และ $c \in C$ ดังแสดงในตารางที่ 2.1 ซึ่ง I = {A,B,C,D,E} และ C={Y,N}

ตารางที่ 1 ตัวอย่างข้อมูลทรานเซคชัน

หมายเลขทรานแซคชัน	รายการ	คลาส
1	*A B D E*	*Y*
2	*B C E*	*Y*
3	*A B D E*	*Y*
4	*A B C E*	*N*
5	*A B C D E*	*N*

นิยามที่ 2.1 เซตรายการ $X = (i_1, i_2, ..., i_k)$ คือ ซับเซตของ I

นิยามที่ 2.2 $|g(X)|$ คือ จำนวนทรานแซคชันทั้งหมดที่ปรากฏเซตรายการ *X*

ตัวอย่างที่ 2.1 จำนวนทรานแซคชันทั้งหมดที่มีเซตรายการ (*AB*) ปรากฏอยู่ คือ 1 3 4 และ 5 ดังนั้น $|g(X)|$ = |{1, 3, 4, 5}| = 4

นิยามที่ 2.3 ค่าสนับสนุนของ X คือ เปอร์เซ็นต์ที่ปรากฏเซตรายการ *X* ต่อจำนวนทรานแซคชันทั้งหมด แทนดังสมการ 2.1

$$supp(X) = \frac{|g(X)|}{|g(D)|} \times 100 \qquad (2.1)$$

ตัวอย่างที่ 2.2 ค่าสนับสนุนของเซตรายการ (*AB*) คือ

supp(AB) = (|{1, 3, 4, 5}|/|{1, 2, 3, 4, 5}|) *100 = (4/5)*100 = 80%

นิยามที่ 2.4 กำหนดให้ค่าสนับสนุนขั้นต่ำ คือ min_supp เซตรายการ *X* เรียกว่าเซตรายการความถี่ก็ต่อเมื่อ $supp(X) \geq min\_supp$

ตัวอย่างที่ 2.3 ถ้ากำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 60% เซตรายการ (*AB*) เป็นเซตรายการความถี่เนื่องมาจากค่าสนับสนุนของเซตรายการ (*AB*) มากกว่าค่าสนับสนุนขั้นต่ำ นั้นก็คือ *supp*(AB) = 80%>60%

นิยามที่ 2.5 กำหนดให้กฎความสัมพันธ์อยู่ในรูป $r : X \rightarrow c$ ค่าสนับสนุนของกฎ *r* คือ เปอร์เซ็นต์ที่ปรากฏเซตรายการ *X* พร้อมกับ *c* ต่อจำนวนทรานแซคชันทั้งหมดแทนได้ดังสมการ 2.2

$$supp(r) = \frac{|g(X \rightarrow c)|}{|g(D)|} \times 100 \qquad (2.2)$$

ตัวอย่างที่ 2.4 ค่าสนับสนุนของกฎ $r : AB \rightarrow Y$ คือ 40% เนื่องจากทรานแซคชันที่ปรากฏเซตรายการ AB พร้อมกับ Y คือ 1 และ 3 ดังนั้น *supp*(*r*) = |{1,3}|/|{1,2,3,4,5}| * 100 = 2/5 *100 = 40%

นิยามที่ 2.6 ค่าความเชื่อมั่นของ r คือ เปอร์เซ็นต์ที่ปรากฏเซตรายการ *X* พร้อมกับ c ต่อจำนวนทรานแซคชันที่ปรากฏ *X* แทนได้ดังสมการ 2.3

$$conf(r) = \frac{|g(X \rightarrow c)|}{|g(X)|} \times 100 \qquad (2.3)$$

ตัวอย่างที่ 2.5 ค่าความเชื่อมั่นของกฎ $AB \rightarrow Y$ = |{1,3}|/|{1,3,4,5}| * 100 = 2/4*100 = 50%

นิยามที่ 2.7 กำหนดให้ค่าความเชื่อมั่นขั้นต่ำ คือ min_conf และ ค่าสนับสนุนขั้นต่ำ คือ min_supp กฎถูกใช้ในการจำแนกก็ต่อเมื่อ $supp(r) \geq min\_supp$ และ $conf(r) \geq min\_conf$

ตัวอย่าง 2.6 สมมติกำหนดให้ค่าสนับสนุนขั้นต่ำเท่ากับ 40% และค่าความเชื่อมั่นขั้นต่ำเท่ากับ 50% กฎ $AB \rightarrow Y$ จะถูกนำไปใช้ในการจำแนกเนื่องจากมีค่าสนับสนุนเท่ากับค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นเท่ากับค่าความเชื่อมั่นขั้นต่ำ

2.2.2 การสร้างโมเดลเพื่อใช้ทำนายข้อมูล (Classifier builder phase)

กฎความสัมพันธ์ที่ได้จากส่วนการสร้างกฎ มาใช้เพื่อสร้างโมเดลในการทำนายข้อมูลโดยในการทำนายข้อมูลนั้นจะมีการพิจารณากฎความสัมพันธ์ทีละกฎ(Single rule) โดยวิธีการพิจารณาแบบนี้จะต้องทำการเรียงลำดับกฎความสัมพันธ์ก่อนโดยทั่วไปแล้วจะเรียงลำดับกฎความสัมพันธ์ดังต่อไปนี้

1. เรียงตามค่าความเชื่อมั่น(Confidence) สูงก่อน
2. ถ้าค่าความเชื่อมั่นของกฎความสัมพันธ์เท่ากันก็จะเรียงลำดับของกฎความสัมพันธ์ตามค่าสนับสนุน(Support)
3. ถ้าค่าสนับสนุนของกฎเท่ากันจะเรียงลำดับกฎโดยดูจากความยาวทางด้านฝั่งซ้ายของกฎ กฎที่มีความยาวน้อยกว่าจะถูกเรียงก่อน
4. ถ้าความยาวทางด้านฝั่งซ้ายของกฎเท่ากันให้เรียงตามลำดับการสร้างกฎ กฎไหนถูกสร้างก่อนให้เรียงก่อน

เมื่อเรียงกฎความสัมพันธ์เป็นที่เรียบร้อยแล้วสามารถนำกฎที่เรียงแล้วไปทำนายข้อมูลโดยการทำนายข้อมูลนั้นจะทำนายตามคลาส(class) ของกฎที่มีศักย์(Precedence)

2.3 การสืบค้นลำดับเหตุการณ์ (Sequential Pattern Mining)

การสืบค้นลำดับเหตุการณ์ เป็นการสืบค้นลำดับเหตุการณ์ที่เกิดร่วมกันบ่อย โดยลำดับเหตุการณ์ดังกล่าวเป็นลำดับเหตุการณ์ที่พิจารณาเรื่องเวลาการเกิดของข้อมูล การสืบค้นลำดับเหตุการณ์จะค้นหาลำดับเหตุการณ์ความถี่ ซึ่งเป็นลำดับเหตุการณ์ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ ปัจจุบันมีขั้นตอนวิธีหลายวิธีที่นำเสนอเพื่อสืบค้นลำดับเหตการณ์ความถี่ เช่น GSP, FreeSpan และ Prefix-Span เป็นต้น

2.3.1 อัลกอริทึม Prefix-Span (Prefix Projected SPM)

อัลกอริทึม Prefix-Span เป็นอัลกอริทึมที่มีประสิทธิภาพในการสืบค้นลำดับเหตุการณ์และเป็นอัลกอริทึมที่มีประสิทธิภาพมากกว่าอัลกอริทึมอื่นเมื่อเทียบกับ GSP และ FreeSpan โดยอัลกอริทึม Prefix-Span ทำการสำรวจ Prefix โปรเจคชั่นในการทำเหมืองข้อมูลแบบต่อเนื่อง ซึ่งเป็นการค้นหารูปแบบที่สมบรูณ์ แต่ลดการสร้างลำดับย่อยของแคนดิเดท และ Prefix โปรเจคชั่นจะลดขนาดของโปรเจคฐานข้อมูลและนำไปสู่การประมวลผลที่มีประสิทธิภาพ ดังนั้นจึงนำอัลกอริทึม Prefix-Span มาใช้ในการแก้ปัญหาในวิจัยฉบับนี้ โดยจะยกตัวอย่างขั้นตอนการทำงานดังนี้

ตารางที่ 2 ฐานข้อมูลลำดับ

SID	Time(EID)	Items
1	10	CD
1	15	ABC
1	20	ABF
1	25	ACDF
2	15	ABF
2	20	E
3	10	ABF
4	10	D G H
4	20	B F
4	25	A G H

ตารางที่ 3 แสดงการจัดรูปแบบฐานข้อมูลแบบแนวนอน

SID	TIME(EID)	ITEMS
1	10,15,20,25	<(CD)(ABC)(ABF)(ACDF)>
2	15,20	< (ABF)(E) >
3	10	< (ABF) >
4	10,20,25	< (DGH)(BF)(AGH) >

ตัวอย่างความหมายของข้อมูล ITEMS ในแถวที่สอง <(ABF)(E)> ในวงเล็บเดียวกันหมายถึงเกิดในครั้งเดียวกันซึ่งครั้งเดียวอาจเกิดหลายเหตุการณ์ เช่น การมารับการรักษาพยาบาลหนึ่งครั้งอาจมีโรคได้หลายโรค เช่น นาย ก มาพบแพทย์วันที่ 1 ถูกระบุว่าเป็นโรค (ABF) ในเวลาต่อมานาย ก มาพบแพทย์ถูกระบุว่าเป็นโรค (E) สามารถแทนข้อมูลด้วย <(ABF)(E)> เป็นต้น

เพื่อแสดงให้เห็นถึงขั้นตอนการทำงานของอัลกอริทึม Prefix-Span จึงกำหนดค่าสนับสนุนขั้นต่ำ 50% หรือความถี่ต่ำสุดเท่ากับ 2 ซึ่งเป็นการกำหนดค่าสนับสนุนแบบสัมบูรณ์ (Absolute Support) ดังต่อไปนี้

1. ขั้นตอนการนับจำนวนของการเกิดแต่ละเหตุการณ์

จากตารางที่ 2.3 แสดงข้อมูลเหตุการณ์ทั้งหมด โดยขั้นตอนแรกจะนับจำนวนของการเกิดแต่ละเหตุการณ์ โดยเหตุการณ์ที่จำนวนความถี่ของเหตุการณ์ผ่านค่าสนับสนุนที่กำหนดไว้ คือ เหตุการณ์A, B, D, F เรียกว่า ลำดับเหตุการณ์ความถี่ 1-ลำดับ (1-sequence)

ตารางที่ 4 ข้อมูลเหตุการณ์

SID	ITEMS
1	<(CD)(ABC)(ABF)(ACDF)>
2	< (ABF)(E) >
3	< (ABF) >
4	< (DGH)(BF)(AGH) >

เมื่อนับจำนวนการเกิดของแต่ละเหตุการณ์สามารถแสดงได้ดังตารางที่ 5

ตารางที่ 5 จำนวนการเกิดแต่ละเหตุการณ์

เหตุการณ์	A	B	C	D	E	F	G	H
จำนวนความถี่ของเหตุการณ์	4	4	1	2	1	4	1	1

2. ขั้นตอนการแบ่งพื้นที่การค้นหา

นำเหตุการณ์ที่จำนวนความถี่ของเหตุการณ์ผ่านค่าสนับสนุนมาแบ่งพื้นที่การค้นหาจะแสดงได้ดังตารางที่ 6

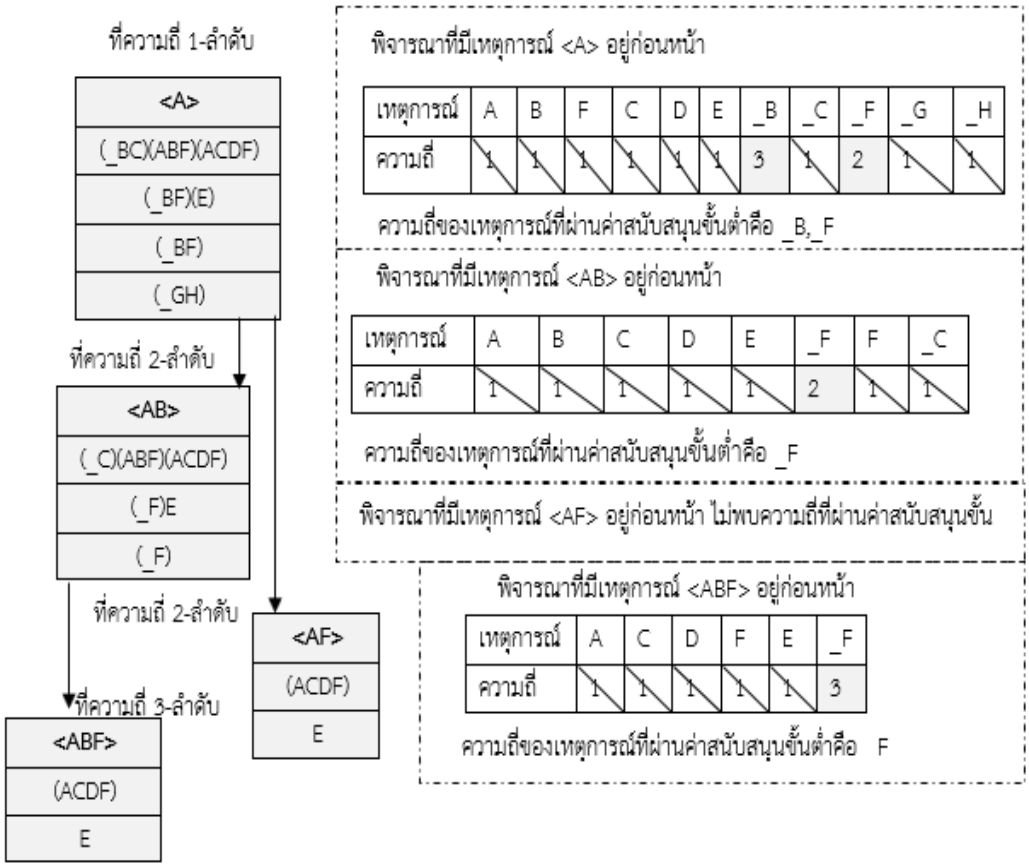
ตารางที่ 6 การแบ่งพื้นที่ค้นหา

<A>	<B>	<D>	<F>
(_BC)(ABF)(ACDF)	(_C)(ABF)(ACDF)	(ABC)(ABF)(ACDF)	(ACDF)
(_BF)(E)	(_F)(E)		(E)
(_BF)	(_F)		
(_GH)	(_F)(AGH)	(_GH)(BF)(AGH)	(AGH)

3. ขั้นตอนการค้นหาชุดเหตุการณ์ย่อย

ค้นหาชุดเหตุการณ์ย่อยของแต่ละเหตุการณ์ที่แบ่งพื้นที่การค้นหา โดยเริ่มจาก <A>

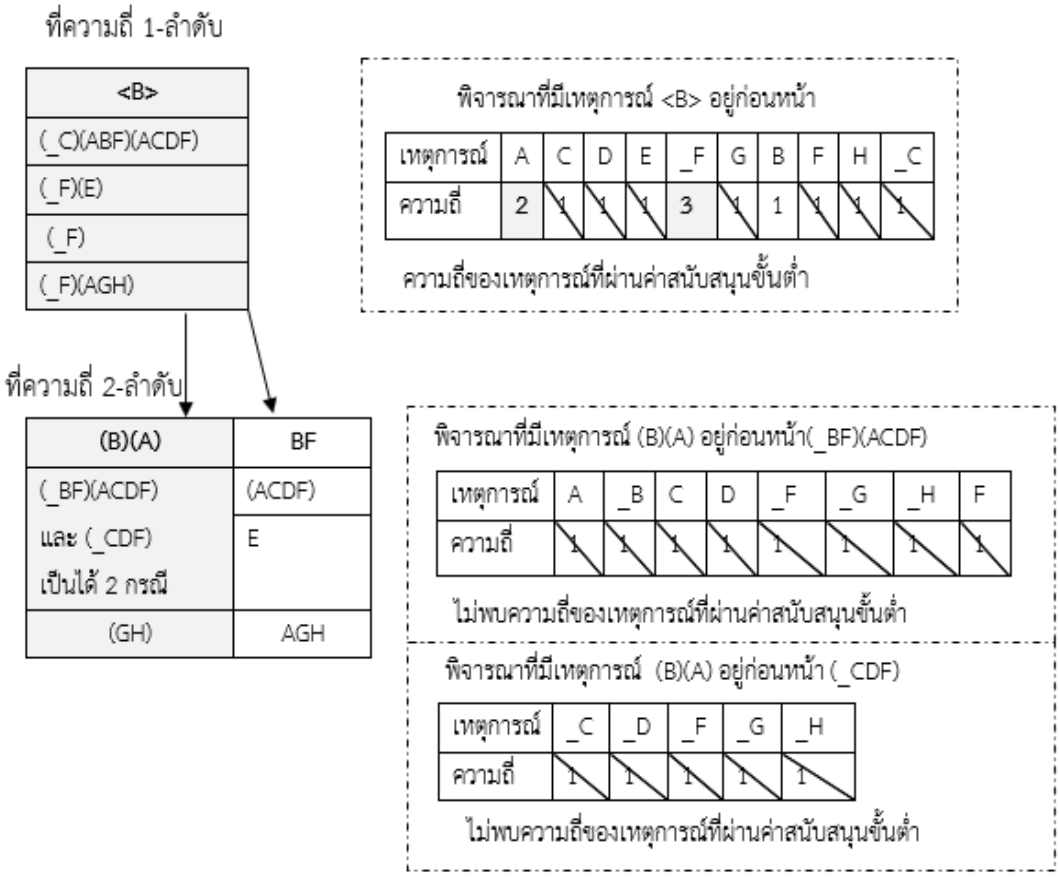
3.1 ชุดเหตุการณ์ย่อยของ A พิจารณาที่มีเหตุการณ์ <A> อยู่ก่อนหน้า



ภาพประกอบที่ 2 ชุดเหตุการณ์ย่อยของ <A>

จากฐานข้อมูลลำดับเมื่อพิจารณาพื้นที่ค้นหาที่มีเหตุการณ์ <A> อยู่ก่อนหน้า คือ เหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ _B, _F ดังนั้นจะได้เหตุการณ์ความถี่ (AB), (AF) เรียกว่า ลำดับเหตุการณ์ความถี่ 2-ลำดับ นำรายการที่ได้มาพิจารณาโดยเริ่มพิจารณาจากเหตุการณ์ที่ (AB) เกิดก่อนจากฐานข้อมูลลำดับ พบว่าเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ เหตุการณ์ _F ดังนั้น เหตุการณ์ความถี่ 3-ลำดับ คือ (ABF) เมื่อพิจารณาที่ (AF) ไม่พบเหตุการณ์ที่มีผ่านค่าสนับสนุนขั้นต่ำ ดังนั้นชุดเหตุการณ์ย่อยของ A คือ (AB), (AF), (ABF)

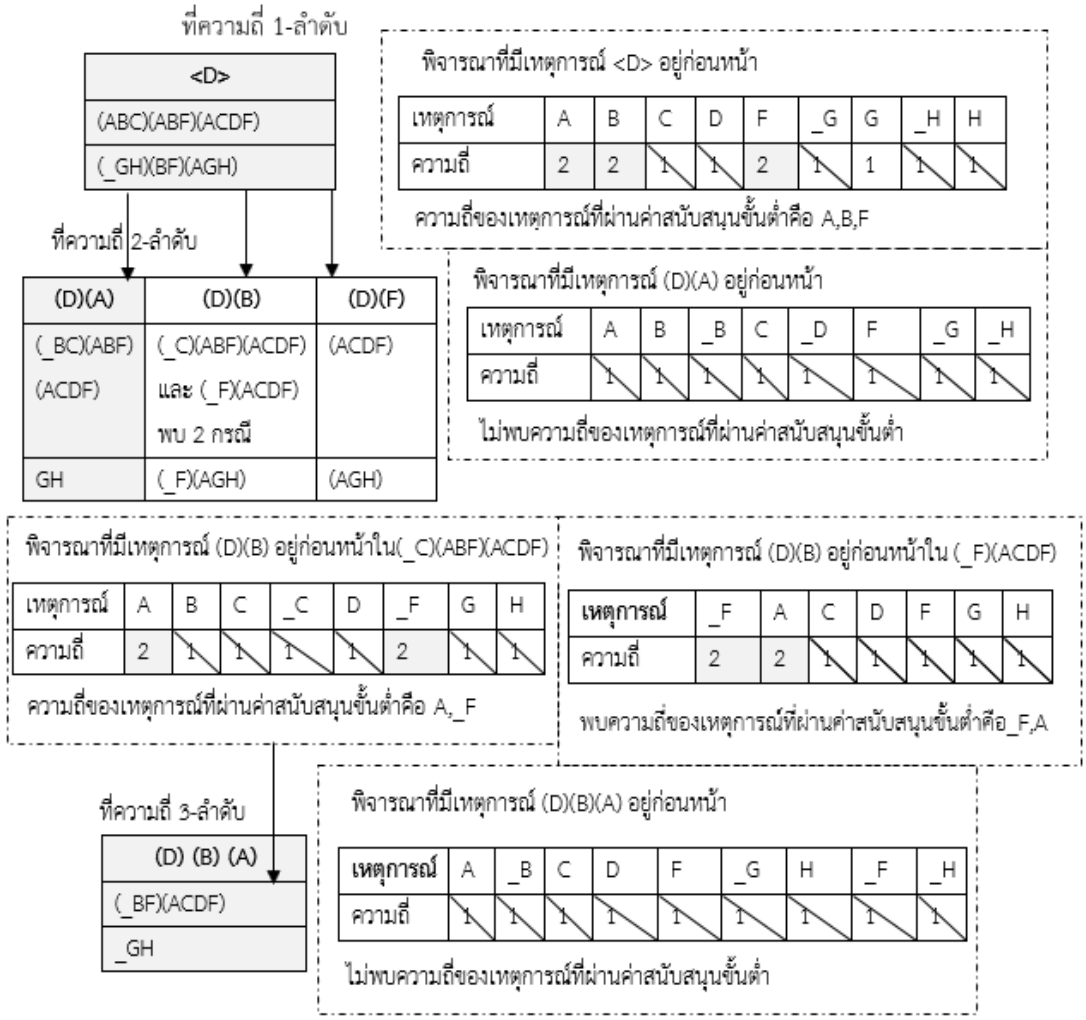
3.2 ชุดเหตุการณ์ย่อยของ B โดยพิจารณาที่มีเหตุการณ์ <B> อยู่ก่อนหน้า



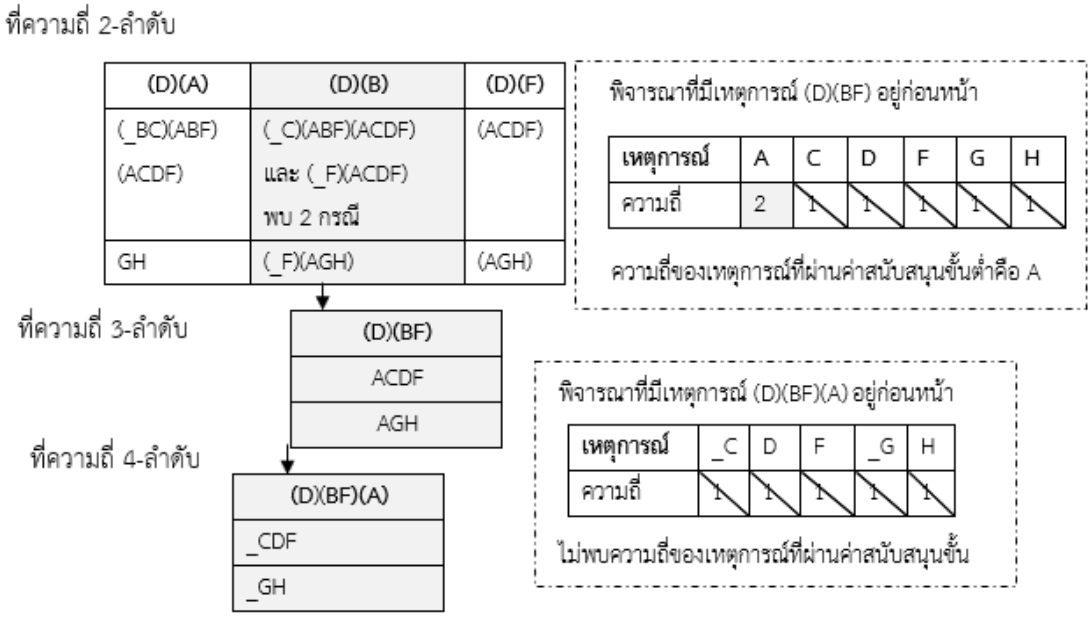
ภาพประกอบที่ 3 ชุดเหตุการณ์ย่อยของ <B>

จากฐานข้อมูลลำดับเมื่อพิจารณาพื้นที่ค้นที่มีเหตุการณ์ <B> อยู่ก่อนหน้า คือ เหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ A,_F ดังนั้น เหตุการณ์ (B)(A) ,(BF) เรียกว่า ความถี่ 2-ลำดับ นำรายการที่ได้มาพิจารณาโดยเริ่มพิจารณาจากเหตุการณ์ (B)(A) ก่อนจากฐานข้อมูลลำดับไม่พบว่าเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ พิจารณาจากเหตุการณ์ที่ BF เกิดก่อนจากฐานข้อมูลลำดับพบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ A ดังนั้นเหตุการณ์ (BF)(A) เรียกว่าความถี่ 3-ลำดับ นำรายการที่ได้มาพิจารณาไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ ดังนั้นชุดเหตุการณ์ย่อยของ B คือ (B)(A),(BF),(BF)(A)

3.3 ชุดเหตุการณ์ย่อยของ D พิจารณาที่มีเหตุการณ์ <D> อยู่ก่อนหน้า



ภาพประกอบที่ 4 ชุดเหตุการณ์ย่อยของ <D>



ภาพประกอบที่ 5 ชุดเหตุการณ์ย่อยของ <D> (ต่อ)

ภาพประกอบที่ 6 ชุดเหตุการณ์ย่อยของ <D> (ต่อ)

จากฐานข้อมูลลำดับเมื่อพิจารณาพื้นที่ค้นหาที่มีเหตุการณ์ <D> อยู่ก่อนหน้า มีเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ A, B, F ดังนั้นจะได้ลำเหตุการณ์ความถี่ คือ (D)(A),(D)(B), (D)(F) เรียกว่า ความถี่ 2-ลำดับ นำรายการที่ได้มาพิจารณาโดยพิจารณาจากเหตุการณ์ที่ (D)(A) เกิดก่อนจากฐานข้อมูลลำดับไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ

เมื่อพิจารณาจากเหตุการณ์ที่ (D)(B) เกิดก่อนจากฐานข้อมูลลำดับแบ่งได้เป็น 2 กรณี เพราะมีลำดับข้อมูลที่พบ (D)(B) เกิดก่อน คือ

กรณีที่ 1 คือ เหตุการณ์ (_C)(ABF)(ACDF) พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำคือ A,_F ดังนั้นเหตุการณ์ (D)(B)A,(D)(BF) เรียกว่า ความถี่ 3-ลำดับ นำรายการที่ได้มาพิจารณาคือ (D)(B)(A) ไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ

กรณีที่ 2 คือเหตุการณ์ (F)(ACDF) พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำคือ _F,A ดังนั้นเหตุการณ์ (D)(BF),(D)(BF)(A) เรียกว่าความถี่ 4-ลำดับ พิจารณา (D)(BF)(A) เกิดก่อนจากฐานข้อมูลลำดับไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ

เมื่อพิจารณาจากเหตุการณ์ที่ (D)(F) พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ A ดังนั้น (D)(F)(A) คือ ความถี่ 3-ลำดับ นำ (D)(F)(A) ไปพิจารณา ไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ ดังนั้นชุดเหตุการณ์ย่อยของ D คือ (D)(A), (D)(B)(A), (D)(BF), (D)(BF), (D)(F)(A)

3.4 ชุดเหตุการณ์ย่อยของ F พิจารณาที่มีเหตุการณ์ <F> อยู่ก่อนหน้า

ภาพประกอบที่ 7 ชุดเหตุการณ์ย่อยของ <F>

จากฐานข้อมูลลำดับเมื่อพิจารณาพื้นที่ค้นหาที่มีเหตุการณ์ <F> อยู่ก่อนหน้า จะได้เหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ A ดังนั้นจะได้ลำดับเหตุการณ์ความถี่ (F)(A) เรียกว่า ความถี่

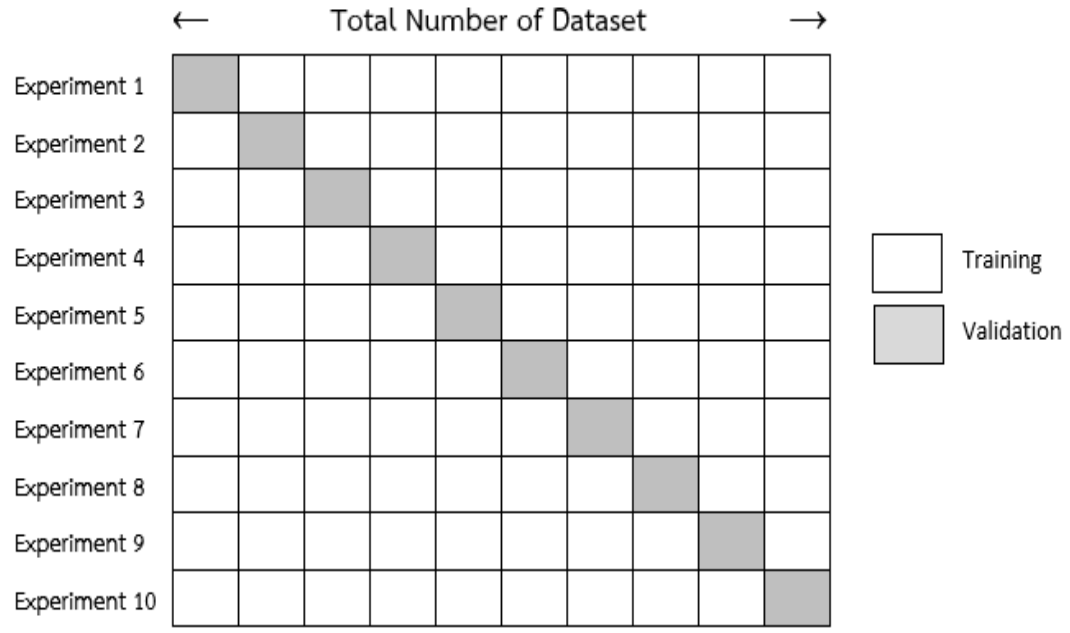
2-ลำดับ นำรายการที่ได้มาพิจารณาโดยเริ่มพิจารณาจากเหตุการณ์ที่ (F)(A) เกิดก่อน จากฐานข้อมูลลำดับไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ ดังนั้นชุดเหตุการณ์ย่อยของ F คือ (F)(A)

ลำดับเหตุการณ์ที่เกิดขึ้นในฐานข้อมูลนี้ทั้งหมดคือ (AB), (AF), (ABF), (B)(A), (BF), (BF)(A), (D)(A), (D)(B)(A), (D)(BF), (D)(BF)(A), (D)(F)(A), (F)(A)

2.4 การประเมิน

2.4.1 แบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพของโมเดล

วิธี Cross-validation Test เป็นที่นิยมใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน โดยจะแสดงด้วยค่า k เช่น 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่าๆกัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำซ้ำๆไปเช่นนี้จนครบจำนวนที่แบ่งไว้เช่น การทดสอบด้วยวิธี 10-fold cross-validation



ภาพประกอบที่ 8 ตัวอย่างการแบ่งข้อมูลแบบ 10-fold cross-validation

จากภาพประกอบที่ 8 แบ่งข้อมูลเทรนนิงออกเป็น 10 ส่วนที่มีจำนวนเท่ากัน หลังจากนั้นทำการทดสอบประสิทธิภาพของโมเดล 10 ครั้ง ดังนี้

- รอบที่ 1 ใช้ข้อมูลส่วนที่ 2,3,4,5,6,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 1
- รอบที่ 2 ใช้ข้อมูลส่วนที่ 1,3,4,5,6,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 2
- รอบที่ 3 ใช้ข้อมูลส่วนที่ 1,2,4,5,6,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 3
- รอบที่ 4 ใช้ข้อมูลส่วนที่ 1,2,3,5,6,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 4
- รอบที่ 5 ใช้ข้อมูลส่วนที่ 1,2,3,4,6,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 5
- รอบที่ 6 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,7,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 6
- รอบที่ 7 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,8,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 7
- รอบที่ 8 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,7,9 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 8
- รอบที่ 9 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,7,8 และ 10 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 9
- รอบที่ 10 ใช้ข้อมูลส่วนที่ 1,2,3,4,5,6,7,8 และ 9 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 10

จะเห็นได้ว่าข้อมูลทุกชุดจะได้เป็นตัวทดสอบประสิทธิภาพของโมเดล โดยในการทดสอบประสิทธิภาพของโมเดลแต่ละรอบจะได้จำนวน TP, TN, FP, FN ใส่ลงไปในตาราง confusion matrix และบวกเพิ่มเข้าไป สุดท้ายจะได้ตาราง confusion matrix ที่เป็นค่ารวมทั้งหมด

2.4.2 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

นำโมเดลไปใช้งานจริงจำเป็นจะต้องทราบประสิทธิภาพของโมเดลก่อน โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยต่างๆ คือ

1. ค่าความแม่นยำ (Precision) เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส
2. ค่าระลึก (Recall) เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส
3. ค่าอัตราการรู้จำ (F-measure) เป็นการวัดค่าความแม่นยำและค่าระลึกพร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส
4. ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส

โดยค่าความแม่นยำ ค่าระลึก ค่าอัตราการรู้จำและค่าความถูกต้องสามารถพิจารณาได้จากตาราง Confusion matrix ซึ่งเป็นตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนคลาส เช่น ในตารางที่ 7 มีคลาส(Class) คำตอบอยู่ 2 ค่า คือ yes และ no ฉะนั้นตาราง confusion matrix นี้จะสร้างได้เป็นตารางขนาด 2x2 โดยข้อมูลด้านคอลัมน์คือ คลาสที่อยู่ในข้อมูลเทรนนิงดาต้า (actual) และข้อมูลในแนวแถว คือ คลาสที่โมเดลทำนายมาได้ (predicted)

ตารางที่ 7 confusion matrix

predicted/actual	yes	no
yes	TP	FP
no	FN	TN

• True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่

• True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่

• False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาสซึ่งกำลังสนใจอยู่

• False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่

แสดงข้อมูลคลาส จากเทรนนิงดาต้า 10 ตัวแรกและค่าที่ทำนายได้ดังตารางที่ 8

ตารางที่ 8 ข้อมูลคลาส

No.	Predicted	Actual
1	*no*	*no*
2	*no*	*no*
3	no	yes
4	**yes**	**yes**
5	no	yes
6	yes	no
7	**yes**	**yes**
8	*no*	*no*
9	no	yes
10	**yes**	**yes**

ตัวอย่างข้อมูลที่อยู่ในเทรนนิงดาต้าและที่ทำนายออกมาโดยที่กำลังพิจารณาคลาส yes ดังนั้นจะสรุปได้ว่า

• True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส yes มีจำนวน 3 ตัว (แถวที่เป็นตัวหนา คือ แถวที่ 4, 7 และ 10)

• True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส no มีจำนวน 3 ตัว (แถวที่เป็นตัวเอียง คือ แถวที่ 1, 2 และ 8)

• False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส yes มีจำนวน 1 ตัว (แถวที่ขีดเส้นใต้ คือ แถวที่ 6)

• False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส no มีจำนวน 3 ตัว (แถวที่ตัวอักษรปกติ คือ แถวที่ 3, 5 และ 9)

ดังนั้นจึงสร้างตาราง Confusion matrix ได้ดังตารางที่ 9

ตารางที่ 9 แสดงตาราง confusion matrix ของข้อมูล

predicted / actual	yes	no
Yes	3	1
No	3	3

ค่าความแม่นยำเป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาสคือ Play=yes และ Play=no ดังแสดงในสมการ 2.4 และ 2.5

$$Precision_{yes} = \frac{TP}{TP+FP} \quad (2.4)$$

$$Precision_{no} = \frac{TN}{TN+FN} \quad (2.5)$$

ดังนั้นค่าความแม่นยำของคลาส yes คือ Precision (yes) = 3/4 = 75% ค่าความแม่นยำของคลาส no คือ Precision (no) = 3/6 = 50%

ค่าระลึกเป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาสดังสมการ 2.6 และ 2.7

$$Recall_{yes} = \frac{TP}{TP+FN} \quad (2.6)$$

$$Recall_{no} = \frac{TN}{TN+FP} \quad (2.7)$$

ดังนั้นค่าระลึกของคลาส yes คือ Recall (yes) = 3/6 = 50% ค่าระลึกของคลาส no คือ Recall (no) = 3/4 = 75%

ค่าอัตราการรู้จำเป็นการวัดค่าความแม่นยำและค่าระลึกพร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาสดังสมการ 2.8 และ 2.9

$$F\text{-}measure_{yes} = \frac{2 \times Precision_{yes} \times Recall_{yes}}{Precision_{yes} + Recall_{yes}} \quad (2.8)$$

$$F\text{-}measure_{no} = \frac{2 \times Precision_{no} \times Recall_{no}}{Precision_{no} + Recall_{no}} \quad (2.9)$$

F-measure (yes) = 2 x 75% x 50% / (75% + 50%) = 60%

F-measure (no) = 2 x 50% x 75% / (50% + 75%) = 60%

ค่าความถูกต้องเป็นการวัดความถูกต้องของโมเดล โดยพิจารณาค่าที่ทำนายถูกของทุกคลาสรวมกันดังสมการ 2.10

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.10)$$

ค่าความถูกต้องของโมเดลหรือ Accuracy = $\frac{6}{10}$ = 60%

2.5 งานวิจัยที่เกี่ยวข้อง

ศุกร์ใจ วุฒิกิจโกศล [3] ได้นำข้อมูลผู้ป่วยมาทำเหมืองข้อมูล โดยใช้ข้อมูลผู้ป่วยที่รับบริการทางกายภาพบำบัดในโรงพยาบาลพระนั่งเกล้าที่รักษาหายแล้วในช่วงปี พ.ศ. 2548-2550 โดยเป็นผู้ป่วยโรคข้อไหล่ติดที่ไม่มีโรคประจำตัวและไม่ได้เกิดจากอุบัติเหตุ โดยนำข้อมูลมาหาปัจจัยที่ส่งผลต่อการรักษาและนำมาสร้างกฎความสัมพันธ์ที่น่าสนใจที่เป็นแนวทางในการช่วยสนับสนุนการตัดสินใจในการเลือกเทคนิคเพื่อรักษาผู้ป่วยให้แก่นักกายภาพบำบัด โดยมีกระบวนการเตรียมข้อมูลและหาปัจจัยที่มีผลต่อการรักษาโดยใช้เทคนิค Cluster และ Association Rule ด้วยปัจจัยดังนี้ คือ ระยะเวลาที่เป็น อายุ องศาการยกแขน ระดับการไขว้หลัง และระดับความเจ็บปวด โดยกำหนดค่าในการหากฎที่น่าสนใจคือ สนับสนุนขั้นต่ำไม่น้อยกว่าร้อยละ 20 และค่าความเชื่อมั่นขั้นต่ำไม่น้อยกว่าร้อยละ 90 ผลการศึกษา แสดงให้เห็นว่า องศาการยกแขนระหว่าง 90-120 ได้กฎความสัมพันธ์การรักษาด้วยแผ่นประคบความร้อนบริเวณคอความร้อนคลื่นเหนือเสียงบริเวณไหล่ด้านหน้า ความร้อนคลื่นเสียงบริเวณกล้ามเนื้อ Infraspinatus และกระตุ้นไฟฟ้าแบบ Surge บริเวณไหล่ด้านหน้า-หลัง จะใช้จำนวนครั้งในการรักษาช่วง 19-24 ครั้ง ด้วยค่าสนับสนุนของกฎคิดเป็น 43.48 % และค่าความเชื่อมั่น 100%

รักถิ่น เหลาหา [4] ศึกษาการพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอดโดยใช้ทฤษฎีของการทำเหมืองข้อมูลเพื่อสร้างระบบที่จัดกลุ่มของผู้ป่วยและพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด โดยข้อมูลที่ใช้ในการทดลองเป็นข้อมูลผู้ป่วยโรงพยาบาลมหาสารคามในเดือนสิงหาคม–ธันวาคม

2552 จำนวนทั้งหมด 2,215 ราย แบ่งผู้ป่วยเป็นสองกลุ่มคือผู้ป่วยที่เป็นโรคมะเร็งปอดจำนวน 118 คน ไม่เป็นมะเร็งปอด 2,097 คน โดยใช้ปัจจัยเฉพาะด้าน ได้แก่ อายุเพศ ประวัติการสูบบุหรี่ ได้รับสารแอสเบสตอส ได้รับแร่เรดอน พันธุกรรม มลภาวะทางอากาศ ประวัติการดื่มสุรา เพื่อคัดเลือกตัวแบบที่เหมาะสมที่สุดในการวิเคราะห์ปัจจัยเสี่ยงของโรคมะเร็งปอด โดยนำค่าปัจจัยเสี่ยงมาวิเคราะห์และพยากรณ์ผู้ป่วยด้วยต้นไม้ตัดสินใจแบบ C4.5 พบว่าปัจจัยที่มีค่าความเสี่ยงที่ทำให้เป็นมะเร็งปอดมากที่สุดคือ ประวัติด้านพันธุกรรม โดยมีค่าความเสี่ยงเป็น 34.59 เท่าของคนที่ไม่มีประวัติด้านพันธุกรรม รองลงมาคือปัจจัยเสี่ยง ประวัติการสูบบุหรี่ ประวัติการดื่มสุราและอายุ ตามลำดับ ผลการพยากรณ์สามารถพยากรณ์ค่าที่แม่นยำที่สุดด้วยการแบ่งข้อมูลแบบ 70:30 เปอร์เซ็นมีค่าเท่ากับ 96.83% ให้ค่าความแม่นยำที่ 79.60% ค่าความระลึก 50% และค่าความถูกต้องของการจำแนกประเภทเฉลี่ยมีค่าร้อยละ 96.83%

อังคณา พิจารโชติ [5] ได้วิเคราะห์หาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆ ที่ก่อให้เกิดโรคเบาหวาน เพื่อพัฒนาระบบสนับสนุนการตัดสินใจโดยใช้เทคนิคดาต้าไมน์นิ่ง โดยได้ข้อมูลจากฐานข้อมูลโปรแกรมระบบงานศูนย์สุขภาพชุมชนที่ใช้งานในสถานีอนามัยในเขตพื้นที่จังหวัดชัยภูมิ เพื่อนำข้อมูลความสัมพันธ์ที่ได้จากปัจจัยเสี่ยงมาใช้วางแผนการป้องกันโรคเบาหวาน ระบบที่พัฒนาประกอบด้วย 3 ส่วนคือ ส่วนแรก คลังข้อมูลที่เก็บปัจจัยเสี่ยงจากการคัดกรองกลุ่มเสี่ยง ส่วนที่สองเหมืองข้อมูลที่เป็นตัวค้นหาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆ ส่วนที่สาม การแสดงข้อมูลรายงาน โดยการวิเคราะห์ความสัมพันธ์ใช้ Weka ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลแบบ Association rule โดยเลือกกฎหรือความสัมพันธ์ 10 กฎ ที่มีความถูกต้องหรือค่าความเชื่อมั่นในระดับที่ดี พบว่าอัตราการเสี่ยงมากที่สุดคือ ประวัติการเป็นเบาหวานของครัวเรือนที่พบคู่กับดัชนีมวลกาย มีค่าความเชื่อมั่นที่ 94% นำไปพัฒนาในรูปแบบเว็บแอพพลิเคชั่นเรียกใช้งานผลการวิจัยได้ค่าสนับสนุนที่ 50% และค่าความเชื่อมั่นที่ 30%

Mullins และคณะ[10] ใช้นำเทคนิคเหมืองข้อมูลและคลังข้อมูลการรักษาผู้ป่วยมาประยุกต์ใช้กับข้อมูลผู้ป่วย 667,000 คน โดยเก็บรวบรวมข้อมูลไว้ในระบบฐานข้อมูลโรงพยาบาลในสถาบันวิทยาศาสตร์สุขภาพ มหาวิทยาลัยเวอร์จิเนีย ในปี ค.ศ. 1993-2005 ประกอบด้วยข้อมูลผู้ป่วยในและผู้ป่วยนอก เพื่อค้นหารูปแบบความสัมพันธ์และการจำแนกดูความสัมพันธ์ของข้อมูลผู้ป่วย รวมถึงการพยากรณ์ด้วยเทคนิควิธีเหมืองข้อมูล พบว่าการจำแนกแนวโน้มข้อมูลด้วย CliniMiner สามารถจำแนกกลุ่มข้อมูลแยกเป็น 3 กลุ่ม คือ established , less well know และ unknown ซึ่งทั้ง 3 กลุ่มนำไปสร้างกฎการจำแนก Prediction โดย กลุ่มแรก established ได้ 73 กฎ จากทั้งหมด 120 กฎคิดเป็น 61% กลุ่มที่สอง less well known 18 กฎ คิดเป็น 15% และกลุ่มที่สาม unknown 29 กฎ คิดเป็น 24% ของข้อมูลที่นำมาศึกษา จากการนำเทคนิคเหมืองข้อมูลมาใช้กับข้อมูลผู้ป่วยนั้น พบว่าการค้นหารูปแบบมีความอย่างเหมาะสมและมีประสิทธิภาพที่จะค้นหาข้อมูล

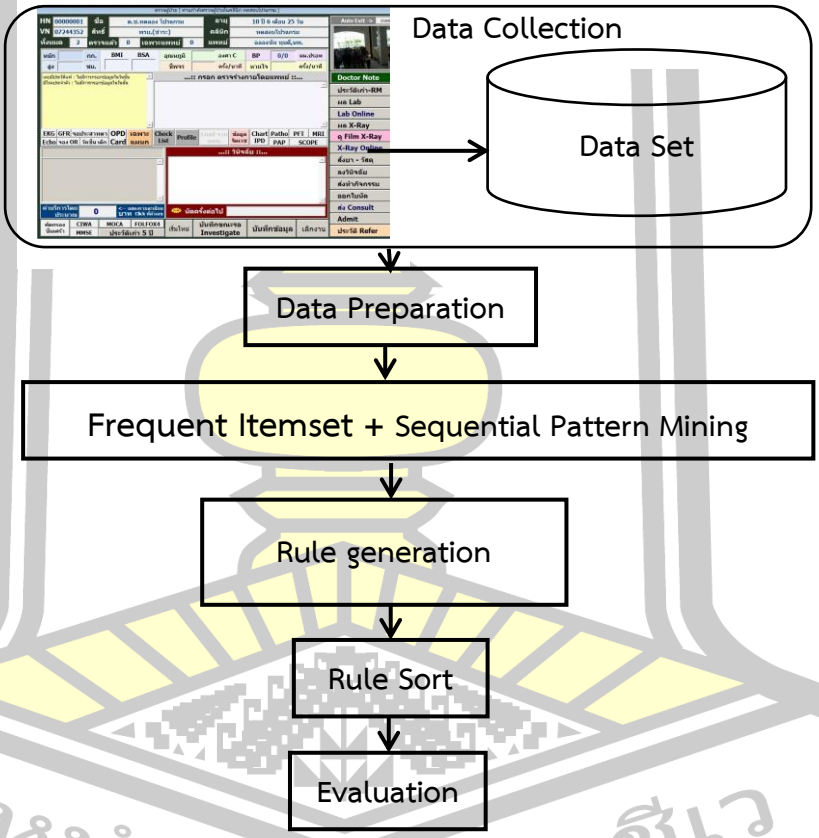
ผู้ป่วยในฐานข้อมูลและขยายขีดความสามารถในการวิจัย โดยระบุความสัมพันธ์ของโรคทางคลินิกที่อาจเกิดขึ้นใหม่ได้ โดยปราศจากความอคติและได้ความเข้าใจทางชีวการแพทย์ จากการศึกษาตัววัดการพยากรณ์ผิด (false positives) และตัววัดการพยากรณ์ถูก (false negatives)

Richard และคณะ [11] ศึกษาตัวชี้วัดการตายก่อนวัยอันควรของผู้ป่วยโรคเบาหวานระหว่างผู้ป่วยใหม่กับผู้ป่วยที่ใกล้ตาย ด้วยเทคนิคเหมืองข้อมูลมาประยุกต์ ซึ่งเทคนิคนำมาที่ใช้คือวิธีกฎความสัมพันธ์และการแบ่งกลุ่มข้อมูล (Classification) โดยใช้วิธีการจำแนกกลุ่ม (Discriminant Analysis) โดยเก็บรวบรวมข้อมูลไว้ในฐานข้อมูลโรงพยาบาลเซ็นโทมัส ตั้งแต่ปี ค.ศ. 1973-2001 ศึกษาพบว่าการนำกฎความสัมพันธ์ (Association Rule) มาใช้สามารถแยกการตายก่อนวัยอันควรของผู้ป่วยโรคเบาหวานด้วยตัวแปรของอายุผู้ป่วยโดยแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มผู้ป่วยอายุตั้งแต่ 60 ปีขึ้นไป จำนวน 275 คน คิดเป็น 6.9% ของข้อมูลทั้งหมด 3,971 คน และกลุ่มผู้ป่วยอายุน้อยกว่า 60 ปี จำนวน 3,696 คน คิดเป็น 93.1% โดยทำการตรวจสอบความถูกต้องของตัวแบบด้วยวิธีทำซ้ำ การกำจัดข้อมูลที่ไม่จำเป็นออกและปรับปรุงข้อมูลด้วยค่า Specified ผลที่ได้คือสามารถขจัดความมีอคติได้มากขึ้น 10% อย่างมีนัยสำคัญที่ระดับ 0.05

จากงานวิจัยต่างๆ เมื่อนำเทคนิคเหมืองข้อมูลมาประยุกต์ใช้ เพื่อช่วยในการวิเคราะห์หาความสัมพันธ์ของปัจจัยต่างๆ พบว่าข้อมูลความสัมพันธ์ของปัจจัยที่ได้เป็นสิ่งที่ช่วยในการรักษาหรือสามารถนำไปใช้วางแผนป้องกันโรคได้ ซึ่งพบว่ามีหลากหลายวิธีในการหาความสัมพันธ์และการจำแนกข้อมูลที่เกี่ยวข้องในทางการแพทย์ รวมถึงการวัดประสิทธิภาพการทำงานของแต่ละเทคนิควิธีที่ใช้ แต่งานวิจัยส่วนมากนำปัจจัยที่เกี่ยวข้องกับผู้ป่วยมาเป็นพื้นฐานในการจำแนกกลุ่มผู้ป่วยหรือพยากรณ์การเกิดโรค แต่ไม่พบว่านำข้อมูลการวินิจฉัยของแพทย์มาร่วมด้วย ซึ่งข้อมูลส่วนนี้เป็นข้อมูลที่สำคัญและมีการเกิดอย่างเป็นลำดับเหตุการณ์ที่มีความสำคัญ

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้นำเสนอวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์ เพื่อใช้กับการจำแนกโรคหลอดเลือดสมองและหาความสัมพันธ์ของปัจจัยกับลำดับเกิดโรคที่นำไปสู่โรคหลอดเลือดสมองของผู้ป่วยสูงอายุที่มารับการรักษาในโรงพยาบาลมหาสารคาม โดยเริ่มจากการรวบรวมข้อมูล การเตรียมข้อมูล การสืบค้นเซตรายการความถี่ร่วมกับลำดับเหตุการณ์ความถี่ การสร้างกฎ การเรียงกฎและการประเมินผล โดยแสดงภาพรวมขั้นตอนวิธีการดำเนินงานวิจัยดังภาพประกอบที่ 9



ภาพประกอบที่ 9 ขั้นตอนการทำวิจัย

3.1 การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้ในการวิจัยคือประวัติสุขภาพของผู้ป่วยที่เข้ารับบริการในโรงพยาบาลผ่านการบันทึกเวชระเบียน ซึ่งข้อมูลการรักษาถูกบันทึกโดยบุคลากรทางการแพทย์ สามารถนำไปใช้ใน

การศึกษาวิจัยทางวิทยาศาสตร์และด้านอื่นๆ ได้โดยการนำเทคโนโลยีและความรู้ทางวิทยาการคอมพิวเตอร์เข้ามาร่วมเพื่อช่วยวิเคราะห์ปัญหาสุขภาพของประชากรได้

โดยงานวิจัยนี้ได้ข้อมูลมาจากฐานข้อมูลผู้ป่วยที่เข้ารับบริการที่โรงพยาบาลมหาสารคาม โดยประชากรกลุ่มตัวอย่างคือผู้ป่วยที่มีอายุ 60 ปีขึ้นไปที่มีปัจจัยเสี่ยงทำให้เกิดโรคหลอดเลือดสมอง จำนวน 1,000 ราย โดยแบ่งข้อมูลออกเป็น 2 ส่วนคือผู้ป่วยที่เป็นโรคหลอดเลือดสมอง 500 ราย และผู้ป่วยที่ไม่เป็นโรคหลอดเลือดสมอง 500 ราย โดยเป็นผู้ป่วยที่มารับบริการในโรงพยาบาล เริ่มเก็บข้อมูลตั้งแต่วันที่ 1 มกราคม พ.ศ. 2555 ถึง 31 ธันวาคม พ.ศ. 2559 ซึ่งข้อมูลที่ใช้ในงานวิจัยนี้ประกอบไปด้วย เพศ สถานะภาพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย อาชีพ ระดับความดันโลหิต(blood pressure:BP) ระดับคอเลสเตอรอล และลำดับการเกิดโรคที่ถูกวินิจฉัยในการมารับบริการทุกครั้ง

3.2 การเตรียมข้อมูล

ในขั้นตอนนี้จะทำการแปลงข้อมูลที่รวบรวมได้ให้อยู่ในรูปแบบของทรานแซคชั่นโดยหนึ่งทรานแซคชั่นหมายถึงข้อมูลของผู้ป่วยหนึ่งคน ซึ่งจะประกอบไปด้วย 3 ส่วนดังตัวอย่างในตารางที่ 3.1 คือ

1) เซตรายการ (itemsets) หรือปัจจัยเสี่ยงที่ทำให้เกิดโรคหลอดเลือดสมองและทำให้ไม่เกิดโรคหลอดเลือดสมอง ซึ่งเป็นข้อมูลที่ไม่เป็นลำดับประกอบไปด้วย เพศ สถานะภาพ อาชีพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย ระดับความดันโลหิต ระดับคอเลสเตอรอล

2) ลำดับการเกิดโรค คือ ลำดับเหตุการณ์ (Sequence) การวินิจฉัยโรคของผู้ป่วยโดยแพทย์ผู้รักษาในการมารับบริการ เป็นข้อมูลแบบลำดับมีเวลาก่อนหลังเข้ามาเกี่ยวข้อง ซึ่งบ่งบอกลำดับการเกิดของโรค เช่น ลำดับการเกิดโรค <(A)(B,C,D)(E)> หมายความว่าผู้ป่วยมารับบริการครั้งที่หนึ่งถูกระบุว่าเป็นโรค A มารับบริการครั้งที่สองถูกระบุว่าเป็นโรค B, C และ D ซึ่งไม่ได้พิจารณาลำดับการเกิดของโรคแต่ระบุว่าพบโรคดังกล่าวในการมารับบริการครั้งเดียวกัน และการมารับบริการครั้งที่สามถูกระบุว่าเป็นโรค E เป็นต้น

3) คลาส (Class) มี 2 คลาส คือ เป็นโรคหลอดเลือดสมองและไม่เป็นโรคหลอดเลือดสมอง

เพื่อให้ข้อมูลสามารถนำเข้าวิธีการวิเคราะห์จำเป็นจะต้องทำการเตรียมข้อมูลก่อน โดยการเตรียมข้อมูลในงานวิจัยนี้ประกอบด้วยการทำงานย่อย 2 กระบวนการคือ การทำความสะอาดข้อมูลและการแปลงข้อมูลจะอธิบายวิธีการดำเนินการของแต่ละขั้นตอนดังนี้

ตารางที่ 10 ตัวอย่างข้อมูล

ลำดับ	เพศ	สถานะภาพ	สูบบุหรี่	ดื่มสุรา	ออกกำลังกาย	อาชีพ	BP	คอเลสเตอรอล	ลำดับการเกิดโรค	คลาส
1	หญิง	สมรส	ผู้สูงอายุไม่ได้ทำงาน	ไม่สูบ	ไม่ดื่ม	ไม่ออก	สูง	สูง	<(N40) (K74,I351) (Z532) (I694)>	เป็น
2	ชาย	สมรส	ทำนา	สูบ	ดื่ม	ไม่ออก	ปกติ	ปกติ	<(A162) (I35,K74,I05) (J44)>	ไม่เป็น
3	ชาย	โสด	ผู้สูงอายุไม่ได้ทำงาน	ไม่สูบ	ไม่ดื่ม	ไม่ออก	สูง	สูง	<(I10,D291, A162) (I694)>	เป็น
4	ชาย	สมรส	ค้าขาย	สูบ	ดื่ม	ไม่ออก	ปกติ	สูง	<(I10)(A162, M10)(J44)>	ไม่เป็น
5	หญิง	สมรส	ทำนา	ไม่สูบ	ไม่ดื่ม	ไม่ออก	ปกติ	ปกติ	<(K52,E11) (G99,D64)>	ไม่เป็น
6	หญิง	สมรส	ทำนา	สูบ	ดื่ม	ออก	สูง	ปกติ	< (A162) (E117) Z532) (J44)>	เป็น
7	ชาย	โสด	ไม่ได้ทำงาน				สูง	ปกติ	<(I10,E78, I51)(Z53) (I07,I25,I37) (S09,S00)>	ไม่เป็น
8	ชาย	โสด	ไม่ได้ทำงาน	ไม่สูบ	ไม่ดื่ม	ไม่ออก	ปกติ	ปกติ		ไม่เป็น

3.2.1 การทำความสะอาดข้อมูล (Data Cleansing)

การทำความสะอาดข้อมูลเป็นกระบวนการตรวจสอบความไม่สมบูรณ์ ความไม่ถูกต้อง ความไม่สัมพันธ์กับข้อมูลอื่นๆ จึงต้องมีการแทนที่ การปรับปรุง หรือการลบข้อมูลที่ไม่ถูกต้องเหล่านี้ออกไป เพื่อให้เหลือเฉพาะข้อมูลที่จะนำไปวิเคราะห์ที่มีความถูกต้อง ครบถ้วนและมีคุณภาพ งานวิจัยนี้ทำความสะอาดข้อมูลโดยทำการตัดข้อมูลที่ไม่สมบูรณ์ออกไป จากตารางที่ 10 ในแถวที่ 7 ไม่พบข้อมูลการซักประวัติ การดื่มสุรา การสูบบุหรี่ การออกกำลังกาย จึงทำการตัดแถวที่ 7 ออกจากฐานข้อมูลที่จะนำไปใช้เพราะข้อมูลไม่สมบรูณ์และแถวที่ 8 ไม่มีข้อมูลลำดับการเกิดโรค จึงตัดออก

จากฐานข้อมูลที่จะนำไปใช้เพราะข้อมูลไม่สมบรูณ์ ข้อมูลที่เหลือในฐานข้อมูลคือข้อมูลที่ทำความสะอาดแล้วดังตารางที่ 11

ตารางที่ 11 ข้อมูลที่ทำความสะอาด

ลำดับ	เพศ	สถานะภาพ	สูบบุหรี่	ดื่มสุรา	ออกกำลังกาย	อาชีพ	BP	คอเลสเตอรอล	ลำดับการเกิดโรค	คลาส
1	หญิง	สมรส	ไม่สูบ	ไม่ดื่ม	ไม่ออก	ผู้สูงอายุไม่ได้ทำงาน	สูง	สูง	<(N40)(K74, I351)(Z532) (I694)>	เป็น
2	ชาย	สมรส	สูบ	ดื่ม	ไม่ออก	ทำนา	ปกติ	ปกติ	<(A162) (I35,K74,I05) (J44)>	ไม่เป็น
3	ชาย	โสด	ไม่สูบ	ไม่ดื่ม	ไม่ออก	ผู้สูงอายุไม่ได้ทำงาน	สูง	สูง	<(I10,D291 ,A162)(I694)>	เป็น
4	ชาย	สมรส	สูบ	ดื่ม	ไม่ออก	ค้าขาย	ปกติ	สูง	<(I10)(A162, M10)(J44)>	ไม่เป็น
5	หญิง	สมรส	ไม่สูบ	ไม่	ไม่ออก	ทำนา	ปกติ	ปกติ	<(K52,E11) (G99,D64)>	ไม่เป็น
6	หญิง	สมรส	สูบ	ดื่ม	ออก	ทำนา	สูง	ปกติ	<(A162)(E117) (Z532) (J44)>	เป็น

3.2.2 การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลเป็นขั้นตอนการแปลงจากกระบวนการข้างต้น ให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมเพื่อนำไปใช้กับขั้นตอนวิธีต่างๆ ในงานวิจัย ดังแสดงในตารางที่ 12-15

ตารางที่ 12 การแปลงค่าข้อมูลปัจจัยพื้นฐาน

ลำดับ	ปัจจัย	ความหมาย
1	เพศ	10=ชาย, 11=หญิง
2	สถานะภาพ	20=โสด, 21=สมรส
3	สูบบุหรี่	40=ไม่สูบบุหรี่, 41=เคยสูบแต่เลิกแล้ว, 42=สูบบุหรี่
4	ดื่มสุรา	50=ไม่ดื่มสุรา, 51=ดื่มสุราแต่เลิกแล้ว, 52=ดื่มสุรา

ตารางที่ 13 การแปลงค่าข้อมูลปัจจัยพื้นฐาน(ต่อ)

ลำดับ	ปัจจัย	ความหมาย
5	ออกกำลังกาย	60=ไม่ออกกำลังกาย, 61=ออกกำลังกาย
6	อาชีพ	700=ไม่ได้ทำงาน, 701=กรรมกร/ผู้ใช้แรงงาน, 702=กสิกรรม ,703=กำนัน, 704=เกษตรกรรม, 705=ข้าราชการการเมือง, 706=ข้าราชการบำนาญ, 707=คนขับรถรับจ้าง, 708=ครูอัตราจ้าง,709=นักท่องเที่ยว, 710=ค้าขาย, 711=ค้าปลีก, 712=ค้าส่ง, 713=เจ้าของกิจการ/ธุรกิจส่วนตัว, 714=ช่างซ่อมเครื่องไฟฟ้า ,715=ช่างตัดผม, 716=ช่างตัดเย็บเสื้อผ้า, 717=ตำรวจ,718= ทนายความ,719=ทหารบก, 720=ทหารอากาศ,721=ทำนา, 722=ทำสวน,723=นักการ-ภารโรง,724=นักบวช, 725=นักเรียน นักศึกษา, 726=นางแบบ,727=ผู้ช่วยผู้ใหญ่บ้าน, 728=ผู้สอนศาสนา, 729=ผู้ใหญ่บ้าน,730=ครูเอกชน, 731=พนักงานรัฐวิสาหกิจ, 732=พนักงานบริษัท,733=พนักงานโฆษณา, 734=พยาบาล RN, 735=แพทย์, 736=พ่อบ้าน, 737=รับจ้างทั่วไป, 738=ข้าราชการพลเรือน, 739=ลูกจ้างชั่วคราว,740=ลูกจ้างประจำ, 741=วิศวกร, 742=แม่บ้าน, 743=พ่อบ้าน,744=อาจารย์มหาวิทยาลัย, 745=ข้าราชการครู, 746=นักโทษ,747=สมณะ, 748=คนพิการ, 749=ผู้สูงอายุไม่ได้ทำงาน,750=ข้าราชการส่วนท้องถิ่น, 751=หมอดู, 752=หาบเร่, 753=นักประพันธ์, 754=ไม่ทราบ
7	ระดับความดันโลหิต	80=ปกติ, 81=สูง
8	ระดับคอเลสเตอรอล	90=ปกติ, 91=สูง

จากตารางที่ 12-13 แสดงการแทนค่าข้อมูลปัจจัยพื้นฐาน ซึ่งประกอบไปด้วย เพศ สถานะภาพ อาชีพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย ระดับความดันโลหิต ระดับคอเลสเตอรอล ด้วยสัญลักษณ์

การแปลงลำดับการเกิดโรคหรือการวินิจฉัยโรคของผู้ป่วยโดยแพทย์ เนื่องจากโรคมีจำนวนหลากหลาย ในงานวิจัยนี้จึงจัดโรคที่เกิดขึ้นเป็นกลุ่มโรคตามมาตรฐาน ICD-10 ซึ่งมีจำนวน 228 กลุ่มโรคโดยเริ่มจาก A00-A09(โรคติดเชื้อที่ลำไส้) , A15-A19(วัณโรค),..., Z00-Z99(ปัจจัยที่มีผลต่อสถานะสุขภาพและการรับบริการสุขภาพ) ตามลำดับ และแทนค่าด้วยสัญลักษณ์ เช่น 8000,8001,8002,...,8227 ตามลำดับ ดังตัวอย่างในตารางที่ 14

ตารางที่ 14 ตัวอย่างการแปลงโรค ICD-10[12]

สัญลักษณ์แทนค่า	**ความหมายกลุ่มโรค**
8000	A00-A09(โรคติดเชื้อที่ลำไส้)
8001	A15-A19(วัณโรค)
...	...
8226	Y90-Y98(ปัจจัยเสริมเกี่ยวกับสาเหตุการเจ็บป่วยและการตายที่
8227	จำแนกไว้ที่อื่น)
	Z00-Z99(ปัจจัยที่มีผลต่อสถานะสุขภาพและการรับบริการสุขภาพ)

หมายเหตุ ตารางอ้างอิงทั้งหมดแสดงในท้ายเล่มวิทยานิพนธ์

ตารางที่ 15 แทนค่าคลาส

ค่า	**แทนค่า**
เป็นโรคหลอดเลือดสมอง	Y
ไม่เป็นโรคหลอดเลือดสมอง	N

แทนค่าการเป็นโรคหลอดเลือดสมองด้วยคลาส Y และแทนค่าการไม่เป็นโรคหลอดเลือดสมองด้วยคลาส N ดังตารางที่ 15

เมื่อแปลงค่าข้อมูลจากตารางที่ 11 ซึ่งประกอบด้วย ปัจจัย ลำดับการเกิดโรค และคลาส ในแต่ละแถวให้อยู่ในรูปแบบของสัญลักษณ์สามารถแสดงได้ดังตารางที่ 16

ตารางที่ 16 แทนค่าด้วยสัญลักษณ์

ลำดับ	เพศ	สถานะภาพ	สูบบุหรี่	ดื่มสุรา	ออกกำลังกาย	อาชีพ	BP	คอเลสเตอรอล	ลำดับการเกิดโรค	คลาส
1	11	21	40	50	60	749	81	90	<(8128) (8105,8083) (8227) (8084)>	Y
2	10	21	42	52	60	721	80	91	<(8001) (8083,8105,8079) (8092)>	N
3	10	20	40	50	60	749	81	91	<(8080,8026,8001) (8084)>	Y
4	10	21	42	52	60	710	80	91	<(8080) (8001,8116) (8092)>	N
5	11	21	40	50	60	721	80	90	<(8102,8035) (8062,8030)>	N
6	11	21	42	52	61	721	81	90	<(8001) (8035) (8227) (8092)>	Y

3.3 การขุดค้นเซตรายการความถี่ร่วมกับลำดับเหตุการณ์ความถี่

โดยงานวิจัยนี้นำปัจจัยที่ทำให้เกิดโรคและลำดับการเกิดโรคมาจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับโรคหลอดเลือดสมอง โดยขั้นตอนแรกจะทำการสืบค้นเซตรายการความถี่ร่วมกับลำดับเหตุการณ์ความถี่ โดยรวมปัจจัยเสี่ยงเข้ากับลำดับการเกิดโรคเข้าด้วยกัน (ดังตารางที่ 17) และเรียกว่า ลำดับเหตุการณ์

ตารางที่ 17 ตัวอย่างข้อมูล Sequence database

NO	Sequence database
1	<(11),(21),(40),(50),(60),(749),(81),(90), (8128), (8105,8083) ,(8227), (8084)>
2	<(10),(21),(42),(52),(60),(721),(80), (91), (8001), (8083,8105,8079), (8092)>
3	<(10),(20),(40),(50),(60),(749),(81,(91), (8080,8026,8001) ,(8084)>
4	<(10),(21),(42),(52),(60),(710),(80),(91), (8080) (8001,8116) (8092)>
5	<(11),(21),(40),(50),(60),(721),(80),(90), (8102,8035), (8062,8030)>
6	<(11),(21),(42),(52),(61),(721),(81),(90), (8001), (8035), (8227),(8092)>

นำส่วนปัจจัยและลำดับการเกิดโรค ดังตารางที่ 17 ไปค้นหาความถี่ด้วยอัลกอริทึม PrefixSpan โดยขั้นตอนของอัลกอริทึมมีขั้นตอนดังนี้

ขั้นตอนที่ 1 อ่านค่าทรานเซคชั่นในฐานข้อมูล เพื่อค้นหารายการความถี่หรือค่าสนับสนุนของแต่ละราย โดยกำหนดค่าสนับสนุนขั้นต่ำคือ 3 หรือ 50% ขั้นตอนการทำงานจะเริ่มจากการสำรวจ Prefix โปรเจคชั่น ในฐานข้อมูลแบบต่อเนื่องเป็นการค้นหารูปแบบที่สมบรูณ์แต่ลดการสร้างลำดับย่อยของแคนดิเดทและ Prefix โปรเจคชั่น จะลดขนาดของโปรเจคฐานข้อมูลและนำไปสู่การประมวลผลที่มีประสิทธิภาพ ซึ่งจะทำให้ได้ลำดับเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ (minimum support) ที่กำหนดเท่านั้น

ขั้นตอนแรกค้นหาความถี่ลำดับเหตุการณ์ทั้งหมดที่เกิดขึ้นจากตัวอย่างข้อมูล Sequence database ในตารางที่ 17 สามารถแสดงความถี่ลำดับเหตุการณ์ทั้งหมดได้ดังตารางที่ 18

ตารางที่ 18 ความถี่ลำดับเหตุการณ์ทั้งหมด

ลำดับเหตุการณ์	10	11	20	21	40	42	50	52
ความถี่	3	3	1	5	3	3	3	3
ลำดับเหตุการณ์	60	61	710	721	749	80	81	90
ความถี่	5	1	1	3	2	3	3	3
ลำดับเหตุการณ์	91	8001	8026	8030	8035	8062	8079	8080
ความถี่	3	4	1	1	2	1	1	2
ลำดับเหตุการณ์	8083	8084	8092	8102	8105	8116	8128	8227
ความถี่	2	2	3	1	2	1	1	2

เซตลำดับเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำที่กำหนดไว้คือความถี่ 3 ดังในตารางที่ 19

ตารางที่ 19 เซตลำดับเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ

ลำดับเหตุการณ์	**ความถี่**	**ค่าสนับสนุน(%)**
10	3	50
11	3	50
21	5	83.33
40	3	50

ตารางที่ 20 เซตลำดับเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ(ต่อ)

ลำดับเหตุการณ์	ความถี่	ค่าสนับสนุน(%)
42	3	50
50	3	50
52	3	50
60	5	83.33
721	3	50
80	3	50
81	3	50
90	3	50
91	3	50
8001	4	66.66
8092	3	50

โดยเหตุการณ์ที่จำนวนความถี่ของเหตุการณ์ผ่านค่าสนับสนุนที่กำหนดไว้คือ เหตุการณ์ (10),(11),(21),(40),(42),(50),(52),(60),(721),(80),(81),(90),(8001),(8092) เรียกว่า ลำดับเหตุการณ์ความถี่ 1-ลำดับ (1-sequence)

ขั้นตอนที่ 2 การแบ่งพื้นที่การค้นหา

ตารางที่ 21 การแบ่งพื้นที่ค้นหา

(10)	(11)	...	(8001)	(8092)
(21),(42),(52),(60),(721),(80), (91), (8001), (8083,8105,8079), (8092)	(21),(40),(50),(60),(749),(81),(90), (8128), (8105,8083) ,(8227), (8084)	...	(8083,8105, 8079), (8092)	-
(20),(40),(50),(60),(749),(81,(91), (8080,8026,8001) ,(8084)	(21),(40),(50),(60),(721),(80),(90), (8102,8035), (8062,8030)	...	(8084)	-
(21),(42),(52),(60),(710),(80),(91), (8080) (8001,8116) (8092)	(21),(42),(52),(61),(721),(81),(90), (8001), (8035), (8227),(8092	...	(_,8116) (8092)	
		...	(8035),(8227), (8092)	-

ตัวอย่างการค้นหาชุดเหตุการณ์ย่อยของ (10) พิจารณาที่มีเหตุการณ์ (10) อยู่ก่อนหน้า

ตารางที่ 22 เหตุการณ์ย่อยที่ 10 อยู่ก่อนหน้า

(10)
(21),(42),(52),(60),(721),(80), (91), (8001), (8083,8105,8079), (8092)
(20),(40),(50),(60),(749),(81,(91), (8080,8026,8001) ,(8084)
(21),(42),(52),(60),(710),(80),(91), (8080) (8001,8116) (8092)

พิจารณาที่มีเหตุการณ์ (10) อยู่ก่อนหน้าจากตารางที่ 21 จะพบความถี่ของเหตุการณ์ที่ 1-ลำดับดังตารางที่ 23

ตารางที่ 23 ความถี่ของเหตุการณ์ที่ 1-ลำดับทั้งหมด

เหตุการณ์	20	21	40	42	50	52	60	710	721	749
ความถี่	1	2	1	2	1	2	3	1	1	1
เหตุการณ์	91	8001	8083	8105	8079	8092	8080	8026	8084	8116
ความถี่	3	3	1	1	1	2	2	1	1	1

ความถี่เหตุการณ์ที่ 1-ลำดับที่มี (10) อยู่ก่อนหน้า ที่ผ่านค่าสนับสนุนขั้นต่ำ คือ (60) ,(91) , (8001) จะได้ลำดับเหตุการณ์ความถี่ 2 ลำดับ คือ (10)(60) (10)(91) และ (10)(8001) จากนั้นพิจารณาลำดับเหตุการณ์ความถี่ 2 ลำดับดังกล่าว เพื่อขยายลำดับเหตุการณ์ สมมติพิจารณาขยายลำดับเหตุการณ์ความถี่ (10)(60) โดยจะค้นหาเหตุการณ์ย่อยที่ (10)(60) อยู่ก่อนหน้า ดังตารางที่ 24

ตารางที่ 24 เหตุการณ์ย่อยที่ (10)(60) อยู่ก่อนหน้า

(10)(60)
(721),(80), (91), (8001), (8083,8105,8079), (8092)
(749),(81,(91), (8080,8026,8001) ,(8084)
(710),(80),(91), (8080) (8001,8116) (8092)

เมื่อพิจารณาที่มีเหตุการณ์ย่อย (10)(60) อยู่ก่อนหน้าจะสามารถแสดงความถี่ของเหตุการณ์ที่ 2-ลำดับทั้งหมดดังตารางที่ 25

ตารางที่ 25 ความถี่ของเหตุการณ์ที่ 2-ลำดับ

เหตุการณ์	721	749	710	80	81	91	8001	8083
ความถี่	1	1	1	2	1	3	2	1
เหตุการณ์	8105	8079	8092	8080	8026	8084	8116	
ความถี่	1	1	2	2	1	1	1	

พบว่าเหตุการณ์ ความถี่ 2-ลำดับที่มี (10)(60) อยู่ก่อนหน้า ที่ผ่านค่าสนับสนุนขั้นต่ำคือ (91) ดังนั้นจะได้ลำดับเหตุการณ์ความถี่ 3 ลำดับคือ (10)(60)(91) และเมื่อพิจารณาเหตุการณ์ย่อยที่ (10)(60)(91) อยู่ก่อนหน้าจะได้ดังตารางที่ 26

ตารางที่ 26 เหตุการณ์ย่อยที่ (10)(60)(91) อยู่ก่อนหน้า

(10)(60)(91)
(8001), (8083,8105,8079), (8092)
(8080,8026,8001) ,(8084)
(8080) (8001,8116) (8092)

พิจารณาที่มีเหตุการณ์ (10)(60)(91) อยู่ก่อนหน้าสามารถค้นหาความถี่ของเหตุการณ์ที่ 3-ลำดับทั้งหมดได้ดังตารางที่ 27

ตารางที่ 27 ความถี่ของเหตุการณ์ที่ 3-ลำดับ

เหตุการณ์	**8001**	**8083**	**8105**	**8079**	**8092**	**8080**	**8026**	**8084**	**8116**
ความถี่	2	1	1	1	2	2	1	1	1

ไม่พบเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำจึงหยุดขยาย

ดังนั้นลำดับเหตุการณ์ความถี่ที่มี (10) อยู่ก่อนหน้า คือ (10)(60), (10)(91), (10)(8001) และ (10)(60)(91) แล้วทำการขยายลำดับเหตุการณ์อื่นจนกว่าจะไม่พบเหตุการณ์ความถี่ที่ผ่านค่าสนับสนุนขั้นต่ำ

ดังนั้นชุดเหตุการณ์ทั้งหมดคือ (10),(11),(21),(40),(42),(50),(52),(60),(721),(80),(81),(90),(8001),(8092),(10)(60),(10)(91),(10)(8001),(10)(60)(91) ...

3.4 การสร้างกฎ

นำรูปแบบที่ได้จากขั้นตอนในหัวข้อ 3.3 มาสร้างกฎ และใช้ทรานเซคชั่นคำนวณหาค่าสนับสนุนและค่าความเชื่อมั่น โดยกฎที่ได้อยู่ในรูปแบบ $r: X \rightarrow c$ โดยที่ X คือลำดับเหตุการณ์ความถี่และ c คือ คลาส กำหนดให้ g(X) คือ ทรานเซคชั่นที่ปรากฏ X , g(c) คือ ทรานเซคชั่น ที่ปรากฏ c โดยสามารถหาทรานเซคชันของกฎได้จากสมการที่ 3.1

$$g(X \rightarrow c) = g(X) \cap g(c) \qquad (3.1)$$

ทำการค้นหาทรานเซคชันที่ปรากฏลำดับเหตุการณ์ความถี่ และค้นหาทรานเซคชั่นของแต่ละคลาส ดังแสดงในตารางที่ 28 และ 29

ตารางที่ 28 คลาสที่สัมพันธ์กับเซตรายการความถี่

คลาส	ทรานเซคชัน
Y	{1,3,6}
N	{2,4,5}

ตารางที่ 29 ทรานเซคชั่นของลำดับเหตุการณ์ความถี่ตัวอย่าง

ลำดับเหตุการณ์ความถี่	ทรานเซคชั่น
(10)(60)(8001)	{2,3,4}
(10)(60)(8092)	{2,4}
(10)(60)(8001)(8092)	{24}

ตัวอย่างการหาทรานเซคชันของกฎ (10)(60)(8001) $\rightarrow$Y สามารถแสดงได้ดังนี้

$$g((10)(60)(8001) \rightarrow c) = g((10)(60)(8001)) \cap g(Y)$$
$$= \{2,3,4\} \cap \{1,3,6\}$$
$$= \{3\}$$

เมื่อค้นหาทรานเซคชันของกฎทั้งหมดจะแสดงได้ดังตารางที่ 30

ตารางที่ 30 กฎและทรานเซคชันของกฎ

กฎ	ทรานเซคชั่น
(10)(60)(8001) →Y	{3}
(10)(60)(8001) →N	{2,4}
(10)(60)(8092) →Y	{}
(10)(60)(8092) →N	{2,4}
(10)(60),(8001)(8092) →Y	{}
(10)(60)(8001)(8092) →N	{2,4}

จากนั้นทำการคำนวณค่าสนับสนุนและค่าความเชื่อมั่นของกฎ ซึ่งสามารถคำนวณได้ดังสมการที่ 3.2 และ 3.3

$$\text{Support(r)} = \frac{|g(X \rightarrow C)|}{T} x100 \qquad (3.2)$$

โดยที่ T คือจำนวนทรานเซคชั่น

$$\text{Confident(X} \rightarrow \text{C)} = \frac{|g(X \rightarrow C)|}{|g(X)|} x100 \qquad (3.3)$$

ตัวอย่างการคำนวณค่าสนับสนุนและค่าความเชื่อมั่นของกฎ $(10)(60)(8001) \rightarrow N$ สามารถแสดงได้ดังนี้

ถ้า g((10)(60)(8001)) ={2,3,4} และ g(N) ={2,4,5} โดย T คือ จำนวนทรานเซคชั่นทั้งหมด

$$\text{Support ((10)(60)(8001)} \rightarrow \text{N)} = \frac{|g((10)(60)(8001) \rightarrow N)|}{T} x100$$

$$= \frac{|\{2,3,4\} \cap \{2,4,5\}|}{6} x100$$

$$= \frac{|\{2,4\}|}{6} x100$$

$$= \frac{2}{6} x100 = 33.3\%$$

$$\text{Confident ((10)(60)(8001)} \rightarrow \text{N)} = \frac{|g((10)(60)(8001) \rightarrow N)|}{|g((10)(60)(8001))|} x100$$

$$= \frac{|\{2,4\}|}{|\{2,3,4\}|} x100$$

$$= \frac{2}{3} x100 = 66.6\%$$

ค่าสนับสนุนและค่าความเชื่อมั่นของแต่ละกฎแสดงได้ดังตารางที่ 31

ตารางที่ 31 การสร้างกฎทั้งหมด

รูปแบบลำดับต่อเนื่องหลายมิติ	ค่าสนับสนุน (%)	ค่าความเชื่อมั่น (%)
10,60,(8001)=>N	33.33	66.66
10,60,(8001)=>Y	16.66	33.33
10,60,(8092)=>N	33.33	66.66
10,60,(8001)(8092)=>N	33.33	66.66

สมมุติค่าสนับสนุนขั้นต่ำ 30% ความเชื่อมั่นขั้นต่ำ 50% จะได้กฎที่ผ่านดังตารางที่ 32

ตารางที่ 32 กฎที่ผ่านค่าที่กำหนด

รูปแบบลำดับต่อเนื่องหลายมิติ	ค่าสนับสนุน (%)	ค่าความเชื่อมั่น (%)
10,60,(8001)=>N	33.33	66.66
10,60,(8092)=>N	33.33	66.66
10,60,(8001)(8092)=>N	33.33	66.66

3.5 การเรียงกฎ

หลังจากที่ได้กฎทั้งหมดแล้วกฎดังกล่าวจะต้องถูกนำมาพิจารณาเพื่อเรียงกฎก่อนนำไปใช้ในการจำแนกจริง ซึ่งจะพิจารณาตามเงื่อนไข โดยกำหนดให้กฎ R_i อยู่ก่อน R_j ก็ต่อเมื่อ

1. *conf*(R_i) > *conf*(R_j) หมายถึงค่าความเชื่อมั่นของ R_i มากกว่าความเชื่อมั่นของ R_j เพราะความเชื่อมั่นสูงแสดงให้เห็นถึงโอกาศสูงที่จะเป็นหรือไม่โรคหลอดเลือดสมอง

2. *conf*(R_i) = *conf*(R_j) ให้พิจารณา *supp*ort(R_i) > *supp*ort(R_j) หมายถึงถ้าค่าความเชื่อมั่นของ R_i เท่ากับความเชื่อมั่นของ R_j ให้พิจารณาค่าสนับสนุนของ R_i ต้องมากกว่าค่าสนับสนุนของ R_j เพราะว่าแสดงให้เห็นจำนวนการเกิดที่มากกว่า ดังนั้นจึงน่าเชื่อถือกว่า

3. *support*(R_i) = *sup*port(R_j) ให้พิจารณา *size*(R_i) < *size*(R_j) โดย size หมายถึงความยาวของ Antecedent หมายถึงถ้าค่าสนับสนุนของ R_i เท่ากับค่าสนับสนุนของ R_j ให้พิจารณาขนาดความยาวของข้อมูลที่อยู่ฝั่งซ้ายของกฎ R_i ต้องน้อยกว่า R_j เนื่องจากกฎที่มีความยาวสั้นเมื่อนำไป

เทียบกับข้อมูลผู้ป่วยจะทำให้สามารถพบผู้ป่วยที่มีโอกาศเกิดโรคหลอดเลือดได้เร็ว ได้จะถูกนำไปใช้เพื่อการป้องกันและเตรียมความพร้อมในการรักษาทางการแพทย์

เมื่อพิจารณาตามเงื่อนไขข้างต้นข้อมูลในตารางที่ 31 สามารถเรียงกฎได้ดังนี้ตารางที่ 33

ตารางที่ 33 การเรียงกฎ

รูปแบบลำดับต่อเนื่องหลายมิติ	ค่าสนับสนุน (%)	ค่าความเชื่อมั่น (%)
10,60,(8001)=>N	33.33	66.66
10,60,(8092)=>N	33.33	66.66
10,60,(8001)(8092)=>N	33.33	66.66

3.6 การประเมินผล

3.6.1 ประสิทธิภาพในการทำนาย

งานวิจัยนี้ใช้ 10-fold cross validation ในการแบ่งข้อมูลเรียนรู้ (Training Set) และข้อมูลทดสอบ (Testing Set) เพราะผลที่ได้มีความน่าเชื่อถือมากกว่าวิธีอื่นเนื่องจากข้อมูลทุกชุดจะถูกนำมาทดสอบเพื่อประเมินผล โดยในแต่ละรอบจะวัดประสิทธิภาพความถูกต้อง ค่าความแม่นยำ ค่าระลึกและค่าอัตราการรู้จำ จากนั้นหาค่าเฉลี่ยของแต่ละค่าเพื่อดูประสิทธิภาพของตัวจำแนก

ตารางที่ 34 Confusion matrix สำหรับจำแนก[9]

		การทำนาย	
		Yes	No
ค่าความจริง	Yes	*a*	*b*
	No	*c*	*d*

วัดความถูกต้อง(*AC*) โดยรวมในการจำแนกค่าที่เป็นโรคหลอดเลือดสมองและไม่เป็นโรคหลอดเลือดสมองซึ่งสามารถคำนวณได้ดังสมการ 3.1

$$AC = \frac{(a+d)}{(a+b+c+d)} \qquad (3.1)$$

โดยที่ a คือจำนวนกฎที่ทำนายได้ถูกต้องว่าเป็นโรคหลอดเลือดสมอง

d คือจำนวนกฎที่ทำนายได้ถูกต้องว่าเป็นไม่เป็นโรคหลอดเลือดสมอง

b คือจำนวนกฎที่ทำนายว่าไม่เป็นโรคหลอดเลือดสมองแต่ความจริงเป็นโรคหลอดเลือดสมอง

c คือจำนวนกฎที่ทำนายว่าเป็นโรคหลอดเลือดสมองแต่ความจริงไม่เป็นโรคหลอดเลือดสมอง

นอกจากนี้ยังได้วัดประสิทธิภาพในการจำแนก โดยใช้ค่าความแม่นยำค่าระลึกและค่าอัตราการรู้จำโดยค่าความแม่นยำในการทำนายว่าเป็นโรคหลอดเลือดสมองสามารถหาได้จากสมการ 3.2

$$precision_{yes} = \frac{a}{(a+c)} \qquad (3.2)$$

ค่าความแม่นยำในการทำนายว่าไม่เป็นโรคหลอดเลือดสมอง สามารถหาได้จากสมการ 3.3

$$precision_{no} = \frac{d}{(d+b)} \qquad (3.3)$$

ค่าความระลึกในการทำนายว่าเป็นโรคหลอดเลือดสมองสามารถหาได้จากสมการ 3.4

$$recall_{yes} = \frac{a}{(a+b)} \qquad (3.4)$$

ค่าความระลึกในการทำนายว่าไม่เป็นโรคหลอดเลือดสมอง สามารถหาได้จากสมการ 3.5

$$recall_{no} = \frac{d}{(c+d)} \qquad (3.5)$$

ส่วนค่าอัตราการรู้จำในการทำนายว่าเป็นโรคหลอดเลือดสมองและไม่เป็นโรคหลอดเลือดสมอง สามารถคำนวณได้ดังสมการ 3.6 และ 3.7 ตามลำดับ

$$\text{F} - \text{Measure}_{yes} = \frac{2\text{x Precision}_{yes}\ \text{x Recall}_{yes}}{\text{Precision}_{yes} + \text{Recall}_{yes}} \qquad (3.6)$$

$$\text{F} - \text{Measure}_{no} = \frac{2\text{x Precision}_{no}\ \text{x Recall}_{no}}{\text{Precision}_{no} + \text{Recall}_{no}} \qquad (3.7)$$

3.6.2 ประสิทธิภาพการประมวลผล

ในงานวิจัยนี้พัฒนาวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมองจะประมวลผลเปรียบเทียบประสิทธิภาพในการจำแนก 2 ส่วนคือ

1. วิธีการสร้างกฎจากเซตปัจจัยร่วมกับลำดับการเกิดโรคเปรียบเทียบกับการสร้างกฎจากเซตปัจจัยอย่างเดียวและการสร้างกฎจากเซตลำดับโรคอย่างเดียวโดยเปรียบเทียบค่าความถูกต้องและจำนวนกฎ
2. เปรียบเทียบค่าค่าความถูกต้อง ความแม่นยำ ค่าระลึก และค่าอัตราการรู้จำ ของวิธีการที่นำเสนอเทียบกับ KNN และ Naïve Baye

บทที่ 4
ผลการวิจัยและการอภิปราย

4.1 ผลการเก็บรวบรวมข้อมูล

งานวิจัยนี้ได้ทำการสร้างวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมอง และหาความสัมพันธ์ของปัจจัยกับลำดับเกิดโรคที่นำไปสู่โรคหลอดเลือดสมอง โดยข้อมูลที่ใช้สำหรับวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์ จะต้องเป็นข้อมูลที่ประกอบด้วยข้อมูลที่ไม่พิจารณาลำดับการเกิดของข้อมูลคือปัจจัย ได้แก่ เพศ สถานะ อาชีพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย ระดับความดันโลหิต ระดับคลอเลสเตอรอล และข้อมูลที่พิจารณาลำดับการเกิดของข้อมูลเป็นสำคัญคือลำดับการเกิดโรค ซึ่งเป็นผลการวินิฉัยทางการแพทย์ด้วย ICD-10 ที่อ้างอิงโรคและอาการของโรค พร้อมกับระบุคลาสที่จะจำแนกว่าเป็นหรือไม่เป็นโรคหลอดเลือดสมอง ซึ่งข้อมูลที่ใช้ในการทดลองในงานวิจัยนี้คือข้อมูลปัจจัยเสี่ยงที่เกี่ยวข้องกับโรคหลอดเลือดสมองและลำดับการเกิดโรค ซึ่งนำมาจากผลการรักษาที่มารับบริการในโรงพยาบาลมหาสารคามจำนวน 1,000 คน ซึ่งเป็นผู้ป่วยที่อายุ 60 ขึ้นไปมารับบริการในโรงพยาบาลตั้งแต่วันที่ 1 มกราคม พ.ศ. 2555 ถึง 31 ธันวาคม พ.ศ. 2559 โดยลักษณะของชุดข้อมูลที่นำมาวิจัยประกอบด้วย ปัจจัยและลำดับการเกิดโรค ปัจจัยที่ใช้ในงานวิจัยนี้ประกอบไปด้วย 8 ปัจจัย ได้แก่ เพศ สถานะภาพ อาชีพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย ระดับความดันโลหิต ระดับคลอเรสเตอรอล ส่วนลำดับการเกิดโรคเป็นจำนวนกลุ่มโรค 169 กลุ่มโรค ซึ่งพิจารณาจากกลุ่มโรคมาตรฐาน ICD-10 ทั้งหมด จำนวน 228 กลุ่มโรค ความยาวเฉลี่ยของลำดับโรค 7 โรค ความยาวสูงสุดของลำดับโรค 14 โรค ความยาวขั้นต่ำของลำดับโรค 3 โรค จำนวนผู้ป่วยโดยเฉลี่ยต่อโรค 7 คน จำนวนผู้ป่วยสูงสุดของโรค 485 คน จำนวนผู้ป่วยน้อยสุดของโรค 1 คน รายละเอียดชุดข้อมูลได้ดังแสดงในตารางที่ 35-36

ตารางที่ 35 ลักษณะของชุดข้อมูล

ลักษณะเฉพาะ	จำนวน	หน่วย
จำนวนของปัจจัย	8	ปัจจัย
จำนวนกลุ่มโรค (จำนวนกลุ่มโรคทั้งหมดตามมาตรฐาน ICD-10 คือ 228 กลุ่มโรค)	169	กลุ่มโรค
ความยาวเฉลี่ยของลำดับโรค	7	ลำดับ
ความยาวสูงสุดของลำดับโรค	14	ลำดับ

ตารางที่ 36 ลักษณะของชุดข้อมูล(ต่อ)

ลักษณะเฉพาะ	จำนวน	หน่วย
ความยาวขั้นต่ำของลำดับโรค	3	ลำดับ
จำนวนผู้ป่วยโดยเฉลี่ยต่อโรค	7	คน
จำนวนผู้ป่วยสูงสุดของโรค	485	คน
จำนวนผู้ป่วยน้อยสุดของโรค	1	คน

4.2 ผลการทดลอง

งานวิจัยนี้ ทำการวัดประสิทธิภาพโดยใช้ 10-fold cross validation ในการแบ่งข้อมูลเรียนรู้และทดสอบและวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าระลึก ค่าอัตราเรียนรู้สำหรับแต่ละคลาส นอกจากนี้ยังแสดงจำนวนกฎเฉลี่ยที่ใช้ในการสร้างตัวจำแนก

1) วัดประสิทธิภาพการจำแนกแบบความสัมพันธ์ด้วยปัจจัยอย่างเดียวโดยเปรียบเทียบค่าความถูกต้องและจำนวนกฎ ผลการทดสอบแสดงดังตารางที่ 37 และ38

ตารางที่ 37 ค่าความถูกต้องในการจำแนกข้อมูลด้วยปัจจัย 8 ปัจจัย

ค่าสนับสนุนขั้นต่ำ(%)	ความเชื่อมั่นขั้นต่ำ(%)						
		50	60	70	80	90	100
	10	62.72	62.72	62.72	72.8	85.37	**95.3**
	20	61.88	61.88	61.88	74.24	87.37	**97.1**
	30	61.89	61.89	61.89	76.69	90.06	**98.4**
	40	61.8	61.8	61.8	76.69	91.56	**100**
	50	61.85	61.85	61.85	78.16	91.56	**100**

ตารางที่ 38 จำนวนกฎในการจำแนกข้อมูลด้วยปัจจัย 8 ปัจจัย

ค่าสนับสนุนขั้นต่ำ(%)	ความเชื่อมั่นขั้นต่ำ(%)					
	50	60	70	80	90	100
10	2,757	2,149	1,436	871	570	325
20	1,674	1,288	766	398	214	85
30	1,232	942	502	208	108	24
40	960	738	360	114	66	11
50	742	569	258	63	43	3

ตารางที่ 37 และ 38 แสดงผลการทดสอบกับข้อมูลปัจจัยที่ค่าสนับสนุนขั้นต่ำ คือ 10% 20% 30% 40% และ 50% และค่าความเชื่อมั่นขั้นต่ำ คือ 50% 60% 70% 80% 90% 100% พบว่าข้อมูลปัจจัยได้ค่าความเชื่อมั่นที่สูงถึง 100% ที่ค่าสนับสนุนขั้นต่ำ 40% และ 50% และค่าความเชื่อมั่นขั้นต่ำ 100% ซึ่งข้อมูลปัจจัยเป็นข้อมูลที่ไม่กระจายตัวมาก เช่น พฤติกรรมสูบบุหรี่มี 3 ประเภท คือ สูบบุหรี่ เคยแต่เลิกสูบแล้ว และไม่สูบบุหรี่ หรือ พฤติกรรมการออกกำลังกาย คือ ออกกำลังกาย และไม่ออกกำลังกาย ทำให้สามารถได้ค่าความถูกต้องที่สูง และพบว่าค่าความเชื่อมั่นต่ำและค่าสนับสนุนต่ำจะได้จำนวนกฎที่มาก ในทางตรงกันข้าม เมื่อค่าสนับสนุนมาก ค่าความเชื่อมั่นสูง ทำให้ได้กฎจำนวนต่ำแต่เป็นกฎที่มีประสิทธิภาพในการจำแนกสูง

2) วัดประสิทธิภาพการจำแนกด้วยลำดับการเกิดโรคโดยเปรียบเทียบค่าความถูกต้องและจำนวนกฎ ผลการทดสอบแสดงดังตารางที่ 39 และ 40

ตารางที่ 39 ค่าความถูกต้องในการจำแนกข้อมูลด้วยลำดับโรค

ค่าสนับสนุนขั้นต่ำ(%)	ความเชื่อมั่นขั้นต่ำ(%)					
	50	60	70	80	90	100
10	39.91	39.98	45.18	45.67	54.64	**68.47**
20	36.89	36.96	49.34	52.91	n/a	n/a
30	37.46	37.62	50.36	n/a	n/a	n/a
40	37.88	38.21	n/a	n/a	n/a	n/a
50	37.17	37.57	n/a	n/a	n/a	n/a

ตารางที่ 40 จำนวนกฎในการจำแนกข้อมูลด้วยลำดับโรค

ค่าสนับสนุนขั้นต่ำ(%)	ความเชื่อมั่นขั้นต่ำ(%)					
	50	60	70	80	90	100
10	249	171	94	35	10	4
20	75	43	11	4	0	0
30	43	20	2	0	0	0
40	27	12	0	0	0	0
50	24	9	0	0	0	0

ตารางที่ 39 และ 40 แสดงผลการทดสอบกับข้อมูลลำดับการเกิดโรค ที่ค่าสนับสนุนขั้นต่ำ คือ 10% 20% 30% 40% และ 50% และค่าความเชื่อมั่นขั้นต่ำ คือ 50% 60% 70% 80% 90% 100% พบว่าข้อมูลลำดับการเกิดโรคได้ค่าความถูกต้องที่สูงคือ 68.47% ที่ค่าสนับสนุนขั้นต่ำ 10% และค่าความเชื่อมั่นขั้นต่ำ 100% แต่เมื่อค่าสนับสนุนเพิ่มขึ้นไม่มีกฎที่ใช้ในการจำแนก ดังตารางที่ 4.4 ข้อมูลลำดับการเกิดโรคเป็นข้อมูลที่กระจายตัวมาก โดยข้อมูลที่เก็บรวบรวมมีจำนวนกลุ่มโรคถึง 169 กลุ่ม และความยาวเฉลี่ยของลำดับโรคถึง 7 ลำดับ ทำให้ได้ค่าความถูกต้องที่น้อยเมื่อเทียบกับการจำแนกด้วยปัจจัย และพบว่าเมื่อพิจารณาความเชื่อมั่นที่ 80% และค่าสนับสนุนที่ 30% จะไม่พบกฎเลย

3) วัดประสิทธิภาพการจำแนกด้วยปัจจัยร่วมกับลำดับการเกิดโรคโดยเปรียบเทียบค่าความถูกต้องและจำนวนกฎ ผลการทดสอบแสดงดังตารางที่ 41 และ 42

ตารางที่ 41 ค่าความถูกต้องการจำแนกข้อมูลปัจจัยร่วมกับลำดับการเกิดโรค

ค่าสนับสนุนขั้นต่ำ(%)	ความเชื่อมั่นขั้นต่ำ(%)					
	50	60	70	80	90	100
10	60.12	60.12	60.12	60.27	61.56	**66.78**
20	58.23	58.23	58.23	58.53	64.58	**89.91**
30	58.36	58.36	58.36	60.12	73.62	**97.64**
40	60.07	60.07	60.07	62.62	77.43	**97.75**
50	60.69	60.69	60.69	65.75	81.05	**97.57**

ตารางที่ 42 จำนวนกฎการจำแนกข้อมูลปัจจัยร่วมกับลำดับการเกิดโรค

ค่าสนับสนุนขั้นต่ำ(%)		ความเชื่อมั่นขั้นต่ำ(%)					
		50	60	70	80	90	100
	10	34,141	28,769	20,783	10,963	4,527	10,870
	20	10,870	8,992	5,931	2,378	595	173
	30	5,929	4,935	3,126	1,000	204	36
	40	3,742	3,126	1,864	503	108	13
	50	2,470	2,049	1,144	273	69	3

ตารางที่ 41 และ 42 แสดงผลการทดสอบกับข้อมูลปัจจัยร่วมกับลำดับการเกิดโรค โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 10% 20% 30% 40% และ 50% และกำหนดค่าความเชื่อมั่นขั้นต่ำเท่ากับ 50% 60% 70% 80% 90% และ 100% เนื่องจากปัจจัยร่วมกับลำดับการเกิดโรคมีรูปแบบข้อมูลที่กระจายตัวน้อยกว่าลำดับการเกิดโรคอย่างเดียว จึงทำให้ประสิทธิภาพในการจำแนกโรคได้ค่าที่สูงเมื่อกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 10% 20% 30% 40% และ 50% และกำหนดค่าความเชื่อมั่นขั้นต่ำเท่ากับ 100% จากผลการทดลองวัดประสิทธิภาพพบว่าที่ค่าสนับสนุนขั้นต่ำ 40% ความเชื่อมั่นขั้นต่ำ 100% ได้ค่าความถูกต้องสูงสุดคือ 97.75%

4) วัดประสิทธิภาพการจำแนกเปรียบเทียบ 3 แบบ ผลการทดสอบแสดงดังตารางที่ 43

ตารางที่ 43 เปอร์เซ็นต์การเปรียบเทียบค่าความถูกต้องของการจำแนก 3 แบบ

วิธีการจำแนก	ค่าตัวแปร	ค่าสนับสนุน(%)				
		10	20	30	40	50
ปัจจัย	**ความเชื่อมั่นขั้นต่ำ=100%**	95.3	97.1	98.4	100	**100**
ลำดับการเกิดโรค		68.47	n/a	n/a	n/a	n/a
ปัจจัยร่วมกับลำดับการเกิดโรค		66.78	89.91	97.64	**97.75**	97.57

จากตารางที่ 43 เมื่อเปรียบเทียบค่าความถูกต้องพบว่าที่ค่าความเชื่อมั่นขั้นต่ำเท่ากับ 100% ให้ค่าความถูกต้องที่สูงที่สุด วิธีการจำแบบปัจจัยอย่างเดียวให้ค่าสูง 100% ในขณะที่วิธีการจำแบบลำดับการเกิดโรคไม่สามารถคำนวณหากฎได้เมื่อค่าสนับสนุนขั้นต่ำมากกว่า 20% และวิธีการจำแบบปัจจัยร่วมกับลำดับการเกิดโรคให้ค่าสูงสุดที่ 97.75% ซึ่งค่าความถูกต้องขึ้นกับลักษณะของข้อมูลที่ใช้รวมถึงปริมาณการกระจายตัวของข้อมูลที่นำเข้า

5) วัดประสิทธิภาพการจำแนกด้วยปัจจัยร่วมกับลำดับการเกิดโรคโดยเปรียบเทียบกับอัลกอริทึมอื่นผลการทดสอบแสดงดังตารางที่ 44

ตารางที่ 44 เปอร์เซ็นต์การเปรียบเทียบผลการดำเนินงานเบื้องต้น

วิธีการจำแนก	ค่าตัวแปร		ค่าความถูกต้อง	เป็นโรคหลอดเลือดสมอง			ไม่เป็นโรคหลอดเลือดสมอง		
				ค่าความแม่นยำ	ค่าระลึก	อัตราการรู้จำ	ค่าความแม่นยำ	ค่าระลึก	อัตราการรู้จำ
วิธีการที่นำเสนอ ค่าเชื่อมั่นขั้น=100%	ค่าสนับสนุนขั้นต่ำ	=10%	66.78	60.97	66.65	63.68	71.72	66.54	69.04
		=20%	89.91	91.38	93.63	92.63	n/a	73.26	n/a
		=30%	97.64	97.64	100	98.81	n/a	n/a	n/a
		=40%	***97.75***	97.75	100	98.86	n/a	n/a	n/a
		=50%	***97.57***	97.57	100	98.77	n/a	n/a	n/a
KNN	K=3 Euclidean distance		***64.00***	63.59	68.00	65.72	64.48	66.00	65.23
Naïve Baye	Mn		***63.10***	61.43	74.00	67.13	65.62	54.00	59.24

เมื่อเปรียบเทียบค่าความถูกต้องกับวิธีการจำแนก KNN และ Naive Baye พบว่าวิธีการที่นำเสนอได้ผลดีกว่า ซึ่งแสดงให้เห็นว่าการพิจารณาปัจจัยร่วมกับลำดับการเกิดโรคที่ต้องพิจารณาลำดับการเกิดของโรคจะให้ประสิทธิภาพการจำแนกที่สูงกว่า แต่อย่างไรก็ตามเนื่องจากคุณลักษณะของข้อมูลลำดับการเกิดโรคมีจำนวนมากและมีการกระจายของข้อมูลที่หลากหลายทำให้ไม่สามารถค้นหากฎที่ผ่านค่าสนับสนุนขั้นต่ำที่มีค่าสูงได้ จากตารางที่ 43 จะเห็นได้ว่าเมื่อกำหนดค่าสนับสนุนขั้นต่ำตั้งแต่ 20% ขึ้นไปจะไม่สามารถค้นหากฎที่ใช้ในการจำแนกผู้ที่ไม่เป็นโรคหลอดเหลือดสมองได้ แต่ยังสามารถสร้างกฎที่ใช้ในการจำแนกโรคหลอดเหลือดสมองที่มีประสิทธิภาพได้

4.3 กฎที่ได้จากงานวิจัย

เมื่อได้ค่าสนับสนุนขั้นต่ำ ค่าความเชื่อมั่นขั้นต่ำที่เหมาะสม คือ ค่าสนับสนุนขั้นต่ำ 40% และ ค่าความเชื่อมั่นขั้นต่ำ 100% จะทำการเรียงกฎ และเลือกกฎ 10 กฎแรกดังตารางที่ 45

ตารางที่ 45 กฎ 10 ลำดับแรกที่นำไปสู่โรคหลอดเลือดสมอง

ลำดับ	กฎความสัมพันธ์	ค่าสนับสนุน (%)	ค่าความเชื่อมั่น (%)
1	สูบบุหรี่ ดื่มสุรา	67.77	100
2	ดื่มสุรา โรคความดันโลหิตสูง	66.66	100
3	สูบบุหรี่ ดื่มสุรา ระดับความดันโลหิตปกติ	61.11	100
4	สูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย	60.00	100
5	สมรส สูบบุหรี่ ดื่มสุรา	58.88	100
6	สมรส ดื่มสุรา โรคความดันโลหิตสูง	57.77	100
7	สูบบุหรี่ ดื่มสุรา ระดับคอเลสเตอรอลปกติ	54.44	100
8	สมรส สูบบุหรี่ ดื่มสุรา ระดับความดันโลหิตปกติ	54.44	100
9	สูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย ระดับความดันโลหิตปกติ	53.33	100
10	ชาย ดื่มสุรา โรคความดันโลหิตสูง	52.22	100

จากตารางที่ 45 พบว่ากฎ 10 กฎแรกมีค่าความเชื่อมั่น100% ทุกกฎ และมีค่าสนับสนุนขั้นต่ำที่ 52.22% โดยกฎแต่กฎแสดงให้เห็นความสัมพันธ์ของปัจจัยและลำดับการเกิดโรคที่นำไปสู่โรคหลอดเลือดสมองดังต่อไปนี้

กฎความสัมพันธ์ลำดับที่ 1

สูบบุหรี่ ดื่มสุรา >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้ามีพฤติกรรมสูบบุหรี่ และดื่มสุราจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 2

ดื่มสุรา โรคความดันโลหิตสูง >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้ามีพฤติกรรมดื่มสุรา และเป็นโรคความดันโลหิตสูงจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 3

สูบบุหรี่ ดื่มสุรา ระดับความดันโลหิตปกติ >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้าพฤติกรรมสูบบุหรี่ และดื่มสุรา ระดับความดันโลหิตปกติจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 4

สูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้ามีพฤติกรรมสูบบุหรี่ ดื่มสุรา และไม่ออกกำลังกายจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 5

สมรส สูบบุหรี่ ดื่มสุรา >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้าสถานะภาพสมรส มีพฤติกรรมสูบบุหรี่ และดื่มสุราจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 6

สมรส ดื่มสุรา โรคความดันโลหิตสูง >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้าสถานะภาพสมรส มีพฤติกรรมดื่มสุราและเป็นโรคความดันโลหิตสูงจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 7

สูบบุหรี่ ดื่มสุรา ระดับคอเลสเตอรอลปกติ >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้ามีพฤติกรรมสูบบุหรี่ ดื่มสุราและระดับคอเลสเตอรอลปกติจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 8

สมรส สูบบุหรี่ ดื่มสุรา ระดับความดันโลหิตปกติ >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้าสถานะภาพสมรส มีพฤติกรรมสูบบุหรี่ ดื่มสุรา และระดับความดันโลหิตปกติจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 9

สูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย ระดับความดันโลหิตปกติ >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้ามีพฤติกรรมสูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย และระดับความดันโลหิตปกติจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

กฎความสัมพันธ์ลำดับที่ 10

ชาย ดื่มสุรา โรคความดันโลหิตสูง >>โรคหลอดเลือดสมอง

ตีความได้ว่า ถ้าเป็น เพศชาย มีพฤติกรรมดื่มสุรา และเป็นโรคความดันโลหิตสูงจะเป็นโรคหลอดเลือดสมองมีค่าความเชื่อมั่นที่ 100%

โดยกฎความสัมพันธ์ที่ได้ไม่จำเป็นต้องเรียงลำดับ เช่น กฎความสัมพันธ์ลำดับที่ 9 สูบบุหรี่ ดื่มสุรา ไม่ออกกำลังกาย ระดับความดันโลหิตปกติ แม้จะดื่มสุรา ก่อนสูบบุหรี่ เพียงแค่มีทุกข้อ แม้ว่าสิ่งใดเกิดก่อนหลังไม่ได้มองว่าสำคัญแต่มองว่าครบทุกข้อก็ทำให้เป็นโรคหลอดเลือดสมอง

สำหรับทางการแพทย์การวินิจฉัยหรือการตรวจสอบโรคของผู้ป่วยเป็นเรื่องที่มีความละเอียดอ่อนและค่อนข้างซับซ้อนมากรวมไปถึงยังต้องใช้ผู้เชี่ยวชาญทางด้านการแพทย์มาวิเคราะห์หรือวินิจฉัยจากอาการของผู้ป่วยที่เป็นโรคต่าง ๆ ดังนั้นการสร้างกฎความสัมพันธ์ที่ดีที่สุดและครอบคลุมการวินิจฉัยให้มากที่สุดรวมถึงให้ค่าความเชื่อมั่นที่ดีที่สุดเป็นสิ่งสำคัญและจำเป็นต่อวงการทางการแพทย์ในปัจจุบัน

บทที่ 5
สรุปผล อภิปราย และข้อเสนอแนะ

5.1 สรุป

ปัจจุบันประชากรผู้สูงอายุในสังคมมีอายุยืนยาวขึ้นแต่ด้วยสภาพความเสื่อมของร่างกายทำให้ผู้สูงอายุเจ็บป่วยได้ง่าย และต้องเข้ารับการรักษาบ่อยครั้ง โดยกระทรวงสาธารณสุขพบสาเหตุที่ทำให้ผู้สูงอายุสูญเสียปีสุขภาวะสูงที่สุดคือโรคหลอดเลือดสมองซึ่งเป็นสาเหตุที่ทำให้เกิดความพิการและเสียชีวิต กระทรวงสาธารณะสุขจึงประกาศให้โรคหลอดเลือดสมองเป็นปัญหาสุขภาพสำคัญของคนไทย ปัจจุบันได้มีการประยุกต์ใช้เหมืองข้อมูลในทางการแพทย์เพื่อการรักษาที่แพร่หลาย โดยวิเคราะห์ปัจจัยเสี่ยงของโรคเพื่อหาความสัมพันธ์ของปัจจัยเสี่ยงต่างๆ เป็นต้น ซึ่งงานวิจัยส่วนใหญ่ใช้ปัจจัยเพียงอย่างเดียวในการวิเคราะห์โดยไม่ได้พิจารณาลำดับการเกิดโรคเข้ามาร่วม ซึ่งการเกิดโรคเป็นการเกิดอย่างต่อเนื่องและโรคหนึ่งอาจนำไปสู่อีกโรคอื่น

งานวิจัยนี้จึงทำการศึกษาการสร้างวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับจำแนกโรคหลอดเลือดสมอง และหาความสัมพันธ์ของปัจจัยร่วมกับลำดับเกิดโรคที่นำไปสู่โรคหลอดเลือดสมอง โดยข้อมูลที่ใช้สำหรับวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์ จะต้องเป็นข้อมูลที่ประกอบด้วยข้อมูลที่ไม่พิจารณาลำดับการเกิดของข้อมูลคือปัจจัยเสี่ยง ได้แก่ เพศ สถานะ อาชีพ การสูบบุหรี่ การดื่มสุรา การออกกำลังกาย ระดับความดันโลหิต ระดับคลอเลสเตอรอล และข้อมูลที่พิจารณาลำดับการเกิดของข้อมูลเป็นสำคัญคือลำดับการเกิดโรคซึ่งเป็นผลการวินิฉัยทางการแพทย์ด้วย ICD-10 ที่อ้างอิงโรคและอาการของโรค พร้อมกับระบุคลาสที่จะจำแนกว่าเป็นหรือไม่เป็นโรคหลอดเลือดสมอง ซึ่งข้อมูลที่ใช้ในการทดลองในงานวิจัยนี้คือข้อมูลปัจจัยเสี่ยงที่เกี่ยวข้องกับโรคหลอดเลือดสมองและลำดับการเกิดโรค ซึ่งได้จากผลการรักษาที่มารับบริการในโรงพยาบาลมหาสารคามจำนวน 1,000 คน ซึ่งเป็นผู้ป่วยที่อายุ 60 ขึ้นไปและมารับบริการในโรงพยาบาลเก็บข้อมูลเริ่มตั้งแต่วันที่ 1 มกราคม พ.ศ. 2555 ถึง 31 ธันวาคม พ.ศ. 2559 จากนั้นทำความสะอาดข้อมูล และแปลงข้อมูลให้อยู่ในรูปแบบตัวเลขและตัวอักษร

การศึกษาครั้งนี้นำเสนอวิธีการที่มีประสิทธิภาพในการจำแนกแบบโรคหลอดเลือดสมองการสร้างกฎความสัมพันธ์สำหรับการคาดการณ์โรคหลอดเลือดสมองที่ถูกสร้างขึ้นตามการสืบค้นลำดับเหตุการณ์และการจำแนกแบบความสัมพันธ์ และเรียงลำดับกฎเพื่อปรับประสิทธิภาพในการคาดการณ์โรค ตรวจสอบค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำสำหรับการคาดการณ์โรค เปรียบเทียบตัวจำแนกที่นำเสนอกับ KNN และ Naïve Baye

จากผลการทดลองพบว่าตัวจำแนกที่นำเสนอให้ประสิทธิภาพในการคาดการณ์โรคได้ดี และมีประสิทธิภาพดีกว่า KNN และ Naïve Bay และในการศึกษาครั้งนี้ได้แสดงกฎความสัมพันธ์ 10 ลำดับแรกสำหรับการทำนายโรคหลอดเลือดสมองให้เห็นว่า การสูบบุหรี่ การดื่มสุรา และโรคความดันโลหิตสูง มีผลต่อการกระตุ้นและนำไปสู่โรคหลอดเลือดสมอง

5.2 ข้อเสนอแนะและงานวิจัยในอนาคต

ข้อเสนอแนะในงานวิจัยนี้มีดังนี้

1. เนื่องจากข้อจำกัดของข้อมูลผู้ป่วยในการบันทึกในโรงพยาบาล เช่น ผู้ให้ข้อมูลให้ข้อมูลเตกต่างกันในแต่ละครั้ง ผู้กรอกข้อมูลไม่ใช่คนเดิมในแต่ละครั้ง ข้อความที่กรอกหรือนำเข้าไปในระบบจะไม่เหมือนกัน ซึ่งจะส่งผลต่อการเตรียมข้อมูลที่ต้องแปรผลให้ถูกต้อง ดังนั้นถ้าต้องการชุดข้อมูลที่อยู่ในรูปแบบเดียวกัน จำเป็นต้องรวบรวมข้อมูลด้วยตัวเอง เช่น ทำแบบสอบถาม ซึ่งใช้ระยะเวลานานในการรวบรวมข้อมูล
2. ปริมาณข้อมูลผู้ป่วยที่เป็นโรคหลอดเลือดสมองมีปริมาณข้อมูลน้อย และลำดับโรคที่เกิดก่อนไม่ชัดเจนเพราะผู้ป่วยเป็นแบบฉับพลัน ซึ่งอาจเป็นผู้ที่ไม่เคยเข้ารับบริการที่โรงพยาบาลก็เป็นไปได้
3. เพิ่มเติมปัจจัยเสี่ยงที่มีผลกระทบต่อผู้ป่วยโรคหลอดเลือดสมอง เช่น การได้รับยา ผลทางห้องปฏิบัติการ เป็นต้น
4. สำหรับงานวิจัยที่เกี่ยวข้องกับข้อมูลทางการแพทย์ในการค้นหากฎความสัมพันธ์ รวมถึงงานของวิทยานิพนธ์ฉบับนี้จำเป็นที่จะต้องมีผู้เชี่ยวชาญทางด้านการแพทย์เพื่อทำการตรวจสอบกฎความสัมพันธ์ที่ได้อีกครั้งก่อนการนำไปใช้งานจริงทางการแพทย์

งานวิจัยในอนาคตมีดังนี้

ทำการศึกษาการสร้างวิธีการจำแนกแบบความสัมพันธ์ร่วมกับการสืบค้นลำดับเหตุการณ์สำหรับโรคอื่น เช่น วัณโรค เป็นต้น

บรรณานุกรม

บรรณานุกรม

[1] กรมอนามัย กระทรวงสาธารณสุข. แผนยุทธศาสตร์สุขภาพกระทรวงสาธารณสุข ด้านส่งเสริมสุขภาพ และป้องกันโรค ปีงบประมาณ พ.ศ.2557;2557[ฉบับที่ 2]:[สืบค้นเมื่อ 10 กรกฎคม 2559];www.anamai.moph.go.th/download/download/แผนยุทธศาสตร์กรมอนามัย2557.pdf

[2] World Health Organization . Mortality database tables 2008; [12 ธันวาคม 2560];www.who.int/healthinfo/mortality_data/en/

[3] ศุกร์ใจ วุฒิกิจโกศล."การใช้เทคนิคการทำเหมืองข้อมูลในผู้ป่วยข้อไหล่ติด โรงพยาบาลพระนั่งเกล้า" [การศึกษาอิสระปริญญาวิทยาศาสตรมหาบัณฑิต]. นครศรีธรรมราช:มหาลัยวลัยลักษณ์;2551;

[4] รักถิ่น เหลาหา. "การพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด โดยใช้ทฤษฎีของการทำเหมืองข้อมูล" [การศึกษาอิสระปริญญาวิทยาศาสตรมหาบัณฑิต].ขอนแก่น: มหาวิทยาลัยขอนแก่น; 2553;

[5] อังคณา พิจารโชติ." ระบบสนับสนุนการตัดสินใจสำหรับวิเคราะห์ปัจจัยเสี่ยงการเป็นโรคเบาหวาน" [การศึกษาอิสระปริญญาวิทยาศาสตรมหาบัณฑิต].ขอนแก่น: มหาวิทยาลัยขอนแก่น; 2552;

[6] โรคหลอดเลือดสมอง อัมพฤกษ์ อัมพาต.[สืบค้นเมื่อ 15 กรกฎคม 2559] ;https://th.yanhee.net/หัตถการ/โรคหลอดเลือดสมอง/

[7] สำนักงานโรคไม่ติดต่อ กรมควบคุมโรค.จำนวนและอัตราตายด้วยโรคไม่ติดต่อและการบาดเจ็บประจำปีปฏิทิน พ.ศ. 2558. [สืบค้นเมื่อ 8 พฤษภาคม 2560]; www.thaincd.com/document/hotWorldStrokeday2016.pdf.

[8] บุญเสริม กิจศิริกุล. "อัลกอริทึมการทำเหมืองข้อมูล". กรุงเทพฯ :มหาวิทยาลัยจุฬาลงกรณ์; 2545.

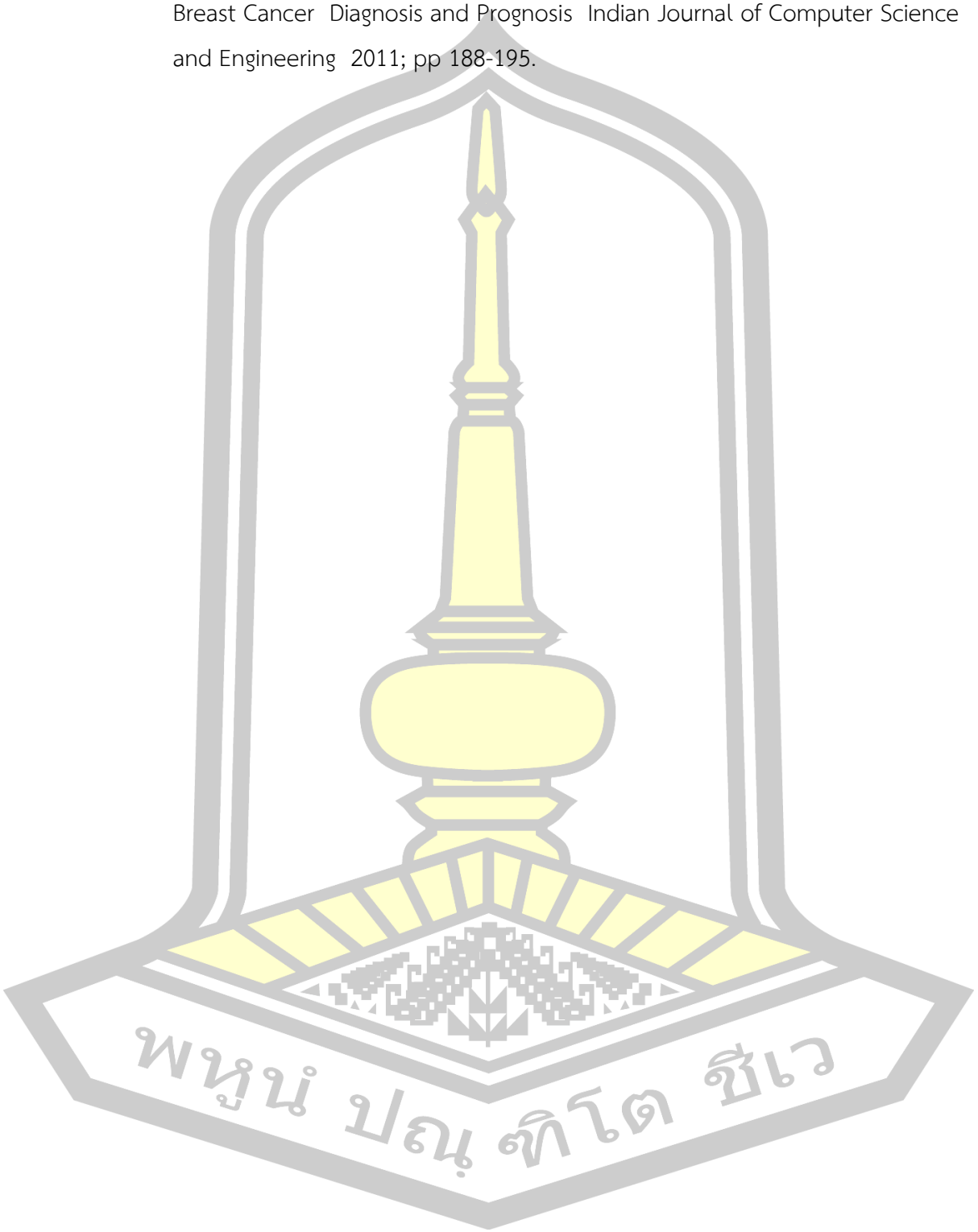
[9] เอกสิทธิ์ พัชรวงศ์ศักดา. การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น. บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด: สิงหาคม 2557.

[10] Mullinset al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in Biology and Medicine 2005; 36(12)1351-1377.

[11] Richards et al. Data Mining for indicator of early mortality in a database of clinical records. ArtifIntell Med 2001; 22(3)215-231.

[12] Organization WH. ICD-10 บัญชีจำแนกโรคระหว่างประเทศ. 2006.

[13] SGupta D, ASharma. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis Indian Journal of Computer Science and Engineering 2011; pp 188-195.



ประวัติผู้เขียน

ชื่อ	นางสาวสุจิตรา นาสิงห์ขันธุ์
วันเกิด	29 มีนาคม พ.ศ. 2528
สถานที่เกิด	อำเภอฆ้องชัย จังหวัดกาฬสินธุ์
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 70 หมู่ที่ 1 บ้านตูม ตำบลเหล่ากลาง อำเภอฆ้องชัย จังหวัดกาฬสินธุ์ รหัสไปรษณีย์ 46130
ตำแหน่งหน้าที่การงาน	นักวิชาการคอมพิวเตอร์
สถานที่ทำงานปัจจุบัน	โรงพยาบาลมหาสารคาม 168 ถนนผดุงวิถี ตำบลตลาด อำเภอเมือง จังหวัดมหาสารคาม
ประวัติการศึกษา	พ.ศ. 2546 มัธยมศึกษาตอนปลาย โรงเรียนกมลาไสย อำเภอกมลาไสย จังหวัดกาฬสินธุ์
	พ.ศ. 2550 ปริญญาวิทยาศาสตรบัณฑิต(วท.บ.) สาขาวิชาระบบสารสนเทศเพื่อการจัดการ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
	พ.ศ. 2562 ปริญญาวิทยาศาสตรมหาบัณฑิต(วท.ม.) สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม