



ใบรับรองวิทยานิพนธ์
บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตร์มหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์

วิทยาการคอมพิวเตอร์

สาขา

ภาควิชา

เรื่อง การทำเหมืองข้อมูลโรคไข้เลือดออก

Data Mining on Dengue Virus Disease

นามผู้วิจัย นางสาวดารณี ฐิติประยูรวงศ์

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์นวลวรรณ สุนทรภิชช์, Ph.D.)

หัวหน้าภาควิชา

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.Sc.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญจนา วีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ _____ เดือน _____ พ.ศ. _____

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

วิทยานิพนธ์

เรื่อง

การทำเหมืองข้อมูลโรคไข้เลือดออก

Data Mining on Dengue Virus Disease

โดย

นางสาวดารณี ฐิติประยูรวงศ์

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

พ.ศ. 2556

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ดรรชนี ฐิติประยูรวงศ์ 2556: การทำเหมืองข้อมูลโรคไข้เลือดออก ปริญญาวิทยาศาสตร
มหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการ
คอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก:
รองศาสตราจารย์นวลวรรณ สุนทรภิชช์, Ph.D. 177 หน้า

งานวิจัยนี้มีจุดประสงค์เพื่อนำขั้นตอนวิธีการทำเหมืองข้อมูลสกัดองค์ความรู้ที่มี
ความสำคัญในการวินิจฉัยโรคไข้เลือดออก รวมถึงตอบ โจทย์ปัญหาจากแพทย์ผู้เชี่ยวชาญ ข้อมูลที่
ใช้ในการทดลองของงานวิจัยคือข้อมูลการรักษาของผู้ป่วยโรคไข้เลือดออกที่มีอายุไม่เกิน 15 ปี
ในช่วงปี พ.ศ. 2544 ถึงปี พ.ศ. 2552 วิธีการทำเหมืองข้อมูลที่ศึกษามี 3 วิธี คือการจำแนกข้อมูล
ด้วยต้นไม้ตัดสินใจ (Dengue Classification with Decision Tree) การทำนาย the day of
defervescence (day0) ของผู้ป่วยโรคไข้เลือดออกด้วยวิธีการแบบฟัซซี่ลอจิก (Fuzzy Logic) และ
วิธีการโครงข่ายประสาทเทียม (Artificial Neural Network) อีกทั้งการหาความสัมพันธ์ระหว่าง
ปัจจัยที่สนใจด้วยวิธีการหาความสัมพันธ์แบบกฎ (Associate Rule) งานวิจัยนี้วัดประสิทธิภาพ
ของผลลัพธ์ด้วยค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าความถูกต้อง
(Accuracy) ผลการศึกษาพบว่าสามารถค้นพบองค์ความรู้ที่มีความสำคัญต่อการจำแนกชนิดของ
โรคไข้เลือดออกได้ด้วยค่าความถูกต้องคิดเป็น 97.4% สำหรับปัญหาการทำนายวัน day0 สามารถ
ทำนายวัน day0 ของผู้ป่วยโรคไข้เลือดออกได้ค่าความถูกต้องคิดเป็น 69.80% และยังสามารถให้คำตอบ
ของโจทย์ปัญหาอื่นๆ

Daranee Thitiprayoonwongse 2013: Data Mining on Dengue Virus Disease.

Master of Science (Computer Science), Major Field: Computer Science, Department of
Computer Science. Thesis Advisor: Associate Professor Nuanwan Soonthornphisaj,
Ph.D. 177 pages.

This study selected data mining algorithms in order to extract knowledge of Dengue Virus Disease datasets and answered the expert's questions. The datasets used in the study were the clinical data which obtained from patients. They were collected from 2001 to 2009. We selected 3 algorithms in this study. There were decision tree, fuzzy logic and artificial neural network for classification and prediction. In addition, we selected association rule for finding relation between attributes and classes. This study used sensitivity, specificity and accuracy as performance measurement. The experimental results showed the interesting attributes and interesting pattern in dengue classification with 97.4% of accuracy. In addition, we got 69.80% of accuracy in day0 prediction and the answers in expert's questions.

Student's signature

Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ดีด้วยความช่วยเหลืออย่างยิ่งของ รศ.ดร. นवलวรรณ
สุนทรภิชช์ ประธานกรรมการที่ปรึกษา ที่กรุณาช่วยให้คำแนะนำขั้นตอนวิธีในการดำเนินงาน
ชี้แนะแนวทางแก้ไขปัญหาและตรวจสอบแก้ไขข้อบกพร่องในงานวิจัยมาโดยตลอด รวมถึงและ
อาจารย์หมอบ ที่กรุณาช่วยสั่งสอนอบรมทั้งการทำงานและเรื่องอื่นๆมาโดยตลอด

ขอกราบขอบพระคุณอาจารย์นายแพทย์ประพัฒน์ สุริยผล ที่ช่วยให้ข้อมูล ให้คำปรึกษา
และคำแนะนำในการทำวิทยานิพนธ์

ขอกราบขอบพระคุณ คุณพ่อเจียง จูติประยูรวงศ์และคุณแม่ประไพ เพศประเสริฐ ที่ให้
กำลังใจในการเรียน ให้การสนับสนุนและให้ความช่วยเหลืออย่างที่สุดมาโดยตลอดจนสำเร็จ
การศึกษา

ขอขอบพระคุณคณาจารย์ทุกท่านที่ให้การอบรม สั่งสอนวิชาความรู้ต่างๆมาจนสำเร็จ
การศึกษา

ขอขอบคุณบัณฑิตวิทยาลัยและภาควิชาวิทยาการคอมพิวเตอร์ ที่ได้สนับสนุนด้านอุปกรณ์
ด้านเงินทุนนำเสนองานในต่างประเทศ และสถานที่ตลอดระยะเวลาในการทำวิทยานิพนธ์

ขอขอบคุณนายวิสิทธิ์ บุญชุม นายชนพล ตันศิริปรีชา ที่ให้ความช่วยเหลือและให้
คำแนะนำการเขียนวิทยานิพนธ์ รวมถึงขอขอบคุณเพื่อนๆปริญญาโทและพี่ๆปริญญาเอกทุกคนที่
คอยให้กำลังใจ และช่วยเหลืออำนวยความสะดวกในการสอบปากเปล่าขั้นสุดท้าย

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์นี้ ขอมอบแด่ผู้มีพระคุณทุกท่าน

ดารณี จูติประยูรวงศ์

กุมภาพันธ์ 2556

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(4)
คำนำ	1
วัตถุประสงค์	2
การตรวจเอกสาร	4
อุปกรณ์และวิธีการ	35
อุปกรณ์	35
วิธีการ	35
ผลและวิจารณ์	57
ผล	57
วิจารณ์	80
สรุปและข้อเสนอแนะ	86
สรุป	86
ข้อเสนอแนะ	88
เอกสารและสิ่งอ้างอิง	89
ภาคผนวก	92
ภาคผนวก ก แบบฟอร์มการบันทึกข้อมูล	93
ภาคผนวก ข ตารางความหมายของแอททริบิวต์ในฐานข้อมูล	122
ภาคผนวก ค ผลการทดลองกฎความสัมพันธ์	141
ภาคผนวก ง ตารางความสัมพันธ์ของแอททริบิวต์ในชุดข้อมูล	146
ภาคผนวก จ ผลงานตีพิมพ์	151
ประวัติการศึกษา และการทำงาน	177

สารบัญตาราง

ตารางที่	หน้า
1 ขั้นตอนวิธีการเรียนรู้ด้วยต้นไม้ตัดสินใจ	8
2 การวัดค่าประสิทธิภาพทางการแพทย์	25
3 สรุปงานวิจัยที่เกี่ยวข้อง เรียงตามประเภทงานวิจัยตามปีที่ตีพิมพ์	33
4 วัตถุประสงค์และจำนวนแอททริบิวต์ในแต่ละเอกสาร	36
5 แอททริบิวต์ที่ได้จากเอกสารการกรอกข้อมูลผู้ป่วย	40
6 แอททริบิวต์ชนิดคุณภาพที่สร้างเพื่อใช้ในการทดลอง	42
7 แอททริบิวต์ชนิดตัวเลขที่สร้างเพื่อใช้ในการทดลอง	46
8 ค่าความถูกต้องของการทดลองโทยที่ 1	58
9 ค่าความถูกต้องของการทดลองโทยที่ 3	61
10 สัดส่วนการพิจารณาจำนวนคืบโตในแต่ละระดับของโรคไข้เลือดออก	62
11 จำนวนผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่าเอนไซม์ ALT	72
12 สรุปจำนวนผู้ป่วยไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาว	74
13 ประสิทธิภาพของการแบ่งกลุ่มด้วยขั้นตอนวิธีเคมีน	77
14 ค่าความถูกต้องของต้นไม้ตัดสินใจโทยที่ 8	79
15 ค่าสหสัมพันธ์ระหว่างคลาสและปัจจัยที่ปรากฏในผลการทดลองโทยที่ 1	80
16 ค่า False Negative และ False Positive เทียบกับเกณฑ์ของผลการทดลองโทยที่ 1	81
17 ค่า False Negative และ False Positive เทียบกับเกณฑ์ขององค์การอนามัยโลก	82
18 ค่าสหสัมพันธ์ระหว่างคลาสและปัจจัยที่ปรากฏในผลการทดลองโทยที่ 5	83
19 อัตราส่วนค่า False Negative เปรียบเทียบโทยที่ 1 และโทยที่ 8	84
20 อัตราส่วนค่า False Negative เทียบโทยที่ 8 กับเกณฑ์ของ Ibrahim	85
21 ค่าสหสัมพันธ์ระหว่างคลาสและปัจจัยที่ปรากฏในผลการทดลองโทยที่ 8	85

สารบัญตาราง (ต่อ)

ตารางผนวกที่		หน้า
ข1	ตารางความหมายของแอททริบิวต์ไฟล์เอกสาร D01-01	123
ข2	ตารางความหมายของแอททริบิวต์เอกสาร D01-02	123
ข3	ตารางความหมายของแอททริบิวต์เอกสาร D02	127
ข4	ตารางความหมายแอททริบิวต์ตามเอกสาร D03	131
ข5	ตารางความหมายแอททริบิวต์ตามเอกสาร D04	133
ข6	ตารางความหมายแอททริบิวต์ตามเอกสาร D05	135
ข7	ตารางความหมายแอททริบิวต์ตามเอกสาร D06	138
ข8	ตารางความหมายแอททริบิวต์ตามเอกสาร D07	139
ค1	ตารางตัวอย่างผลการหาความสัมพันธ์โจทย์ที่ 5	142
ค2	ตารางตัวอย่างผลการหาความสัมพันธ์โจทย์ที่ 7	143
ง1	ตารางความสัมพันธ์ของแอททริบิวต์เทียบกับคลาสข้อมูลที่ปรากฏบนต้นไม้	147
ง2	ตารางความสัมพันธ์ของแอททริบิวต์เทียบกับคลาสข้อมูลที่ไม่ปรากฏบนต้นไม้	147
ง3	ตารางอธิบายหมายเลขและความหมายของแอททริบิวต์ในภาพผนวกที่ ง1	149

สารบัญภาพ

ภาพที่	หน้า
1 สถาปัตยกรรมของโครงข่ายประสาทเทียม	11
2 ฟังก์ชันซิกมอยด์ (Log-Sigmoid Transfer Function)	12
3 ตัวอย่างข้อมูลคลื่นหัวใจ	21
4 ตัวอย่างฐานข้อมูลชนิด Temporal Database	21
5 ตัวอย่างโปรแกรมเวก้า	25
6 หน้าต่างการเตรียมข้อมูล	26
7 หน้าต่างตัวจำแนกข้อมูล	26
8 ภาพตัวอย่างผลลัพธ์ต้นไม้ตัดสินใจ	27
9 แผนภาพการทำเหมืองข้อมูล	39
10 ต้นไม้ตัดสินใจแสดงปัจจัยที่ช่วยในการจำแนกระดับของโรคไข้เลือดออก	57
11 ต้นไม้ตัดสินใจเฉพาะผู้ป่วยที่มีอาการตับโตในโรคที่ 4 (เพิ่มเติม)	63
12 ต้นไม้ตัดสินใจโรคที่ 5 ตัดผู้ป่วยที่มีค่าเอนโทรปีว่า je เป็น Unknown ออก	68
13 กราฟการกระจายตัวของผลต่างของ AST มากกว่า ALT	73
14 แผนภูมิแสดงอัตราส่วนของข้อมูลแต่ละคลัสเตอร์	76
15 ต้นไม้ตัดสินใจ คลาส Group1 และ Group2	79
ภาพผนวกที่	
ง1 ตารางความสัมพันธ์ระหว่างเอนโทรปีว่าในชุดข้อมูลทดลอง	148

การทำเหมืองข้อมูลโรคไข้เลือดออก

Data Mining on Dengue Virus Disease

คำนำ

โรคไข้เลือดออกเป็นโรคติดต่อที่มีพาหะคือยุงลายตัวเมียที่มีเชื้อไวรัสเดงกี จากการศึกษาพบว่าอุบัติการณ์ของโรคไข้เลือดออกพบมากในภูมิภาคเขตร้อนชื้น โรคไข้เลือดออกมี 2 ชนิด ได้แก่ ไข้เดงกี (Dengue Fever: DF) และ ไข้เลือดออก (Dengue Hemorrhagic Fever: DHF) โดยโรคไข้เลือดออกชนิด DHF แสดงอาการแบ่งได้เป็น 4 ระดับ ได้แก่ ระดับ 1 (DHF grade1: DHF I) เป็นระดับที่มีความรุนแรงน้อยที่สุดของโรคไข้เลือดออก ระดับ 2 (DHF grade2: DHF II) เป็นระยะที่ตรวจพบว่ามีอาการเลือดออก เช่น พบจุดเลือดที่ผิวหนัง หรือถ่ายดำเนื่องจากมีเลือดออกในระบบทางเดินอาหาร ระดับ 3 (DHF grade3: DHF III) เป็นระดับรุนแรง ผู้ป่วยมีอาการช็อก และสามารถเสียชีวิตได้ หากดำเนินไปถึงความรุนแรงระดับ 4 (DHF grade4: DHF IV) ผู้ป่วยมีอาการช็อกอย่างรุนแรง (Profound shock) และการทำงานของอวัยวะหลายระบบล้มเหลวทำให้ผู้ป่วยเสี่ยงต่อการเสียชีวิตสูงมาก ในกรณีการวินิจฉัยของแพทย์คลาดเคลื่อนอาจทำให้เกิดผลกระทบร้ายแรง กล่าวคือ ในกรณีที่ผู้ป่วยเป็นโรคไข้เลือดออกชนิด DHF แต่แพทย์วินิจฉัยคลาดเคลื่อนและผู้ป่วยไม่ได้รับการรักษาอย่างทันท่วงที อาจทำให้อาการของผู้ป่วยพัฒนาความรุนแรงจนถึงแก่ความตายได้

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำข้อมูลผู้ป่วยโรคไข้เลือดออกมาประมวลผลด้วยการทำเหมืองข้อมูลเพื่อหาความสัมพันธ์ของข้อมูลทางคลินิกที่มีความซับซ้อน วิธีการที่ใช้เรียกว่า การทำเหมืองข้อมูล (Data mining) เช่น การจัดกลุ่มข้อมูล (Data clustering) ซึ่งเป็นวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning), การจำแนกประเภทข้อมูล (Data classification) ซึ่งเป็นวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning), การหาความสัมพันธ์ในข้อมูล (Data association) เพื่อค้นหาแอตทริบิวต์ที่สัมพันธ์กัน เป็นต้น

วัตถุประสงค์

งานวิจัยนี้ศึกษาและประมวลผลข้อมูลผู้ป่วยโรคไข้เลือดออก เพื่อตอบโจทย์วิจัยดังนี้

1. แพทย์ต้องการทราบว่าแอทริบิวต์ใด ส่งผลและไม่ส่งผลต่อการจำแนกประเภทของโรคไข้เลือดออก
2. รูปแบบ (pattern) ของแอทริบิวต์ในการจำแนกระดับความรุนแรงของโรคไข้เลือดออกเป็นแบบใด
3. แพทย์ต้องการทราบวันที่ผู้ป่วยมีอาการไข้ลงเพื่อหาวิธีป้องกัน (day0 problem) เนื่องจากวันดังกล่าวอาจสามารถบอกได้ว่าผู้ป่วยรายนั้นมีโอกาสช็อกจนถึงขั้นเสียชีวิตได้
4. อาการดับโตเป็นปัจจัยที่ช่วยในการจำแนกประเภทของไข้เลือดออกได้หรือไม่
5. บุคคลใดที่ได้รับวัคซีนป้องกันโรคไข้สมองอักเสบ (Japanese Encephalitis: JE) แล้วสามารถเป็นโรคไข้เลือดออกได้อีกหรือไม่
6. สำหรับผู้ป่วยโรคไข้เลือดออก ค่าเอนไซม์ Aspartate aminotransferase: AST ซึ่งบ่งบอกการทำงานของตับมีค่าสูงกว่าค่าของเอนไซม์ Alanine aminotransferase: ALT เสมอหรือไม่
7. กลุ่มผู้ป่วยที่เป็นโรคไข้เลือดออกมีค่าความผิดปกติของเม็ดเลือดขาวทุกรายหรือไม่
8. ปัจจัยที่สามารถจำแนกผู้ป่วยที่เป็นโรค DHF I ออกจาก DHF II และ DHF III สามารถค้นหาพบได้หรือไม่ ถ้าพบ อะไรบ้าง

ประโยชน์ที่คาดว่าจะได้รับ

ผลการวิจัยนี้สามารถค้นพบปัจจัยที่ส่งผลต่อการเกิดไข้เลือดออกเพื่อการวินิจฉัยประเภทของโรค และตอบคำถามจากแพทย์ผู้เชี่ยวชาญได้

ขอบเขตและข้อจำกัด

ข้อมูลที่ใช้ในการทำวิทยุกรวบรวมมาจากโรงพยาบาลสงขลานครินทร์ จังหวัดสงขลา และโรงพยาบาลศรีนครินทร์ จังหวัดขอนแก่น ข้อมูลของผู้ป่วยอายุตั้งแต่แรกเกิดถึง 15 ปี จำนวน 1001 คนและข้อมูลของการทำนายวัน day0 จำนวน 968 เรคคอร์ด

การตรวจเอกสาร

ทฤษฎีที่เกี่ยวข้อง

1. การค้นพบความรู้ (Knowledge Discovery)

การค้นหาคำความรู้และความสัมพันธ์ระหว่างข้อมูลที่ซ่อนอยู่ในชุดข้อมูล มีขั้นตอนในการทำงานหลายขั้นตอน ดังต่อไปนี้

1.1. ขั้นตอนการรวมข้อมูล (Integrated Data) ขั้นตอนการรวมข้อมูลจากทุกฐานให้เป็นข้อมูลชุดเดียวเพื่อนำไปประมวลผลตามขั้นตอนวิธีต่อไป

1.2. ขั้นตอนการทำความสะอาดข้อมูล (Cleaning Data) เป็นขั้นตอนที่สำคัญอย่างมากในการทำเหมืองข้อมูล เนื่องจากคุณภาพของการทำเหมืองข้อมูลขึ้นอยู่กับคุณภาพของข้อมูลที่นำเข้ามาประมวลผล ขั้นตอนนี้ประกอบด้วยการจัดการค่าข้อมูลที่บันทึกผิดพลาด (noise) หรือการจัดการค่าของข้อมูลที่หายไป (missing value) โดยวิธีการจัดการกับปัญหาที่ข้อมูลหายไปมีให้เลือกดังนี้

1.2.1. การตัดข้อมูลผู้ปวยรายนั้นออกจากการทำเหมืองข้อมูล

1.2.2. วิธีเติมค่าที่ขาดหายไปให้ถูกต้องโดยอาศัยการตรวจสอบข้อมูลจากต้นฉบับ ซึ่งขั้นตอนนี้ใช้เวลาค่อนข้างมากในการจัดการ

1.2.3. ถ้าค่าข้อมูลที่หายไปเป็นค่าข้อมูลเชิงปริมาณ อาจสามารถแทนค่านั้นได้ด้วยค่าเฉลี่ยของเขตข้อมูลนั้น

1.2.4. วิธีการสมการถดถอย (regression)

1.3. ขั้นตอนการคัดเลือกข้อมูล (Feature selection) ขั้นตอนนี้มีวัตถุประสงค์เพื่อเลือกขอบเขตและคุณลักษณะข้อมูล (แอททริบิวต์) ที่นำมาใช้ในการทำเหมืองข้อมูล

1.4. ขั้นตอนการแปลงข้อมูล (Transformation) ขั้นตอนการแปลงข้อมูลเพื่อปรับข้อมูลให้อยู่ในรูปแบบเดียวกันหรืออาจแบ่งข้อมูลออกเป็นช่วง ยกตัวอย่างเช่น ถ้ามีการเก็บข้อมูลในกระดาษโดยใช้ตัวอักษรแทนข้อมูลของเพศ เช่น F หรือ M แต่ฐานข้อมูลมีค่าเป็น 1 หรือ 2 จึงต้องมีการ

แปลงข้อมูลให้อยู่ในรูปแบบเดียวกันก่อนนำไปประมวลผล การแปลงข้อมูลมีได้หลายรูปแบบ ยกตัวอย่างเช่น

1.4.1. การแบ่งค่าข้อมูลออกเป็นช่วง (Discretization) วิธีนี้ใช้กับข้อมูลเชิงปริมาณที่มีค่าของข้อมูลเป็นจำนวนจริงในกรณีที่ขั้นตอนวิธีการบางวิธีที่ไม่สามารถประมวลผลค่าของข้อมูลที่เป็นค่าต่อเนื่อง

1.4.2. การนอร์มัลไลซ์เขตข้อมูล (Normalization) วิธีนี้มีวัตถุประสงค์เพื่อปรับค่าของข้อมูลให้อยู่ในช่วงหรือขอบเขตข้อมูลเพื่อให้เหมาะสมกับขั้นตอนวิธีที่เลือกใช้

1.5. ขั้นตอนการทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนส่วนหนึ่งของวิธีการค้นหาองค์ความรู้ ขั้นตอนนี้นำข้อมูลเข้ามาประมวลผลด้วยวิธีการที่ต้องการ การทำเหมืองข้อมูลมี 4 ประเภทดังต่อไปนี้

1.5.1. จำแนกประเภทข้อมูล

1.5.2. การจัดกลุ่มของข้อมูล

1.5.3. การจัดความสัมพันธ์ของข้อมูล

1.5.4. การทำนายข้อมูล

ในปัจจุบันได้มีการนำซอฟต์แวร์สำเร็จรูปมาใช้ในขั้นตอนนี้อย่างแพร่หลาย เช่น โปรแกรม MATLAB และ โปรแกรม WEKA เป็นต้น

1.6. ขั้นตอนการประเมินคุณภาพของผลลัพธ์ (Measure Accuracy) เป็นขั้นตอนการประเมินคุณภาพของผลลัพธ์หรือองค์ความรู้ที่ได้ ขั้นตอนนี้ควรมีผู้เชี่ยวชาญเฉพาะด้านของข้อมูลเข้ามามีส่วนร่วมในการตรวจสอบเพื่อเพิ่มความน่าเชื่อถือของผลลัพธ์

1.7. การนำเสนอองค์ความรู้แก่ผู้ใช้ (Present to User) เป็นขั้นตอนการนำเสนอผลลัพธ์ไปยังผู้ใช้หรือผู้เชี่ยวชาญเพื่อนำไปใช้ประโยชน์

2. การทำเหมืองข้อมูล

ขั้นตอนวิธีที่ใช้ในการทำเหมืองข้อมูลมีหลายวิธี (Ho *et al.*, 2003; Ratanamahatana *et al.*, 2005) เช่น การจัดกลุ่มข้อมูล, การจำแนกข้อมูล เป็นต้น

2.1. การจำแนกข้อมูล วิธีการจำแนกข้อมูลมีหลายวิธี เช่น เทคนิคการจำแนกและการสร้างแผนภาพต้นไม้ (CART: Classification and Regression Tree) เพื่อค้นหากฎในชุดข้อมูลโรคไวรัสตับอักเสบและสามารถบอกได้ว่าผู้ป่วยรายนั้นเป็นโรคไวรัสตับอักเสบหรือไม่ (Lin , 2009) อีกทั้งวิธีการของการค้นคืนกรณี (Case Based Reasoning: CBR) เข้ามาช่วยในการแบ่งแยกว่าชนิดของไวรัส ขั้นตอนวิธีที่นิยมใช้ในการจำแนกข้อมูลได้แก่ โครงข่ายประสาทเทียม (Delen *et al.*, 2005) ต้นไม้การตัดสินใจ (Tanner *et al.*, 2008) นาอ์ฟเบย์ เคนเนย์เรสเน-เบอร์ (Ratanamahatana *et al.*, 2005) เป็นต้น

ขั้นตอนวิธีที่เลือกใช้งานวิจัยนี้ได้แก่ ขั้นตอนวิธีต้นไม้ตัดสินใจ ขั้นตอนวิธีการอนุมานฟัซซี่ลอจิก และขั้นตอนวิธีโครงข่ายประสาทเทียม

2.1.1. การเรียนรู้ด้วยต้นไม้ตัดสินใจ

หลักการทำงานของต้นไม้ตัดสินใจ (Decision Tree) คือการเลือกปัจจัยสำคัญหรือแอตทริบิวต์ของข้อมูลที่มีอำนาจในการจำแนกสูงนำมาสร้างเป็น โหนดให้แก่ต้นไม้เพื่อแยกประเภทของข้อมูลโดยให้มีการปะปนกันน้อยที่สุด การเลือกแอตทริบิวต์ดังกล่าวอาศัยค่าความไร้ระเบียบของข้อมูล (Entropy) และค่าเกน (Gain) ของแอตทริบิวต์แต่ละตัว ตามสูตรเอ็นโทรปีหรือสูตรจินีอินเด็กซ์

ก. สูตรเอ็นโทรปี กำหนดให้ข้อมูลมีจำนวน S เรคคอร์ด คลาสของข้อมูลประกอบด้วยคลาสบวกและคลาสลบ สามารถคำนวณระดับการปะปนของข้อมูล S ด้วยสมการ (1)

$$\text{Entropy}(S) = -\frac{P}{P+N} \log_2 \frac{P}{P+N} - \frac{N}{P+N} \log_2 \frac{N}{P+N} \quad (1)$$

โดยที่ S คือ ชุดข้อมูลสอนระบบ
P คือ จำนวนข้อมูลฝึกที่เป็นคลาสบวก
N คือ จำนวนข้อมูลฝึกที่เป็นคลาสลบ

ในกรณีที่ข้อมูลมีมากกว่าสองคลาส สูตรเอ็นโทรปีข้างต้นสามารถเขียนอยู่ในรูปสูตรอย่างง่ายได้ดังนี้

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

โดยที่ p_i คือ สัดส่วนของข้อมูลที่อยู่ในคลาส i เทียบกับข้อมูลทั้งหมด
 C คือ จำนวนคลาสทั้งหมด

ถ้าค่าเอ็นโทรปีมีค่าเป็นศูนย์ หมายถึงข้อมูลมีความเป็นระเบียบคือข้อมูลไม่
 ปะปนกัน ค่าสูงสุดเท่ากับหนึ่ง หมายถึงข้อมูลมีการปะปนกันสูงที่สุดคือมีข้อมูลคลาสละห้าสิบ
 เปอร์เซ็นต์

การวัดความสามารถในการแยกแยะข้อมูลของแอตทริบิวต์สามารถคำนวณ
 ได้จากสูตรเอนตังสมการ (3)

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (3)$$

โดยที่ S คือ ข้อมูลก่อนที่ใช้แอตทริบิวต์ A เป็นตัวแบ่ง
 $|S_v|$ คือ จำนวนเรคคอร์ดที่มีค่าแอตทริบิวต์เป็น A
 $|S|$ คือ จำนวนเรคคอร์ดทั้งหมดในชุดข้อมูลก่อนการแยกแยะ
 $\text{Entropy}(S)$ คือ ค่าความไร้ระเบียบของชุดข้อมูลตอนระบบ S
 $\text{Entropy}(S_v)$ คือ ค่าความไร้ระเบียบของข้อมูลที่ถูกแยกแยะด้วย
 แอตทริบิวต์ A ที่มีค่าเป็น v

ข. สูตรจินีอินเด็กซ์ เป็นสูตรที่นิยมใช้ในการวัดระดับความปะปนของข้อมูล
 สูตรที่ใช้คำนวณแสดงดังสมการ (4)

$$\text{Gini}(S) = 1 - \sum_{i=1}^k p_i^2 \quad (4)$$

โดยที่ p_i คือ สัดส่วนข้อมูลคลาส i เทียบกับข้อมูลทั้งหมด
 ถ้าค่าจินีอินเด็กซ์มีค่าน้อยคือข้อมูลมีความปะปนน้อย ถ้ามีค่าเท่ากับ 1 คือปะปนกันมาก
 การวัดความสามารถในการแยกแยะข้อมูลของแอตทริบิวต์ใช้สูตรเอนตังสมการ (5)

$$\text{Gain}(S, A) = \text{Gini}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \left(1 - \sum_{i=1}^k p_i^2\right) \quad (5)$$

ค่าเอนถ้ามีค่าสูงแสดงว่ามีอำนาจการจำแนกสูง ถ้าแอตทริบิวต์ใดมีค่าสูงที่สุด แอตทริบิวต์
 นั้นถูกเลือกเป็นโหนดของต้นไม้

ตารางที่ 1 ขั้นตอนวิธีการเรียนรู้ด้วยต้นไม้ตัดสินใจ

Algorithm DecisionTree (examples, attributes)

```

if all example have the same classification then
    return the classification
else if attributes is empty then
    return Mode (examples)
bestAttribute = SelectBestAttribute (attributes, examples)
tree = a new decision tree with root node best
for each value  $v_i$  of best do
    examplesi = {element of example with best =  $v_i$ }
    subtree = DecisionTree (examplesi, attributes-best)
    add a branch to tree with label  $v_i$  and subtree tree
return tree
  
```

ขั้นตอนวิธีการพื้นฐานของการสร้างต้นไม้ช่วยตัดสินใจ คือ ขั้นตอนวิธีเชิงละโมภ (Greedy Algorithm) โดยการสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (Recursive) ด้วยวิธีการแบ่งปัญหาใหญ่ให้เป็นปัญหาย่อย (Divide-And-Conquer) วิธีการทำงานคือ ขั้นแรกต้นไม้จะสร้างโหนดแรกด้วยการทดสอบอำนาจการแบ่งแยกข้อมูล ถ้าข้อมูลมีค่าข้อมูลเดียวกันทั้งหมด โหนดนั้นกลายเป็นโหนดใบ แต่ถ้าไม่ใช่จะมีการคำนวณค่าเกณฑ์ของแต่ละแอททริบิวต์เก็บไว้ ต่อมาขั้นตอนวิธีจะเลือกแอททริบิวต์ที่มีค่าเกณฑ์สูงที่สุด เพื่อใช้แอททริบิวต์นั้นเป็นโหนดที่ใช้แบ่งแยกข้อมูล หลังจากนั้นสร้างกิ่งของต้นไม้และแบ่งข้อมูลออกตามคลาส งานการทำงานอย่างนี้ไปเรื่อยๆจนกว่าไม่สามารถแบ่งความแตกต่างของข้อมูลได้ หลังจากนั้นแสดงผลลัพธ์สุดท้ายออกมาในรูปแบบของต้นไม้ตัดสินใจ

2.1.2. ฟัซซี่ลอจิก (Fuzzy Logic)

ฟัซซี่ลอจิก หรือ ตรรกะความคลุมเครือมีรากฐานมาจากทฤษฎีบทของเซตวัตถุประสงค์เพื่อแก้ปัญหของเซตแบบดั้งเดิมที่มีเพียงค่าถูกหรือผิดอย่างเดียว แต่ในฟัซซี่เซตเป็นการให้ค่าระดับความเป็นสมาชิก (degree of membership - μ) โดยมีค่าตั้งแต่ 0 ถึง 1 หรือเขียนเป็น

สัญลักษณ์คือ $[0,1]$ ยกตัวอย่างเช่น การเปิดน้ำร้อน ผู้ใช้จะรู้ว่าต้องปรับอุณหภูมิในการเปิดน้ำมากหรือน้อยต้องวัดความร้อนของน้ำก่อน แต่การที่จะบอกได้น้ำร้อนหรือเย็นนั้นเป็นเพียงการประมาณของแต่ละบุคคล ซึ่งต่างคนอาจมีเกณฑ์ที่ต่างกันขึ้นกับประสบการณ์การเรียนรู้ของแต่ละคน ซึ่งเรียกเหตุการณ์ที่ไม่แน่นอนนี้ว่า ฟัซซีเซต (Fuzzy Set)

องค์ประกอบในการสร้างฟัซซีลอจิกประกอบด้วย 3 ส่วนดังนี้

ก. ฟังก์ชันการเป็นสมาชิก (Membership Function) คือ ฟังก์ชันที่ใช้ในการกำหนดค่าความเป็นสมาชิกของเอททริบิวต์ซึ่งค่าความเป็นสมาชิกขึ้นอยู่กับรูปแบบของกราฟ ฟังก์ชันการเป็นสมาชิก ซึ่งโดยทั่วไปมีหลายชนิดด้วยกันยกตัวอย่างเช่น ฟังก์ชันความเป็นสมาชิกรูปสามเหลี่ยม (Triangular Membership Function) ฟังก์ชันความเป็นสมาชิกรูปเกาส์เซียน (Gaussian membership function) ฟังก์ชันความเป็นสมาชิกแบบรูปประฆังคว่ำ (Bell-Shaped membership function) เป็นต้น การเลือกฟังก์ชันความเป็นสมาชิกมาใช้งานขึ้นอยู่กับลักษณะของข้อมูลที่น่ามาประมวลผล

ข. ตัวดำเนินการฟัซซีลอจิก มี 3 ตัวดำเนินการคือ คำสั่งยูเนียน (Union) อินเตอร์เซกชัน (Intersection) และคำสั่งคอมพลีเมนต์ (Complement) ซึ่งทั้งสามคำสั่งมีการทำงานเหมือนคำสั่งในเซตทั่วไป

ค. กฎฟัซซี (Fuzzy Rule) ขั้นตอนนี้เป็นส่วนสำคัญสำหรับการตัดสินใจเลือกข้อมูล ซึ่งกฎต่างๆ ในขั้นตอนี้ต้องได้รับความร่วมมือจากผู้เชี่ยวชาญในการใส่กฎลงในฟัซซีลอจิก โดยใส่กฎลงไปในรูปแบบของคำสั่งเงื่อนไข ถ้า...แล้ว (If...then)

กระบวนการทำงานของฟัซซีลอจิกมีขั้นตอน 4 ขั้นตอนคือ

(1) Fuzzification : ขั้นตอนการแปลงข้อมูลเข้าให้เป็นค่าระดับความเป็นสมาชิก โดยค่าที่ได้ขึ้นกับฟังก์ชันความเป็นสมาชิกที่เลือกใช้

(2) Rule Inference : ขั้นตอนการสร้างความสัมพันธ์ระหว่างข้อมูลนำเข้าทั้งหมดด้วยคำสั่งและ (and) หรือ คำสั่งหรือ (or)

(3) Rule Composition : การนำค่าความเชื่อมั่นของกฎที่เกี่ยวข้องมาหาข้อสรุป หรือเปรียบได้กับการนำพื้นที่ใต้กราฟจากสองขั้นตอนด้านบนมาหาผลลัพธ์ด้วยการอินเตอร์เซกกัน

(4) Defuzzification : ขั้นตอนการแปลงระดับความเป็นสมาชิกกลับสู่ความเชื่อมั่น วิธีการหนึ่งที่ใช้คือการหาจุดศูนย์กลางของพื้นที่ใต้กราฟซึ่งเป็นจุดที่แบ่งพื้นที่ใต้กราฟ ออกเป็นสองส่วนเท่าๆกัน ซึ่งมีสูตรในการหาดังนี้

$$COG = \frac{\sum_{i=1}^N \alpha_i w_i}{\sum_{i=1}^N \alpha_i} \quad (6)$$

โดยที่ COG คือ จุดศูนย์กลางของพื้นที่ใต้กราฟ

N คือ ค่าตั้งแต่ตำแหน่งที่ 1 ถึงตำแหน่งที่ i

α_i คือ ค่าความเป็นสมาชิกซึ่งมาจากฟังก์ชันที่เลือก ณ ตำแหน่งที่ i

w_i คือ พื้นที่ใต้กราฟของเซตฟัซซี่ตำแหน่งที่ i

2.1.3. โครงข่ายประสาทเทียม (Artificial Neural Networks)

โครงข่ายประสาทเทียมเป็นวิธีการที่ถูกสร้างขึ้นเลียนแบบการทำงานของสมองมนุษย์ มีคุณสมบัติเด่นในด้านการเรียนรู้และการจดจำ ยกตัวอย่างเช่นสมองมนุษย์นั้นเมื่อได้เรียนรู้ บางอย่างหลายๆครั้ง ก็ทำให้เกิดการจดจำและเมื่อพบเห็นสิ่งที่ไม่เคยได้เรียนรู้มาก่อนก็สามารถ อนุมานได้ว่าสิ่งนั้นคืออะไรจากการเรียนรู้ก่อนหน้านี้

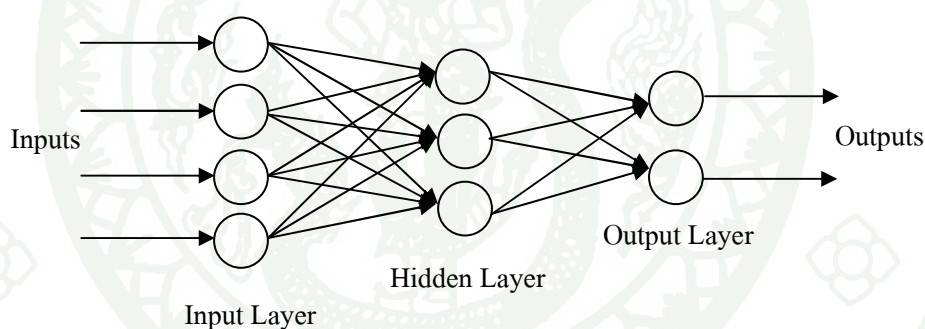
การเรียนรู้ของโครงข่ายประสาทเทียมคือ การคำนวณข้อมูลเอาต์พุต (output) พร้อมปรับปรุงค่าน้ำหนักโดยใช้ข้อมูลอินพุต (input) หรือข้อมูลนำเข้าที่ป้อนให้กับโครงข่าย สามารถแบ่งเป็น 3 ประเภทคือ

(1) การเรียนรู้แบบมีผู้สอน (supervised learning) ข้อมูลที่ใช้เรียนรู้ต้องมีอินพุต และเอาต์พุตเพื่อการเรียนรู้ โดยโครงข่ายจะเรียนรู้ข้อมูลทั้งสองโดยปรับค่าน้ำหนักเพื่อลดความแตกต่างหรือค่าความผิดพลาดให้มีค่าน้อยที่สุด การปรับค่าน้ำหนักจะปรับทีละน้อยโดย กระบวนการทำซ้ำกับข้อมูลที่ระบุ จนกระทั่งค่าน้ำหนักในโครงข่ายลู่เข้า ซึ่งในงานวิจัยนี้ใช้การ เรียนรู้แบบมีผู้สอนเท่านั้น

(2) การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) ข้อมูลที่ใช้ในการเรียนรู้เป็นเพียงข้อมูลอินพุตเท่านั้น กระบวนการเรียนรู้ใช้หลักทางสถิติ และจัดกลุ่มข้อมูลออกเป็นระดับต่างๆ โดยโครงข่ายหาค่าเอาต์พุตเองจากข้อมูลอินพุต

(3) การเรียนรู้เชิงบังคับ (reinforcement learning) เป็นการเรียนรู้แบบมีผู้สอน และไม่มีผู้สอน โดยใช้การเรียนรู้แบบไม่มีผู้สอนในระหว่างการสอนที่มีเพียงชุดข้อมูลอินพุต และใช้การเรียนรู้แบบมีผู้สอนเพื่อบอกว่าถูกหรือผิดแต่ไม่บอกว่าเอาต์พุตที่ถูกต้องคืออะไร

วิธีการโครงข่ายประสาทเทียมที่ใช้ในงานวิจัยนี้คือ เพอร์เซปตรอนแบบหลายชั้น (Multilayer Perceptron) ซึ่งวิธีการนี้มีข้อดีคือมีการทำงานแบบไม่เชิงเส้น (non-linear) ที่มีความเหมาะสมต่อการแก้ปัญหาที่ไม่ชัดเจน การทำงานของวิธีการประกอบด้วยหนึ่งชั้นอินพุต (Input Layer) หนึ่งชั้นเอาต์พุต (Output Layer) และอย่างน้อยหนึ่งชั้นซ่อน (Hidden Layer) แต่ละชั้นเชื่อมต่อกันดังภาพที่ 1



ภาพที่ 1 สถาปัตยกรรมของโครงข่ายประสาทเทียม

ขั้นตอนการเรียนรู้โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP) มีการเรียนรู้แบบแพร่ย้อนกลับ (Backpropagation Learning) และมีขั้นตอนการทำงานดังนี้

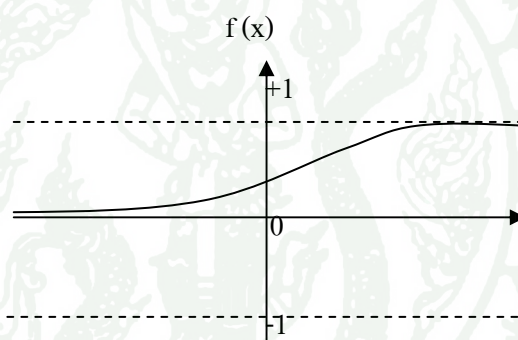
- (1) กำหนดจำนวน โหนดข้อมูลเข้า จำนวนโหนดซ่อน จำนวนโหนดผลลัพธ์ และจำนวนรอบสูงสุดที่เรียนรู้
- (2) สุ่มค่าน้ำหนักเริ่มต้นให้กับทุกๆเส้นเชื่อมภายในโครงข่ายประสาทเทียมมีค่าอยู่ในช่วง $[-1,1]$
- (3) รับค่าข้อมูลเข้าของข้อมูลแถวแรก เพื่อใช้ในการคำนวณหาค่าผลลัพธ์ของโครงข่ายประสาทเทียม

(4) คำนวณค่าผลรวมของโหนดซ่อน (S_m) ด้วยสมการ (7) จากนั้นปรับค่ารวมด้วยฟังก์ชันกระตุ้น (Activation Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) ดังสมการ (8) ซึ่งได้ค่าโหนดซ่อน (y_m) หลังการปรับค่าดังสมการ (9) โดยค่าที่ได้อยู่ในช่วง $[0,1]$ ดังภาพด้านล่าง

$$S_m = \sum_{i=1}^N \alpha_i w_i \quad (7)$$

$$f(x) = \frac{1}{1 + e^{-x}} \quad (8)$$

$$y_m = f(s_m) \quad (9)$$



ภาพที่ 2 ฟังก์ชันซิกมอยด์ (Log-Sigmoid Transfer Function)

(5) คำนวณค่าผลรวมของโหนดผลลัพธ์ จากนั้นปรับค่าผลรวมด้วยฟังก์ชันกระตุ้น (Activation Function)

(6) คำนวณค่าความผิดพลาดของค่าของโหนดผลลัพธ์

(7) คำนวณค่าความผิดพลาดของโหนดซ่อน

(8) ปรับค่าน้ำหนักของเส้นเชื่อมระหว่างโหนดซ่อนและโหนดผลลัพธ์จากค่าความผิดพลาดของโหนดผลลัพธ์ โดยใช้ค่าอัตราการเรียนรู้ที่มีค่าอยู่ในช่วง $[0,1]$ ช่วยในการคำนวณ

(9) ปรับค่าน้ำหนักของเส้นเชื่อมระหว่างโหนดข้อมูลเข้าและโหนดซ่อนจากค่าความผิดพลาดของโหนดซ่อน

(10) รับค่าข้อมูลเข้าของข้อมูลแถวถัดไปและกลับไปทำข้อ 4 แต่ถ้าเป็นข้อมูลแถวสุดท้ายให้ไปทำข้อ 11

(11) คำนวณค่าความผิดพลาดเฉลี่ย (Mean Squared Error: MSE) ในทุกแถวของข้อมูล ถ้าค่าความผิดพลาดเฉลี่ยมีค่าน้อยกว่าค่าที่ยอมรับได้ที่กำหนดไว้ ให้จบการเรียนรู้ แต่ถ้าค่าความผิดพลาดเฉลี่ยมากกว่าค่าที่ยอมรับได้ ให้ตรวจสอบจำนวนรอบของการเรียนรู้ที่กำหนดไว้ ถ้าครบแล้วให้จบการเรียนรู้ แต่ถ้ายังไม่ครบให้กลับไปทำข้อ 3 ใหม่อีกครั้ง

2.2. การจัดกลุ่มข้อมูลหรือคลัสเตอร์ กลุ่มข้อมูลที่ศึกษาประกอบด้วยข้อมูลที่มีคุณสมบัติเหมือนกันหรือคล้ายคลึงกัน ดังนั้นข้อมูลที่อยู่ต่างกลุ่มข้อมูลย่อมมีความแตกต่าง การจัดกลุ่มข้อมูลเป็นกระบวนการที่สามารถนำไปใช้เพื่อค้นพบความรู้ในกลุ่มข้อมูล และขั้นตอนวิธีการจัดกลุ่มเป็นการเรียนรู้จากข้อมูลฝึกที่ไม่ต้องมีคำเฉลย

ในปัจจุบันขั้นตอนวิธีที่ใช้ในการจัดกลุ่มมีหลายวิธี วิธีที่มีผู้ใช้นิยมอย่างแพร่หลายคือ ขั้นตอนวิธีเคมีน ซึ่งมีรายละเอียดดังนี้

2.2.1. ขั้นตอนวิธีเคมีน เป็นขั้นตอนที่ได้รับความนิยมมากเพราะเป็นขั้นตอนวิธีที่ไม่ซับซ้อน ค่า k คือ จำนวนกลุ่ม ข้อมูล means หมายถึงค่าเฉลี่ยซึ่งถูกใช้เป็นตัวแทนของกลุ่มข้อมูลหรือเป็นศูนย์กลางของคลัสเตอร์หนึ่งๆ

หลักการทำงานคือ ผู้ใช้กำหนดค่าเริ่มต้น k คือ ระบุจำนวนกลุ่มคลัสเตอร์ที่ต้องการ จากนั้นขั้นตอนวิธีเคมีนเริ่มทำงานดังนี้

- (1) สร้างกลุ่มข้อมูลที่ 1 โดยสุ่มเลือกเรคคอร์ดเป็นศูนย์กลางของกลุ่มข้อมูลที่หนึ่ง
- (2) สร้างกลุ่มข้อมูลจนครบ k กลุ่ม โดยแต่ละกลุ่มมีข้อมูลศูนย์กลางของกลุ่มนั้น
- (3) วัดระยะห่างด้วยวิธีการหาระยะห่างแบบยูคลิดีเนียนระหว่างข้อมูลทั้งหมดกับศูนย์กลางของกลุ่มที่หนึ่ง $d(i, \text{centroid}_1)$ และศูนย์กลางของกลุ่มที่สอง $d(i, \text{centroid}_2)$ ไปเรื่อยๆ จนครบ k กลุ่ม กำหนดคลัสเตอร์ให้แก่ข้อมูลเหล่านี้โดยใช้ค่าระยะห่างที่น้อยกว่าเป็นตัวกำหนด
- (4) เมื่อข้อมูลที่เหลือทั้งหมดได้รับการกำหนดความเป็นสมาชิกให้แก่ทั้งสองกลุ่มข้อมูลแล้ว คำนวณค่าศูนย์กลางของกลุ่มข้อมูลทั้งสองอีกครั้ง สมมติให้กลุ่มข้อมูลประกอบด้วยสมาชิก n จำนวน ข้อมูลแต่ละเรคคอร์ดประกอบด้วยแอตทริบิวต์ สำหรับเรคคอร์ดที่ i ใดๆ คือ $(x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})$ ค่าศูนย์กลางสามารถคำนวณจากค่าเฉลี่ยของแต่ละแอตทริบิวต์ตามสูตรในสมการ (10)

$$\text{Centroid}_k = (\text{Mean}_{x_1}, \text{Mean}_{x_2}, \dots, \text{Mean}_{x_n}) \quad (10)$$

(5) กำหนดกลุ่มข้อมูลให้กับข้อมูลทั้งหมดโดยการทำซ้ำกระบวนการที่ 3 ถึง 5 จนกว่าค่ากลุ่มของข้อมูลที่ระบุนั้นไม่มีการเปลี่ยนแปลง

ข้อดีของขั้นตอนวิธีการเคมิน

1) วิธีการนี้ใช้เวลาในการทำงานเป็น $O(tkn)$ โดยที่ n คือจำนวนข้อมูลที่ต้องการจัดกลุ่ม k คือจำนวนกลุ่มข้อมูลที่ต้องการ และ t คือจำนวนรอบที่มีการตรวจสอบความเป็นสมาชิกของกลุ่มข้อมูล

2) เงื่อนไขในการจัดกลุ่มข้อมูลของขั้นตอนเคมินคือ เมื่อการจัดกลุ่มเสร็จสิ้นต้องพบว่าข้อมูลทุกตัวอยู่ในกลุ่มข้อมูลที่เหมาะสม โดยวัดจากค่าระยะห่างระหว่างสมาชิกข้อมูลนั้นกับศูนย์กลางของกลุ่มข้อมูลต้องเป็นระยะทางที่สั้นที่สุดเมื่อเทียบกับศูนย์กลางของกลุ่มข้อมูลอื่นๆ ดังสูตร mean square error สมการ (11)

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (11)$$

โดยที่

E	หมายถึง ค่าความผิดพลาด
k	หมายถึง จำนวนคลัสเตอร์
p	หมายถึง ข้อมูลที่เป็นสมาชิกของคลัสเตอร์
C_i	หมายถึง คลัสเตอร์กลุ่มที่ i
m_i	หมายถึง ศูนย์กลางของคลัสเตอร์ที่ i

ข้อเสียของขั้นตอนวิธีการเคมิน

1) ขั้นตอนวิธีการเคมินไม่สามารถใช้กับข้อมูลเชิงคุณภาพได้เพราะทำให้เกิดความไม่สมเหตุสมผล

2) ผู้ใช้ต้องกำหนดกลุ่ม (ค่า k) ซึ่งในบางกรณีผู้ใช้ไม่สามารถทราบจำนวนกลุ่มของข้อมูลก่อนล่วงหน้า

3) วิธีการนี้ไม่สามารถจัดการกับข้อมูลที่มีค่าของข้อมูลที่แตกต่างกันมาก ส่งผลให้การจัดกลุ่มมีประสิทธิภาพที่ไม่ดี

4) วิธีการนี้สามารถทำงานได้ดีในกรณีที่ข้อมูลมีการกระจายตัวรูปร่างเป็นวงกลม แต่ถ้าเป็นรูปร่างอื่นๆ อาจทำให้คุณภาพในการจัดกลุ่มลดลง

การวัดประสิทธิภาพของการแบ่งกลุ่ม ผู้วิจัยได้วัดประสิทธิภาพด้วยการวัดค่า

Entropy และค่า Purity (Hanburg, 2008; Cambridge University, 2008)

ค่า Entropy คือการวัดค่าความปะปนของคลัสเตอร์ที่แบ่งได้ เทียบกับคลาสของข้อมูล ถ้าค่าที่คำนวณได้มีค่าเข้าใกล้ศูนย์หมายถึงมีความปะปนน้อย สูตรการคำนวณค่า Entropy มีดังนี้

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^q \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \quad (12)$$

เมื่อ S_r คือ คลัสเตอร์กลุ่มที่ r
 q คือ จำนวนคลาสในชุดข้อมูล
 n_r^i คือ จำนวนข้อมูลคลาส i ที่อยู่ในคลัสเตอร์กลุ่มที่ r

$$\text{Entropy} = \sum_{r=1}^k \frac{n_r}{n} E(S_r) \quad (13)$$

เมื่อ k คือ จำนวนคลัสเตอร์

ค่า Purity คือการวัดค่าคลาสส่วนใหญ่ที่ปรากฏอยู่จริงในคลัสเตอร์ที่แบ่งได้ ถ้าค่าที่คำนวณได้มีค่ามากหมายถึงผลลัพธ์การแบ่งกลุ่มดีกว่า สูตรการคำนวณค่า Purity มีดังนี้

$$\text{Pu}(S_r) = \frac{1}{n_r} \max_i n_r^i \quad (14)$$

เมื่อ S_r คือ คลัสเตอร์กลุ่มที่ r
 n_r^i คือ จำนวนข้อมูลคลาส i ที่อยู่ในคลัสเตอร์กลุ่มที่ r

$$\text{Purity} = \sum_{r=1}^k \frac{n_r}{n} \text{Pu}(S_r) \quad (15)$$

เมื่อ k คือ จำนวนคลัสเตอร์ที่ต้องการแบ่ง

2.3. การหาความสัมพันธ์ของข้อมูล

ข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่อาจมีความสัมพันธ์บางอย่างซ่อนอยู่ วิธีการหาความสัมพันธ์ของข้อมูลสามารถทำได้หลายวิธียกตัวอย่างเช่น การหาความสัมพันธ์ และการหาค่าสหสัมพันธ์ เป็นต้น

2.3.1. การหากฎความสัมพันธ์ (Association Rule)

วัตถุประสงค์ของวิธีการเพื่อค้นหาความรู้และความสัมพันธ์ของข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่ ส่วนใหญ่มักนำผลลัพธ์ไปใช้ในการทำธุรกิจ ยกตัวอย่างเช่นมีการค้นพบว่า ถ้าซื้อหอมหัวใหญ่และมันฝรั่งแล้วผู้ซื้อน่าจะมีการซื้อเนื้อวัวด้วย

การสร้างกฎความสัมพันธ์ประกอบด้วย 2 กระบวนการดังนี้

(1) ขั้นตอนการค้นหาข้อมูลที่เกิดร่วมกันบ่อยครั้งในฐานข้อมูล (Frequent itemsets) ซึ่งเป็น Itemsets ที่มีความถี่หรือค่าสนับสนุน (Support) มากกว่าหรือเท่ากับค่าความถี่ขั้นต่ำ (Minimum Support) ที่กำหนด

(2) ขั้นตอนสร้างกฎความสัมพันธ์จาก Frequent itemsets จากผลขั้นตอนที่ 1 ขอมรับกฎความสัมพันธ์ที่สร้างขึ้นก็ต่อเมื่อกฎที่สร้างมีค่าความเชื่อมั่น (Confidence) มากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ (minimum confidence) ที่กำหนด

ขั้นตอนที่นิยมเลือกใช้ได้แก่ ขั้นตอนวิธี Apriori ซึ่งมีขั้นตอนการทำงานดังนี้

1. สร้าง Candidate Itemset ขึ้นมาให้มีขนาดเท่ากับ k-itemsets โดยที่ $k = \{1, 2, \dots, n\}$ โดยที่ Candidate Itemsets ของระดับ k ได้มาจากการรวมระหว่าง Frequent Itemsets ในระดับที่ k-1
2. หา Frequent Itemsets ในระดับ k ด้วยการนับความถี่ (support) ของ Candidate Itemsets แต่ละตัวในระดับที่ k
3. ตรวจสอบ Candidate Itemsets ถ้าตัวใดมีความถี่มากกว่าหรือเท่ากับค่าความถี่ขั้นต่ำ (minimum support) ถือว่าตัวนั้นมีคุณสมบัติเป็น Frequent Itemsets ของระดับ k
4. สร้าง Candidate Itemset หา Frequent Itemsets ในระดับที่ k+1 ทำเช่นนี้ไปเรื่อยๆจนกว่าไม่สามารถสร้าง Candidate Itemsets ในระดับที่ k+1 ได้

โปรแกรมเวก้ามีขั้นตอนวิธี Apriori ในส่วนของ Association ค่าพารามิเตอร์ในโปรแกรมเวก้าที่กำหนดมีค่าระหว่าง 0 ถึง 1 โดยทั่วไปนิยมกำหนดค่าสนับสนุนขั้นต่ำ

(lowerBoundMinSupport) ประมาณไม่เกิน 0.25 ในขณะที่ค่าความเชื่อมั่นขั้นต่ำ (minMetric) จะตั้งค่าสูงคือมากกว่า 0.8 ถึง 1 (นวลวรรณ, 2553)

2.3.2. การคำนวณค่าสหสัมพันธ์ (Correlation)

ค่าสหสัมพันธ์ (Correlation) เป็นสถิติที่ใช้หาความสัมพันธ์ระหว่างตัวแปร ค่าสหสัมพันธ์ที่คำนวณได้เรียกว่า ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ซึ่งการคำนวณค่ามีหลายแบบ ขึ้นอยู่กับเงื่อนไขของสิ่งที่เราต้องการหา

ในการวิเคราะห์ความสัมพันธ์ระหว่างสองตัวแปร ถ้าหากสองตัวแปร มีระดับการวัดอันดับ (Interval scale) หรืออัตราส่วน (Ratio scale) เรียกว่าการวิเคราะห์โดยใช้พารามเมตริก (Parametric procedure) แต่ถ้ามีระดับการวัดมาตราแบบบัญญัติ (Nominal scale) หรือมาตราเรียงอันดับ (Ordinal scale) เรียกว่าการวิเคราะห์แบบไม่ใช้พารามเมตริก (Nonparametric procedure) ในงานวิจัยนี้เลือกใช้การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's Correlation) หรือบางครั้งเรียกว่า สหสัมพันธ์อย่างง่าย (Simple Correlation) ซึ่งเป็นการทดสอบความสัมพันธ์แบบเชิงเส้น วิธีการนี้สามารถบอกความสัมพันธ์ระหว่างตัวแปร 2 ตัวได้ แต่ไม่สามารถบอกเหตุและผลได้ กล่าวคือ ไม่สามารถบอกได้ว่าตัวแปรใดเกิดก่อนหรือหลัง ค่าสหสัมพันธ์เพียร์สัน (r) ที่คิดได้มีค่าระหว่าง -1 ถึง 1 สามารถคำนวณได้จาก

$$r_{xy} = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \quad (16)$$

เมื่อ	r_{xy}	เป็นค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน
	$\sum X$	เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)
	$\sum Y$	เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)
	$\sum XY$	เป็นผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2
	$\sum X^2$	เป็นผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1
	$\sum Y^2$	เป็นผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2
	N	เป็นขนาดของกลุ่มตัวอย่าง

ทิศทางความสัมพันธ์ระหว่างตัวแปรมี 3 แบบคือ

(1) สหสัมพันธ์ทางบวก (Positive Correlations) คือเมื่อตัวแปรหนึ่งเพิ่ม ตัวแปรอีกตัวหนึ่งเพิ่มด้วย (ค่า r มีค่ามากกว่า 0 แต่น้อยกว่า 1)

(2) สหสัมพันธ์ทางลบ (Negative Correlations) คือเมื่อตัวแปรหนึ่งเพิ่ม ตัวแปรอีกตัวหนึ่งลด (ค่า r มีค่าน้อยกว่า 0 แต่มากกว่า -1)

(3) สหสัมพันธ์เป็นศูนย์ (Zero Correlations) หมายถึงตัวแปรสองตัวไม่มีความสัมพันธ์กัน (ค่า r เท่ากับ 0)

ถ้าตัวแปรสองตัวมีความสัมพันธ์กันแล้ว สามารถบอกระดับความสัมพันธ์ตามค่าสัมประสิทธิ์ที่คิดได้ดังนี้ (Hinkle D.E. 1998, p.118)

ค่า $ r $	ระดับของความสัมพันธ์
0.90 - 1.00	มีความสัมพันธ์กันในระดับสูงมาก
0.70 - 0.90	มีความสัมพันธ์กันในระดับสูง
0.50 - 0.70	มีความสัมพันธ์กันในระดับปานกลาง
0.30 - 0.50	มีความสัมพันธ์กันในระดับต่ำ
0.00 - 0.30	มีความสัมพันธ์กันในระดับต่ำมาก

ลักษณะของสมมติฐานที่ทดสอบว่าตัวแปรสองตัวมีความสัมพันธ์กันหรือไม่ ใช้สัญลักษณ์ภาษากรีกคือ ρ (rho) แทน r ซึ่งสามารถเขียนเป็นสมมติฐานทางสถิติได้ดังนี้

$H_0: \rho = 0$ หมายถึง ตัวแปรสองตัวไม่มีความสัมพันธ์กัน

$H_1: \rho \neq 0$ หมายถึง ตัวแปรสองตัวมีความสัมพันธ์กัน

สถิติทดสอบ เป็นการทดสอบแบบสองทาง

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, df = n-2 \quad (17)$$

อาณาเขตวิกฤตและการสรุปผลจะปฏิเสธ H_0 เมื่อค่า t ที่คำนวณได้มีค่ามากกว่าหรือเท่ากับค่า $t_{\alpha, n-2}$ ที่เปิดจากตาราง หรือถ้าค่า t ที่คำนวณได้มีค่าน้อยกว่าหรือเท่ากับค่า $-t_{\alpha, n-2}$ หรือนำค่า r_{xy} ที่ได้เปรียบเทียบกับค่าวิกฤต r_{xy} จากตารางสำเร็จรูปโดยใช้ $df = n-2$ ถ้าค่า r_{xy} มีค่ามากกว่าหรือเท่ากับค่าวิกฤต r_{xy} จะปฏิเสธ H_0

2.4. การทำนายข้อมูล ขั้นตอนวิธีนี้มีวัตถุประสงค์เพื่อการทำนายผลหรือสิ่งที่จะเกิดขึ้นในอนาคตเพื่อหาวิธีการรักษาหรือป้องกัน จากงานวิจัยที่ศึกษาพบหลายเทคนิคเช่น ต้นไม้ตัดสินใจ เป็นต้น งานวิจัยของ Tanner (Tanner *et al.*, 2008; Lavrac N., 1999) ใช้วิธีต้นไม้ตัดสินใจเพื่อทำนายว่าผู้ป่วยเป็นโรคไข้เลือดออกชนิดไม่ร้ายแรง หรือชนิดร้ายแรงเพื่อบอกว่าควรให้ผู้ป่วยรายนั้นรักษาที่โรงพยาบาลต่อหรือไม่

3. โรคไข้เลือดออก

ปัจจุบันโรคที่พบมากในพื้นที่เขตร้อนชื้นคือโรคไข้เลือดออกโดยพาหะนำโรคคือ ยุงลาย (*Aedes aegyti*) ตัวเมีย กัดผู้ป่วยที่เป็นโรคไข้เลือดออกโดยเฉพาะช่วงที่มีไข้สูง เชื้อโรคจะเพิ่มอยู่ในตัวยุงประมาณ 8-10 วัน เชื้อไวรัสเดงกีจะอยู่ที่ผนังกระเพาะและต่อมน้ำลายของยุง ดังนั้นเมื่อยุงกัดคนจึงทำให้แพร่เชื้อสู่คนได้ และในปัจจุบันยังไม่มีวัคซีนป้องกัน โรคไข้เลือดออกแบ่งออกเป็นสองประเภทใหญ่ๆ ได้แก่ ไข้เดงกีและไข้เลือดออก ไข้เลือดออกนั้นสามารถแยกย่อยออกได้เป็น 4 ระดับคือ ระดับ 1 เป็นขั้นที่ตรวจพบว่ามีเชื้อ ระดับ 2 เป็นขั้นที่ตรวจพบว่ามีจุดเลือดที่ผิวหนัง ระดับ 3 เป็นระดับที่พบว่ามีอาการเปลี่ยนแปลงความดันอย่างรวดเร็วและมีช่วงความแตกต่างของความดัน (pulse pressure) น้อยกว่า 20 มิลลิเมตรปรอท โดยมีอาการแสดงคือผิวหนังเย็นและกระสับกระส่าย และระดับ 4 ระดับนี้จะมีอาการช็อก สังเกตได้จากไม่สามารถวัดค่าความดันเลือดได้ ซึ่งในบางครั้งเรียกระดับ 3 และ 4 รวมกันว่า DSS (Dengue Shock Syndrome) โดยอาการของโรคแบ่งเป็น 3 ระยะคือ

(1) ระยะไข้สูง ผู้ป่วยมีอาการเบื่ออาหาร อาจอาเจียน ปวดศีรษะมีไข้สูง 38-40 องศาเซลเซียส ปวดกล้ามเนื้อ หน้าแดง ตัวแดง อาจมีผื่นหรือจุดเลือดแดงออกทางผิวหนัง ตรวจพบตับโต และมีผลตรวจ Tourniquet test ให้ผลเป็นบวก

(2) ระยะวิกฤติ (ระยะช็อกและเลือดออก) ไข้ลด อาการทรุดจนอาจเข้าสู่ภาวะช็อก มีอาการกระสับกระส่าย มือเท้าเย็น ชีพจรเต้นเร็ว ความดันโลหิตต่ำ อาเจียนมาก ปวดท้อง ผู้ป่วยบางคนอาจมีอาการซึมเศร้าหรือมีเลือดออกในกระเพาะอาหาร ถ้าได้รับการรักษาอย่างถูกต้องและทันถ่วงทีรวมถึงไม่มีอาการแทรกซ้อน ระยะของอาการนี้ประมาณ 24-48 ชั่วโมง หลังจากนั้นจะเข้าสู่ระยะฟื้นตัวของร่างกาย

(3) ระยะฟื้น ในระยะนี้อาการโดยทั่วไปผู้ป่วยมีอาการดีขึ้น ความดันและชีพจรกลับเป็นปกติ ปัสสาวะออกมากขึ้น ตับโตลดขนาดลง และเริ่มรับประทานอาหารได้ตามปกติ

องค์การอนามัยโลก (WHO, 1999) นิยามการวินิจฉัยโรคที่สามารถบอกได้ว่าเป็น ไข้เลือดออก ถูกบัญญัติโดยองค์การอนามัยโลก โดยต้องมีลักษณะครบทั้ง 4 ประการคือ

1. มีไข้ มีปัญหาทางกระเพาะปัสสาวะ ปวดศีรษะเรื้อรัง และเบื่ออาหาร
2. มีแนวโน้มน้ำมีเลือดออก อาจดูได้จากการถ่ายเป็นเลือดหรืออาเจียนเป็นเลือดได้
3. เกล็ดเลือดต่ำ (มีค่าน้อยกว่า 100,000 ตัวต่อหนึ่งไมโครลิตร)
4. ปรากฏหลักฐานการสูญเสียพลาสมาจากหลอดเลือด (เลือดขึ้นขึ้น ดูจากการเพิ่มขึ้นของเลือดเท่ากับหรือมากกว่า 20% เมื่อเทียบกับความเข้มข้นเลือดเดิม หรือมีหลักฐานการรั่วของพลาสมา เช่น มี pleural effusion และ ascites หรือมี ระดับสัดส่วนโปรตีนหารด้วยอัลบูมินในเลือดต่ำ) มีภาวะสมองอักเสบ

นอกจากนั้นถ้าผู้ป่วยอยู่ในระยะช็อกให้พิจารณาเพิ่มจากที่กล่าวมาแล้วอีก 3 ข้อคือ

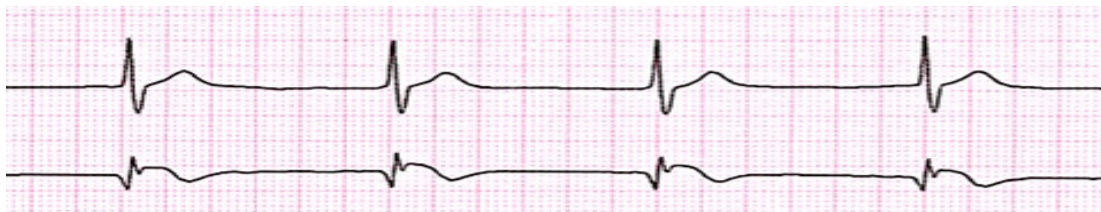
1. ชีพจรเต้นเร็วและเบา
2. ผลต่างค่าความดัน (Pulse pressure minimum) มีค่าน้อยกว่า 20 มิลลิเมตรปรอท
3. ผิวหนังเย็น ชื้น และไม่สลายตัว ซึ่งข้อมูลทางการแพทย์ที่ใช้ในการวิเคราะห์การเป็นโรคมียาค่า เช่น อาการไข้ (Fever), อาเจียน (Vomit), อาการช็อก (Shock), การมีของเหลวในช่องท้อง (Pleural Effusion) เป็นต้น ดังนั้นถ้ามีวิธีที่ช่วยในการวินิจฉัยโรคของแพทย์ ให้สามารถระบุประเภทของโรคไข้เลือดออกได้รวดเร็ว สามารถลดความรุนแรง ภาวะแทรกซ้อนของโรคไข้เลือดออก และอัตราการตายของโรคนี้ได้

4. ลักษณะข้อมูลทางการแพทย์

ข้อมูลทางการแพทย์มีหลายรูปแบบ สามารถแบ่งออกเป็น 3 กลุ่มใหญ่ๆดังนี้

4.1. ข้อมูลทางคลินิก (Clinical Data) คือข้อมูลทางการแพทย์ทั่วไปที่ถูกเก็บในฐานะข้อมูลตัวอย่างข้อมูลเช่น ข้อมูลอุณหภูมิของผู้ป่วย, ข้อมูลทางกายภาพทั่วไป (มีไข้, ปวดหัว) เป็นต้น โดยข้อมูลดังกล่าวถูกเก็บเมื่อผู้ป่วยเข้ามาตรวจโรคโดยมีแพทย์หรือพยาบาลเป็นผู้เก็บข้อมูล

4.2. ข้อมูลอนุกรมเวลา (Time Series Data) คือข้อมูลที่มีค่าของข้อมูลเป็นค่าต่อเนื่อง โดยส่วนใหญ่เป็นค่าที่อ่านได้จากอุปกรณ์ทางการแพทย์ เพื่อวัดค่าความผิดปกติต่างๆ เช่น คลื่นหัวใจ ที่สามารถวัดได้จากเครื่องวัด ค่าข้อมูลเป็นค่าต่อเนื่องที่แสดงออกมาอยู่ในรูปแบบของกราฟ



ภาพที่ 3 ตัวอย่างข้อมูลคลื่นหัวใจ

ที่มา: <http://dmsrsu.thaikm4u.com/blog/chaisit-s/74> (2553)

4.3. ข้อมูลที่เก็บตามช่วงเวลา (Temporal Data) คือข้อมูลที่มีลักษณะการเก็บเป็นช่วงเวลา ต่อผู้ป่วยหนึ่งราย ซึ่งระยะเวลาอาจไม่เท่ากัน ดังนั้นจะเห็นว่าข้อมูลของผู้ป่วยแต่ละคนมีหลายรายการ

EmpID	Name	Department	Salary	ValidTimeStart	ValidTimeEnd
10	John	Research	11000	1985	1990
10	John	Sales	11000	1990	1993
10	John	Sales	12000	1993	INF
11	Paul	Research	10000	1988	1995
12	George	Research	10500	1991	INF
13	Ringo	Sales	15500	1988	INF

ภาพที่ 4 ตัวอย่างฐานข้อมูลชนิด Temporal Database

ที่มา: <http://www.timeconsult.com/TemporalData/TemporalDB.html> (2553)

5. การทำเหมืองข้อมูลทางการแพทย์

เนื่องจากข้อมูลส่วนใหญ่ในปัจจุบันมีความสัมพันธ์ระหว่างข้อมูลซ่อนอยู่ และผู้ใช้งาน ต้องการสกัดความสัมพันธ์หรือองค์ความรู้เพื่อนำมาใช้ประโยชน์ในหลายๆด้าน เช่น ประโยชน์ทางธุรกิจ ยกตัวอย่างผลลัพธ์การทำเหมืองข้อมูลคือ เมื่อลูกค้าซื้อผ้าอ้อมสำเร็จรูปและนมแล้ว ลูกค้ารายนั้นจะซื้อเบียร์ด้วย เห็นว่าองค์ความรู้นี้สามารถช่วยเพิ่มผลกำไรทางธุรกิจได้

ในทางการแพทย์ ข้อมูลในปัจจุบันมีแอททริบิวต์ที่มีความหลากหลายและลักษณะข้อมูลที่เป็นแบบอนุกรมเวลาถือเป็นข้อมูลที่เก็บเป็นช่วงเวลา ซึ่งเป็นการยากในการนำข้อมูลเหล่านั้นมาประมวลผลเพื่อให้ได้มาซึ่งองค์ความรู้หรือความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ในฐานข้อมูล ดังนั้นจึงนำวิธีการทำเหมืองข้อมูลเข้ามาใช้กับข้อมูลที่เป็นแบบชนิดช่วงเวลา ซึ่งเราเรียกว่า Temporal Data Mining (Laxman Srivatsan and Sastry P.S., 2006) หรือข้อมูลที่เป็นแบบอนุกรมเวลาที่เรียกว่า Mining Time Series Data ซึ่งมีเทคนิคในการทำงานหลากหลายวิธีซึ่งเลือกใช้ตามลักษณะของข้อมูล

ปัญหาต่อมาคือ ปัญหาในเรื่องระดับความรุนแรงของโรค ทำให้เกิดความยุ่งยากต่อการวิเคราะห์เพื่อหาวิธีการรักษาที่เหมาะสม อีกทั้งระยะเวลาการแสดงอาการของโรคในผู้ป่วยแต่ละคนอาจใช้เวลาไม่เท่ากัน ทำให้ในบางครั้งอาจไม่ได้รับการรักษาที่ทันท่วงที ดังนั้นจึงมีแนวคิดในการหาเครื่องมือมาช่วยในการวิเคราะห์หรือทำนายสิ่งที่จะเกิดขึ้นในอนาคต โดยอาศัยข้อมูลที่มีอยู่ อาจช่วยแก้ไขปัญหาการวินิจฉัยระดับความรุนแรงของโรคได้ดียิ่งขึ้น ยกตัวอย่างเช่น ถ้ามีผู้ป่วยที่เป็นโรคเมอเร็งเม็ดเลือดกระษัตริ์ แล้วเราเปรียบเทียบอาการในปัจจุบันด้วยองค์ความรู้ที่มีอยู่ อาจทำให้เราสามารถรู้วิธีการรักษาที่เหมาะสมกับผู้ป่วยรายนั้น เพื่อหลีกเลี่ยงกรณีที่ต้องเสียค่ารักษาแต่ไม่หายจากโรคที่รักษาอยู่ ซึ่งอาจก่อให้เกิดค่าใช้จ่ายที่เกินความจำเป็น

ลักษณะข้อมูลทางการแพทย์มีสองลักษณะใหญ่ ได้แก่ ข้อมูลชนิดอนุกรมเวลาหรือข้อมูลชนิดช่วงเวลา จากงานวิจัยที่ศึกษาสามารถแยกการทำเหมืองข้อมูลตามลักษณะของข้อมูลได้ดังนี้

5.1. การทำเหมืองข้อมูลอนุกรมเวลา (Mining Time Series Data) ข้อมูลแบบอนุกรมเวลาเป็นข้อมูลต่อเนื่องที่ถูกเก็บตามช่วงระยะเวลาที่เท่ากัน ข้อมูลของผู้ป่วยแต่ละคนอาจมีลักษณะของค่าข้อมูลที่คล้ายกัน ยกตัวอย่างเช่น ผู้ป่วย ก มีค่าข้อมูลอุณหภูมิที่ถูกเก็บข้อมูลเพิ่มขึ้นตั้งแต่วันที่ 1 (วันที่รับการรักษาครั้งแรก) จนถึงวันที่ 5 ผู้ป่วย ข ถูกเก็บเป็นจำนวน 5 วันเท่ากัน แต่ไม่จำเป็นต้องเป็นวันเดียวกัน มีค่าข้อมูลเพิ่มขึ้นในลักษณะเดียวกัน เป็นต้น ดังนั้นการทำเหมืองข้อมูลอนุกรมเวลาจำเป็นต้องปรับค่าของข้อมูลให้เป็นค่าตัวแทนของข้อมูลของผู้ป่วยรายนั้น วิธีการหาค่าตัวแทนของข้อมูลอนุกรมเวลาเพื่อบอกว่าค่าข้อมูลในช่วงเวลาที่ต่างกันของผู้ป่วยแต่ละคนมีลักษณะข้อมูลคล้ายกันหรือไม่ สามารถทำได้หลายแบบ (Ratanamahatana *et al.*, 2005) ยกตัวอย่างเช่น

5.1.1. Euclidean Distances and L_p Norms เป็นวิธีการหาความคล้ายของข้อมูลอนุกรมเวลา 2 ชุดด้วยการหาระยะห่างระหว่างลักษณะของข้อมูล 2 ชุด แต่ถ้าลักษณะของข้อมูลมีค่าของข้อมูลต่างกันมากเกินไปทำให้การวัดระยะห่างนั้นมีความคลาดเคลื่อน วิธีการนี้จึงมีการปรับค่าหรือการทำนอร์มอลไลซ์ค่าข้อมูลให้อยู่ในช่วงค่าข้อมูลเดียวกันก่อนการหาระยะห่างระหว่างข้อมูล ดังนั้นถ้าระยะห่างของลักษณะข้อมูลของผู้ป่วยที่นำมาเปรียบเทียบกันมีค่าน้อย แสดงว่าข้อมูลของผู้ป่วยคู่หนึ่งมีความคล้ายกันมากและอาจสรุปได้ว่ามีค่าตัวแทนข้อมูลเหมือนกัน แต่วิธีการนี้อาจไม่มีประสิทธิภาพกับข้อมูลที่คล้ายกันในช่วงระยะเวลาที่ต่างกัน วิธีการแก้ปัญหาคือจะกล่าวถึงในหัวข้อต่อไป

5.1.2. Dynamic Time Warping วิธีการนี้มีวัตถุประสงค์เพื่อหาแก้ไขปัญหาวិธีของ Euclidean Distances โดยใช้ลักษณะการบิดงอของกราฟเข้ามาพิจารณาความเหมือนเพิ่มเติม ดังนั้นประสิทธิภาพของวิธีนี้สูงกว่าวิธีการข้างต้น

5.2. การทำเหมืองข้อมูลชนิดช่วงเวลา (Temporal Data Mining) ข้อมูลชนิดช่วงเวลามีเทคนิคการทำเหมืองข้อมูลที่ต่างออกไปที่เรียกว่า การค้นหารูปแบบ (Pattern Discovery) โดยมีวิธีการดังนี้ (Ibrahim *et al.*, 2005 ; Ho *et al.*, 2003)

- (1) Sequential Pattern: การดูรูปแบบเป็นลำดับ
- (2) Frequent episodes: สังเกตช่วงที่เกิดบ่อยๆหรือมีความถี่มาก
- (3) Patterns with explicit time constraints: การดูรูปแบบจากช่วงระยะเวลาจำกัด

6. การวัดประสิทธิภาพทางการแพทย์

งานวิจัยด้านการแพทย์นิยมใช้การวัดค่าประสิทธิภาพด้วยการคำนวณค่าความไวและค่าความจำเพาะดังนี้

- | | |
|----------|--|
| กำหนดให้ | Positive (คลาสบวก) หมายถึง คลาสของผู้ป่วยที่เป็นโรค |
| | Negative (คลาสลบ) หมายถึง คลาสของผู้ป่วยที่ไม่เป็นโรค (ปกติ) |

6.1. ค่าความไว (Sensitivity) ค่าที่บ่งบอกประสิทธิภาพความไวของขั้นตอนวิธีในการตรวจจับคนที่เป็โรคได้ถูกต้อง มีสูตรในการคำนวณดังนี้

$$\text{Sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}} \quad (18)$$

Number of True Positives (TP) คือ จำนวนผู้ป่วยที่ระบบสามารถทำนายได้ถูกต้องว่าผู้ป่วยเป็นโรค

Number of False Negatives (FN) คือ จำนวนผู้ป่วยที่ระบบทำนายว่าไม่เป็นโรคแต่ผู้ป่วยเหล่านั้นเป็นโรค

6.2. ค่าความจำเพาะ (Specificity) ค่าบ่งบอกความสามารถของขั้นตอนวิธีในการคัดกรองผู้ป่วยที่ไม่ได้เป็นโรค (Rule out) มีสูตรในการคำนวณดังนี้

$$\text{Specificity} = \frac{\text{number of True Negatives}}{\text{number of False Positives} + \text{number of True Negatives}} \quad (19)$$

Number of True Negatives (TN) คือ จำนวนผู้ป่วยที่ระบบสามารถทำนายได้ถูกต้องว่าผู้ป่วยไม่เป็นโรค

Number of False Positives (FP) คือ จำนวนผู้ป่วยที่ระบบทำนายว่าเป็นโรคแต่ผู้ป่วยเหล่านั้นไม่เป็นโรค

6.3. ค่าความถูกต้อง (Accuracy) บ่งบอกถึงประสิทธิภาพโดยรวมของตัวแบบผลลัพธ์ มีสูตรดังนี้

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (20)$$

โดยที่ True Positive หมายถึง ผลบวกแท้, ผลบวกจริง

False Positive หมายถึง ผลบวกเทียม, ผลบวกหลง

True Negative หมายถึง ผลลบแท้, ผลลบจริง

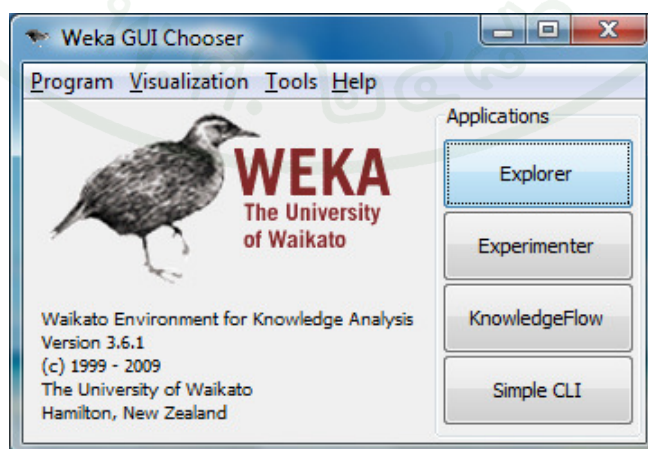
False Negative หมายถึง ผลลบเทียม, ผลลบหลวง

ตารางที่ 2 การวัดค่าประสิทธิภาพทางการแพทย์

		คำเฉลย	
		Positive	Negative
ผลการทำนาย	Positive	True Positive	False Positive <Type I Error>
	Negative	False Negative <Type II Error>	True Negative
		Sensitivity	Specificity
		$\frac{TP}{TP + FN}$	$\frac{TN}{FP + TN}$

7. โปรแกรมเวก้า (Waikato Environment for Knowledge Analysis: WEKA)

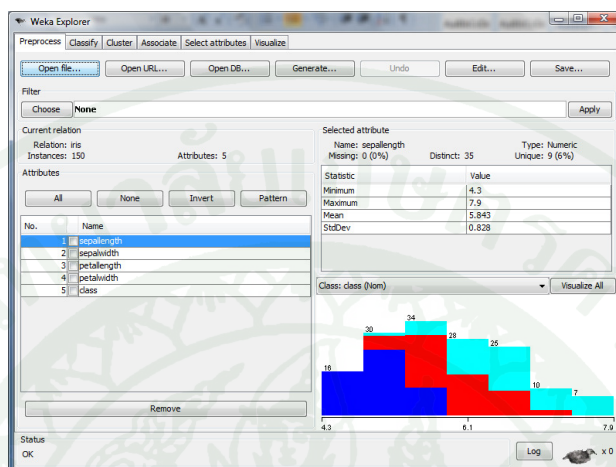
โปรแกรมเวก้าเป็นโปรแกรมที่พัฒนาโดยทีมนักวิจัยจากมหาวิทยาลัยไวกาโต ประเทศนิวซีแลนด์ สามารถเลือกมาใช้งานได้ฟรี ฟังก์ชันในการทำเหมืองข้อมูลประกอบด้วย การจำแนกข้อมูล การจัดกลุ่มข้อมูล และการค้นหาความสัมพันธ์ระหว่างตัวแปรที่ผู้วิจัยสนใจ



ภาพที่ 5 ตัวอย่างโปรแกรมเวก้า

ฟังก์ชันการทำงานในโปรแกรมเวก้าประกอบด้วย

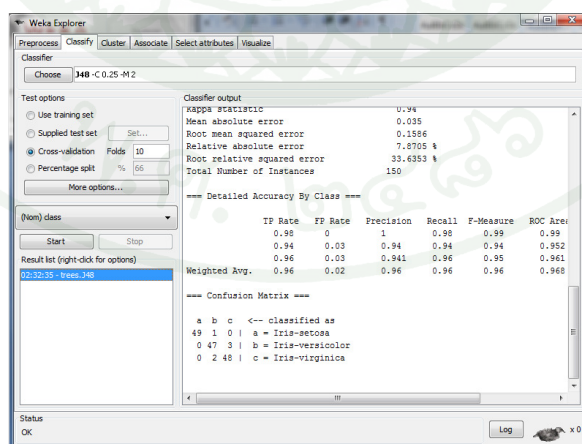
ก. การจัดการข้อมูลก่อนการประมวลผล



ภาพที่ 6 หน้าต่างเตรียมข้อมูล

ถ้าผู้ใช้งานโปรแกรมต้องการตัดแอททริบิวท์บางตัวหรือต้องการเปลี่ยนชนิดของแอททริบิวท์ รวมถึงการประมาณค่าข้อมูลส่วนที่ขาดหายจากการเก็บข้อมูล สามารถทำได้ในส่วนการทำงานนี้

ข. การจำแนกข้อมูล



ภาพที่ 7 หน้าต่างตัวจำแนกข้อมูล

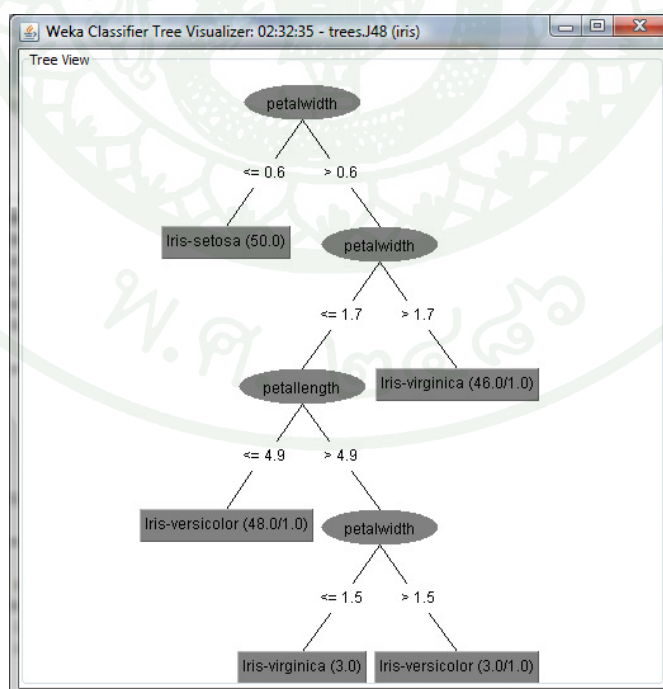
ส่วนของการจำแนกข้อมูลมีขั้นตอนวิธีให้เลือกใช้ตามความเหมาะสมและวัตถุประสงค์ของข้อมูลที่ผู้ใช้งานสนใจ รวมถึงสามารถตั้งค่าการประเมินผลการทดลองด้วยการกำหนดวิธีการวัดประสิทธิภาพด้วยค่าของครอสวาไลเดชัน (fold-cross validation) เพื่อเพิ่มความน่าเชื่อถือของผลการทดลอง

งานวิจัยนี้สนใจขั้นตอนวิธีต้นไม้ตัดสินใจ ดังนั้นจึงเลือกขั้นตอนวิธีต้นไม้ชื่อว่า J48 ซึ่งขั้นตอนวิธีนี้เป็นขั้นตอนวิธีที่พัฒนาแล้ว ข้อมูลที่นำเข้าไปทดลองสามารถนำเข้าได้ทั้งข้อมูลชนิดคุณภาพและข้อมูลชนิดปริมาณ ตัวอย่างผลลัพธ์ของขั้นตอนวิธีต้นไม้ตัดสินใจดังแสดงในภาพที่ 8

ค. การจัดกลุ่มข้อมูล

ส่วนของการจัดกลุ่มข้อมูลคล้ายกับส่วนของการจำแนก แต่แตกต่างกันตรงข้อมูลที่นำเข้ามาใช้งานของส่วนนี้ไม่จำเป็นต้องเป็นข้อมูลที่มีผลเฉลย ผู้ใช้ต้องการแบ่งกลุ่มตามธรรมชาติของข้อมูล

ขั้นตอนวิธีในการแบ่งกลุ่มมีหลายวิธีให้เลือกตามความต้องการของผู้ใช้งาน เช่น SimpleKMeans ซึ่งเป็นชื่อขั้นตอนวิธีเคมีน เป็นต้น



ภาพที่ 8 ภาพตัวอย่างผลลัพธ์ต้นไม้ตัดสินใจ

ง. การค้นหาความสัมพันธ์

ส่วนนี้เป็นการค้นหาความสัมพันธ์ระหว่างแอททริบิวต์ หรือความสัมพันธ์ระหว่างแอททริบิวต์กับคลาสที่สนใจ ข้อมูลที่ใช้ต้องเป็น Nominal Attributes เท่านั้น

จ. การเรียกดูภาพ (visualize)

ส่วนนี้สำหรับแสดงการกระจายตัวของข้อมูลซึ่งแสดงออกเป็นรูปแบบของกราฟ จุดข้อมูล ผลการกระจายตัวที่ได้อาจส่งผลต่อการเลือกขั้นตอนวิธีเพื่อเพิ่มประสิทธิภาพของการทำเหมืองข้อมูลในแบบต่างๆได้

เนื่องจากโปรแกรมเวก้ามีฟังก์ชันการทำงานหลายอย่างเช่น การจำแนกข้อมูล การแบ่งข้อมูล และการหาความสัมพันธ์ของข้อมูล จึงเป็นสาเหตุให้โปรแกรมนี้ถูกนำไปประยุกต์ใช้กับข้อมูลหลายสาขงานในการทำเหมืองข้อมูล ยกตัวอย่างเช่น การนำขั้นตอนวิธีต้นไม้ตัดสินใจในโปรแกรมเวก้าเพื่อทำนายการซ้ำของผลแอปเปิลจากอิทธิพลการกระแทก รัศมีของการซ้ำและตำแหน่งที่เกิดการซ้ำในขั้นตอนบรรจุหีบห่อเตรียมส่งขายบนสายพานเพื่อคัดเลือกผลิตภัณฑ์ที่ดีที่สุด (Holmes *et al.*, 1998) การใช้ต้นไม้ตัดสินใจทำนายประสิทธิภาพของเห็ดแบบตามเวลา ณ ตอนนั้น โดยอาศัยภาพถ่ายเพื่อจัดลำดับของเห็ด (Kusabs *et al.*, 1998) การทำเหมืองข้อมูลตัวอักษรและการสกัดข้อความจากข้อความโดยอัตโนมัติ (Frank *et al.*, 1999) และการแบ่งชนิดของเอกสารด้วยขั้นตอนของโปรแกรมเวก้า (Sauban and Pfahringer, 2003) นอกจากนี้ โปรแกรมเวก้ายังถูกนำไปใช้ในสาขาทางการแพทย์อีกด้วย ยกตัวอย่างเช่นการจำแนกข้อมูลลักษณะของยีนและสกัดกฎจากข้อมูลนั้น (Li *et al.*, 2003) เป็นต้น

8. โปรแกรมแมตแล็บ (Matrix Laboratory: MATLAB)



โปรแกรมนี้เป็นโปรแกรมในการคำนวณและการเขียนโปรแกรมชนิดหนึ่งที่สามารถในการสร้างแบบจำลองทางคณิตศาสตร์ รวมถึงโปรแกรมนี้ยังเป็นเครื่องมือที่ใช้ในการพัฒนาอัลกอริทึม การสร้างระบบและการจัดการเรื่องของ image processing และเป็นเครื่องมือที่ช่วยให้ผู้ใช้งานสามารถแก้ปัญหาตัวแปรที่มีลักษณะเป็นเมทริกซ์ได้ง่ายขึ้น

งานวิจัยที่เกี่ยวข้อง

1. งานวิจัยเกี่ยวกับการทำเหมืองข้อมูลทางการแพทย์

ในปัจจุบันพบว่าม้งงานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อมูลทางการแพทย์มากมายหลายโรค ยกตัวอย่างเช่น โรคมะเร็งทรวงอก โรคไวรัสตับอักเสบบ เป็นต้น เทคนิคการทำเหมืองข้อมูลทางการแพทย์สามารถแบ่งได้ตามวัตถุประสงค์ที่ต้องการศึกษา เช่น การจำแนกข้อมูล การทำนายข้อมูล และวิธีที่เลือกใช้ควรเหมาะสมกับลักษณะข้อมูลและผลลัพธ์ที่สามารถนำไปประยุกต์ใช้

งานวิจัยของ Delen D. และคณะ (Delen *et al.*, 2005) ได้ศึกษาการวัดประสิทธิภาพการจำแนกข้อมูลการรอดชีวิตของผู้ป่วยโรคมะเร็งทรวงอก โดยข้อมูลที่ใช้ นำมาจาก The American Cancer Society โดยมีข้อมูลที่ใช้ในการทดลองคือ ข้อมูลผู้ป่วย 202,932 คน และตัวแปรที่ใช้ในการพิจารณาทั้งหมด 17 ตัวแปร ผลลัพธ์จำแนกผลออกเป็น ไม่รอดชีวิต และรอดชีวิตโดยเทคนิคที่ใช้เปรียบเทียบมีทั้งหมด 3 เทคนิคคือ โครงข่ายประสาทเทียม, ขั้นตอนวิธีต้นไม้ตัดสินใจ และขั้นตอนวิธีการวัดค่าความเสื่อมถอยทางตรรกะ การวัดประสิทธิภาพของผลลัพธ์มี 2 วิธี วิธีแรกคือการวัดค่าความถูกต้องรวม ค่า sensitivity และค่า specificity ผลการทดลองพบว่าขั้นตอนวิธีโครงข่ายประสาทเทียมให้ค่าเท่ากับ 91.2% 94.37% และ 87.48% ตามลำดับ ขั้นตอนวิธีการวัดค่าความเสื่อมถอยทางตรรกะให้ค่าเท่ากับ 89.20% 90.71% และ 87.86% ตามลำดับ และขั้นตอนวิธีต้นไม้ตัดสินใจให้ค่าเท่ากับ 93.62% 96.02% และ 90.66% ตามลำดับ และวิธีการวัดประสิทธิภาพแบบที่สองคือ วิธีการ k-Fold cross-validation ขั้นตอนวิธีต้นไม้ตัดสินใจให้ผลค่าประสิทธิภาพการทำงานได้ดีที่สุด โดยให้ค่าความถูกต้องถึง 93.62% รองลงมาคือ ขั้นตอนวิธีโครงข่ายประสาทเทียม โดยให้ค่าความถูกต้องถึง 91.21% และสุดท้ายคือขั้นตอนวิธีการวัดค่าความเสื่อมถอยทางตรรกะ โดยให้ค่าความถูกต้องเพียง 89.20%

งานวิจัยของ Lin Rang-Ho (Lin, 2009) ค้นหาองค์ความรู้ที่ซ่อนอยู่ในฐานข้อมูลของโรคไวรัสตับอักเสบบ โดยใช้ 2 ขั้นตอนวิธี ได้แก่

(1) การจำแนกและการสร้างแผนภาพต้นไม้

ขั้นตอนวิธีการจำแนกและการสร้างแผนภาพต้นไม้เป็นขั้นตอนวิธีการเรียนรู้เพื่อสร้างต้นไม้ตัดสินใจเพื่อสกัดกฎหรือองค์ความรู้ที่ซ่อนอยู่ในข้อมูลโรคตับอักเสบบและแสดงผลได้ว่าผู้ป่วยคนใดเป็นโรค หรือ ไม่เป็นโรค

(2) การค้นคืนกรณี

องค์ความรู้ที่ได้ถูกนำมาต่อยอดในการสร้างโปรแกรมการค้นคืนกรณีเพื่อค้นหาและหาคำตอบได้ว่าผู้ป่วยรายนั้นเป็นโรคไวรัสตับอักเสบชนิดใด (โรคตับอักเสบประกอบด้วย 4 ชนิด คือ โรคตับอักเสบขั้นรุนแรง (chronic hepatitis), โรคแอลกอฮอล์ในตับ (alcohol hepatitis), โรคตับแข็ง (liver cirrhosis) และ โรคตับอักเสบชนิด B (B hepatitis))

จากผลการทดลองพบว่าวิธีการของการจำแนกและการสร้างแผนภาพต้นไม้ให้ค่าความถูกต้องในการบอกว่าผู้ป่วยคนใดเป็นโรคหรือไม่เป็นโรคได้เท่ากับ 92.94% และในขั้นตอนการค้นคืนกรณีของผลลัพธ์ที่ได้สามารถบอกชนิดของโรคตับอักเสบที่เป็นได้โดยให้ค่าความถูกต้องในเท่ากับ 90.0%

ในงานวิจัยฉบับนี้ ผู้วิจัยได้สนใจทำเหมืองข้อมูลโรคไข้เลือดออกเพื่อศึกษาปัจจัยสำคัญที่ส่งผลต่อการเกิดโรค อีกทั้งต้องการทราบรูปแบบของการเกิดโรค ซึ่งในปัจจุบันได้มีผู้สนใจศึกษาโรคไข้เลือดออกแล้ว แต่ยังไม่แพร่หลาย อีกทั้งการศึกษาโรคใดๆจำเป็นต้องมีข้อมูลที่จำเพาะต่อพื้นที่ศึกษาและปริมาณข้อมูลที่เพียงพอต่อการเรียนรู้ เพื่อให้ผลลัพธ์มีประสิทธิภาพและสามารถนำไปใช้ได้จริง

งานวิจัยของ Tanner Lukas และคณะได้ศึกษาวิจัยการทำเหมืองข้อมูลโรคไข้เลือดออกเพื่อทำนายผลของโรค โดยแบ่งคลาสไข้เลือดออกเป็น 2 คลาสคือ ไข้เดงกีและไข้เลือดออก ข้อมูลผู้ป่วยที่ใช้ในงานวิจัยมีจำนวน 1,200 คนจากโรงพยาบาลดองเทป (Dong Thap Hospital) ประเทศสิงคโปร์ โดยมีจำนวนผู้ป่วยที่ติดเชื้อโรคไข้เลือดออก (ทุกชนิด) จำนวน 364 คน และผู้ป่วยที่ไม่ได้เป็นโรคไข้เลือดออกจำนวนทั้งสิ้น 836 คน โดยงานวิจัยนี้แบ่งขั้นตอนการทดลองออกเป็นสองส่วนคือ ส่วนที่ 1 ทดสอบโดยวิธีการทางการแพทย์ และส่วนที่ 2 คือ ใช้ขั้นตอนวิธีในการวิเคราะห์คือ ขั้นตอนวิธีต้นไม้ตัดสินใจ และกำหนดค่าการตรวจสอบเป็น 10-fold cross validation เพื่อตรวจสอบความถูกต้องของขั้นตอนวิธีที่ใช้ ผลลัพธ์จากการทดลองแสดงเป็นแผนภาพต้นไม้ที่บ่งบอกแอททริบิวต์ที่สำคัญต่างๆที่ขั้นตอนวิธีเลือกมา และความสัมพันธ์ในรูปของกฎ ผลลัพธ์ได้ค่า sensitivity เท่ากับ 71.2% และค่า specificity เท่ากับ 90.1% และให้ค่าความถูกต้องรวมเท่ากับ 84.7% ตามลำดับ (Tanner *et al.*, 2008)

งานวิจัยของ Ibrahim Fatimah และคณะ (Ibrahim *et al.*, 2005) ศึกษาเกี่ยวกับโรคไข้เลือดออก ผู้วิจัยต้องการทำนายวันที่ไข้ลดจนอาจทำให้เกิดอาการช็อกได้ หรือทางการแพทย์เรียกวันดังกล่าวว่า day0 กระบวนการในงานวิจัยมีสองส่วนคือ ส่วนการวิเคราะห์ข้อมูลเพื่อวิเคราะห์ข้อมูลโดยใช้โปรแกรม SPSS และส่วนที่สองใช้ขั้นตอนวิธีโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) โดยโปรแกรมที่ใช้ในขั้นตอนวิธีโครงข่ายประสาทเทียมคือ โปรแกรม MATLAB ตัวแบบที่ใช้คือ Multilayer feed forward neural network (MFNN) และมีการหาค่าที่เหมาะสมที่ใช้ในขั้นตอนวิธีโครงข่ายประสาทเทียมทั้งหมด เช่น จำนวนนิวรอน คำนวณน้ำหนักที่ให้แก่ตัวแบบ และอัตราการเรียนรู้ เป็นต้น ข้อมูลที่ใช้ในการทดลองคือ ข้อมูลผู้ป่วยจำนวน 252 คนจากโรงพยาบาล HUKM (Hospital Universiti Kebangsaan Malaysia) ประเทศมาเลเซีย โดยในข้อมูลประกอบด้วยผู้ป่วยไข้เดงกี (DF) 4 คน และผู้ป่วยไข้เลือดออก (DHF) 248 คน ผลการทำนายมีทั้งหมด 3 แบบคือ 2 วัน 3 วัน และ 4 วัน ผลลัพธ์ที่ได้มีค่าความถูกต้องโดยเฉลี่ยตั้งแต่ 81.3% ถึง 90%

ต่อมาผู้วิจัยต้องการทราบความเสี่ยงของการเป็นโรคไข้เลือดออก งานวิจัยของ Faisal Tarig และคณะ (Tarig Faisal *et al.*, 2010a) กล่าวถึงตัวแบบที่ทำนายความเสี่ยงของการเป็นโรคไข้เลือดออกโดยใช้ขั้นตอนวิธีร่วมกัน 2 วิธี วิธีแรกคือ Self-organizing map (SOM) ซึ่งเป็นวิธีการเรียนรู้แบบไม่มีผลเฉลยเพื่อแบ่งกลุ่ม โดยใช้หลักการการหาความสัมพันธ์แบบ Neighborhood Relation ซึ่งวิธีนี้สามารถระบุกลุ่มของผู้ป่วยตามความเสี่ยง วิธีที่สองคือ multilayer feed-forward neural network เพื่อสร้างตัวแบบการทำนาย ข้อมูลผู้ป่วยที่ใช้ในการทดลองมีทั้งหมด 210 คน โดยนำข้อมูลมาจากงานวิจัยของ Ibrahim (Ibrahim *et al.*, 2005) โดยขั้นตอนในการทำวิจัยมี 3 ขั้นตอนคือ ขั้นแรกนำข้อมูลการตรวจเลือด (Blood test) โดยมีพารามิเตอร์คือ Hemotocrit concentration: HCT และ Platelet counts: PLT อีกทั้งข้อมูลการตรวจค่าของตับ(liver function test) โดยมีพารามิเตอร์คือ Alanine aminotransferase: ALT และ Aspartate aminotransferase: AST นำพารามิเตอร์ข้างต้นตั้งแต่ Day0 ถึง Day4 ไปเข้ากระบวนการ SOM เพื่อแบ่งกลุ่มผู้ป่วยที่มีความเสี่ยงสูงและกลุ่มผู้ป่วยที่มีความเสี่ยงต่ำ ขั้นตอนที่สองใช้ข้อมูลพารามิเตอร์ทั้งหมด 37 ตัว และปรับค่าในพารามิเตอร์ให้มีความตั้งแต่ 0 ถึง 1 หลังจากนั้นนำไปเข้าขั้นตอน SOM อีกครั้ง ขั้นตอนที่สามใช้วิธี ANN โดยใช้ข้อมูลเข้าที่ผ่านกระบวนการจากขั้นตอนสองขั้นแรก หลังจากนั้นแบ่งข้อมูลออกเป็นสองส่วนคือ ข้อมูลสอน 70% (ผู้ป่วย 147 คน) และข้อมูลทดสอบ 30% (ผู้ป่วย 63 คน) ตั้งค่าพารามิเตอร์ในการเรียนรู้ ผลการทดลองได้ว่า SOM คัดกรองพารามิเตอร์ออกมาได้ 7 พารามิเตอร์ หลังจากนั้นนำเข้า MFNN เพื่อสร้างตัวแบบการทำนายโดยใช้ค่าจำนวนโหนดใน

โหนดซ่อนเท่ากับ 10 ค่าน้ำหนักที่ใช้คือ 0.99 ค่าอัตราการเรียนรู้ 0.1 และค่าการทำซ้ำ 20,000 ครั้ง ค่าความถูกต้องของการทำนายมีค่าเท่ากับ 70% และค่า sum square error มีค่าเท่ากับ 0.121 ต่อมา Faisal Tarig และคณะ (Tarig Faisal *et al.*, 2010b) ทำงานวิจัยในการตรวจสอบเกณฑ์ในการวัดความเสี่ยงของโรคไข้เลือดออกด้วยวิธีการ Self-organizing map อีกครั้งโดยใช้ข้อมูลผู้ป่วยทั้งหมด 195 คน ผลการวิจัยพบว่าสามารถระบุเกณฑ์ใหม่ที่ใช้ในการวัดความเสี่ยงของโรคไข้เลือดออกคือ เกณฑ์ที่ 1 ค่าจำนวนเกล็ดเลือด (HCT) น้อยกว่าหรือเท่ากับ 40,000 เซลล์ต่อลูกบาศก์มิลลิเมตร เกณฑ์ที่ 2 ค่าความเข้มข้นของเม็ดเลือดในพลาสมา (PLT) มากกว่าหรือเท่ากับ 25% เกณฑ์ที่ 3 ค่าอะมิโนในตับ (AST,ALT) เกินค่าปกติ โดยครั้งนี้ได้เปรียบเทียบการทำงานระหว่าง SOM และ ขั้นตอนวิธีเคมิน ผลการทดลองสรุปว่าวิธีการ SOM สามารถแยกแยะเกณฑ์ข้างต้นออกจากกันได้ดี มีประสิทธิภาพมากกว่าขั้นตอนวิธีเคมินถึง 23.8% (Tarig Faisal *et al.*, 2010b) และในปีเดียวกัน คณะวิจัยนี้ (Tarig Faisal *et al.*, 2010c) เลือกใช้ขั้นตอนวิธี Multilayer Perceptron : MLP เพื่อใช้ในการจำแนกความเสี่ยงของโรค อัลกอริทึมที่ใช้มีสองอัลกอริทึมคือ Levenberg-Marquardt (MLPLM) และ Scaled Conjugate Gradient (MLPSCG) การทดลองมุ่งเน้นการวัดค่าประสิทธิภาพเพื่อค้นหาค่าเริ่มต้นที่เหมาะสมที่ใช้ในแต่ละอัลกอริทึม ผลการทดลองคือ อัลกอริทึม MLPLM ให้ค่าความถูกต้องคือ 75% ซึ่งดีกว่าอัลกอริทึม MLPSCG ที่มีค่าความถูกต้องเท่ากับ 70.7% แต่ค่าความถูกต้องที่ได้มีค่าไม่มากพอที่จะรับรองผลการทดลองว่าสามารถนำไปใช้งานได้จริง จึงสามารถปรับปรุงผลของการทำงานให้ดีขึ้นด้วยวิธีการอื่นได้ เช่น ใช้ฟัซซี่ลอจิกเข้ามาช่วยในการจำแนก เป็นต้น

จากงานวิจัยที่เกี่ยวข้องพบว่าเทคนิคในการทำเหมืองข้อมูลโรคทางการแพทย์มีหลายวิธี ขึ้นกับวัตถุประสงค์ของงานวิจัย สำหรับงานวิจัยนี้ ผู้วิจัยศึกษาปัจจัยสำคัญที่ส่งผลต่อการจำแนก ระดับของไข้เลือดออก และศึกษาข้อสันนิษฐานที่เกี่ยวข้องกับโรคไข้เลือดออกพร้อมด้วย ดังนั้นผู้วิจัยจึงเลือกขั้นตอนวิธีในการทำเหมืองข้อมูล ได้แก่ วิธีต้นไม้ตัดสินใจ เพื่อค้นหาปัจจัยและรูปแบบการเกิดโรคของผู้ป่วยโรคไข้เลือดออก วิธีฟัซซี่ลอจิกเพื่อปรับปรุงผลลัพธ์จากขั้นตอนวิธีต้นไม้ตัดสินใจให้มีประสิทธิภาพมากขึ้น วิธีโครงข่ายประสาทเทียมใช้ทำนายวัน day0 ของผู้ป่วยโรคไข้เลือดออกเพื่อเป็นข้อเสนอแนะให้กับแพทย์ผู้เชี่ยวชาญ วิธีเคมินเพื่อทดสอบการแบ่งกลุ่มตามความรุนแรงของโรค และขั้นตอนวิธีอื่นๆ เพื่อตอบข้อสงสัยของแพทย์ผู้เชี่ยวชาญตามวัตถุประสงค์ที่ต้องการค้นหาคำตอบ

ตารางที่ 3 สรุปงานวิจัยที่เกี่ยวข้อง เรียงตามประเภทงานวิจัยตามปีที่ตีพิมพ์

ที่	ปี	ชื่องานวิจัย	สิ่งที่นำเสนอ	ผลการทดลอง
1	2005	Predicting breast cancer survivability: a comparison of three data mining methods	ประสิทธิภาพการจำแนกข้อมูลผู้รอดชีวิตจากมะเร็งทรวงอกโดยเปรียบเทียบสามอัลกอริทึม คือ โครงข่ายประสาทเทียม, ต้นไม้การตัดสินใจ และการวัดค่าความเสื่อมถอยทางตรรกะ	ขั้นตอนวิธีการต้นไม้การตัดสินใจให้ผลค่าประสิทธิภาพการทำงานได้ดีที่สุด
2	2005	A novel dengue fever (DF) and dengue hemorrhagic fever (DHF) analysis using artificial neural network (ANN)	ใช้ขั้นตอนวิธีโครงข่ายประสาทเทียมเพื่อทำนายวันที่ไข้ลด (day0) ซึ่งอาจทำให้เกิดอาการช็อกได้	ให้ค่าความถูกต้องในการทำนายได้มากถึง 81.3% ถึง 90%
3	2008	Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase	การใช้ตัวแบบต้นไม้การตัดสินใจมาใช้ในการทำนายความเป็นโรคของผู้ป่วยในระยะเริ่มแรก	ให้ค่าความถูกต้องในการทำนายรวม 84.7%
4	2009	An intelligent model for liver disease diagnosis	การค้นหาค่าความรู้ในฐานข้อมูลโรคไวรัสตับอักเสบด้วยสองวิธีคือ CART เพื่อบอกว่าผู้ป่วยคนใดเป็นโรคและ CBR เพื่อบอกว่าผู้ป่วยเป็นโรคตับอักเสบชนิดใด	CART ให้ค่าความถูกต้อง 92.94% และ CBR ให้ค่าความถูกต้อง 90.0%

ตารางที่ 3 (ต่อ)

ที่	ปี	ชื่องานวิจัย	สิ่งที่นำเสนอ
5	2010	A noninvasive intelligent approach for predicting the risk in dengue patients	ขั้นตอนวิธีการทำนายความเสี่ยงของผู้ป่วยโรคไข้เลือดออกโดยใช้วิธีการ SOM และ MFNN ค่าความถูกต้องของตัวแบบในการทำนายเท่ากับ 70% และค่า sum square error เท่ากับ 0.121
6	2010	Reexamination of risk criteria in dengue patients using the self-organizing map	ขั้นตอนวิธี SOM สามารถจำแนกกลุ่มระดับความเสี่ยงออกจากกันได้ดีมีประสิทธิภาพมากกว่าขั้นตอนวิธีเคมีน 23.8%
7	2010	Neural network diagnostic system for dengue patients risk classification	การเปรียบเทียบอัลกอริทึมในขั้นตอน MLP ที่ใช้จำแนกโรคไข้เลือดออก ระหว่างอัลกอริทึม MLPLM และอัลกอริทึม MLPSCG อัลกอริทึม MLPLM ทำงานได้ดีกว่าอัลกอริทึม MLPSCG โดยให้ความถูกต้อง 75% และ 70.7% ตามลำดับ

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องคอมพิวเตอร์หน่วยประมวลผล Intel Core2 Duo processor T5550 1.83 GHz หน่วยความจำ 3 GB และฮาร์ดดิสก์บรรจุข้อมูล 160 GB
2. โปรแกรมเวก้า version 3-6-1
3. โปรแกรมเมตแล็บ
3. ระบบปฏิบัติการ Microsoft Windows Vista Home Premium

วิธีการ

วิทยานิพนธ์ฉบับนี้แบ่งการทดลองออกเป็น 8 ชุด ตามวัตถุประสงค์ของโจทย์วิจัย โดยปัญหาหลักคือการจำแนกประเภทของไข้เลือดออกด้วยขั้นตอนวิธีต้นไม้ตัดสินใจ และปัญหารองคือการทำนายวันที่ไข้ของผู้ป่วยด้วยขั้นตอนวิธีต้นไม้ตัดสินใจ วิธีพีชชีลอจิกและวิธีโครงข่ายประสาทเทียม สำหรับการหาความสัมพันธ์ระหว่างปัจจัยในโจทย์ย่อยที่สนใจ ผู้วิจัยเลือกวิธีการหาความสัมพันธ์ระหว่างแอททริบิวต์ร่วมกับการพิจารณาสัดส่วนจากข้อมูลจริงเพื่อสรุปความสัมพันธ์และตอบโจทย์ปัญหาย่อยอื่นๆ ในแต่ละการทดลองใช้ชุดของแอททริบิวต์ (ปัจจัย) และคลาสต่างกัน ขึ้นอยู่กับขอบเขตของโจทย์ที่ต้องการศึกษา

1. ข้อมูลที่ใช้ในการทดลอง

ข้อมูลในงานวิจัยได้รับความร่วมมือจากโรงพยาบาลศิริราช ซึ่งโรงพยาบาลได้รวบรวมข้อมูลจาก 2 แหล่งคือ โรงพยาบาลศรีนครินทร์ จังหวัดขอนแก่น และโรงพยาบาลสงขลานครินทร์ จังหวัดสงขลา ข้อมูลผู้ป่วยอายุไม่เกิน 15 ปี ค่าข้อมูลดิบได้จากแบบฟอร์มการกรอกรายละเอียดของผู้ป่วยโรคไข้เลือดออก (รายละเอียดภาคผนวก ก) โดยเอกสารแต่ละฉบับมีวัตถุประสงค์และจำนวนคำถามที่แตกต่างกันออกไป ดังแสดงในตารางที่ 4 ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่รวบรวมข้อมูลตั้งแต่ พ.ศ. 2544 ถึง พ.ศ. 2552 ก่อนการทดลอง ผู้ทดลองได้เตรียมข้อมูลตามขั้นตอนวิธีการทำเหมืองข้อมูลโดยได้รับความร่วมมือจากผู้เชี่ยวชาญในการคัดกรองและเลือกปัจจัยที่นำมาใช้ในการทดลอง ดังนั้นข้อมูลผู้ป่วยที่ผ่านขั้นตอนการเตรียมข้อมูลแล้วมีจำนวน 1001 คน

ตารางที่ 4 วัตถุประสงค์และจำนวนแอมพริบในแต่ละเอกสาร

ชื่อเอกสาร	วัตถุประสงค์ของเอกสาร	จำนวนแอมพริบที่ (ต่อหนึ่งเอกสาร)
Form Inclusion criteria & Form 1 Enroll (D01)	เอกสารเก็บข้อมูลทั่วไปของผู้ป่วยและอาการเบื้องต้น โดยข้อมูลเป็นข้อมูลในวันที่ผู้ป่วยเข้าร่วมโครงการกับทางโรงพยาบาล (วันตรวจรักษา)	109 แอมพริบ
Form 2 : Drug Taken Before Admission (D02)	เอกสารเก็บข้อมูลประวัติเกี่ยวกับการใช้ยาหรือว่าการรักษาก่อนที่ผู้ป่วยเข้ารับการรักษาส่วนใหญ่เป็นข้อมูลทางกายภาพ เช่น เคยได้รับวัคซีน JE หรือไม่	118 แอมพริบ
Form 3 : Master (D03)	เอกสารบันทึกข้อมูลการรักษาตัวของผู้ป่วยรายหนึ่ง ซึ่งต้องตรวจวัดและบันทึกค่าทุกวันจนกระทั่งจบการรักษา ซึ่งสังเกตว่าข้อมูลในเอกสารนี้ ผู้ป่วยหนึ่งรายมีข้อมูลหลายเรคคอร์ดตามจำนวนวันที่ผู้ป่วยรายนั้นเข้ารับการรักษาที่โรงพยาบาล	31 แอมพริบ
Form 4/1 : Admission Clinical (D04)	เอกสารบันทึกข้อมูลทางด้านกายภาพของผู้ป่วยรายหนึ่ง เช่น สุขภาพจิตหรือการมีผื่นขึ้นที่ผิวหนัง เป็นต้น ซึ่งเอกสารนี้แพทย์หรือพยาบาลต้องบันทึกทุกวัน ดังนั้นผู้ป่วยรายหนึ่งมีข้อมูลหลายเรคคอร์ด	64 แอมพริบ
Form 4/2 : Admission : Routine Laboratory (D05)	เอกสารบันทึกผลค่าทางห้องปฏิบัติการของผู้ป่วย เช่น จำนวนเม็ดเลือดขาว, ผลการตรวจวัดเอนไซม์, ผลทางเคมีของตับ เป็นต้นเป็นค่าที่บันทึกทุกวัน	57 แอมพริบ

ตารางที่ 4 (ต่อ)

ชื่อเอกสาร	วัตถุประสงค์ของเอกสาร	จำนวนแอมทริบิวท์ (ต่อหนึ่งเอกสาร)
Form 5 : Clinical Dx Vs Final Dx (D06)	เอกสารที่เก็บการวินิจฉัยของแพทย์และ ผลการตัดสินใจผู้ป่วยรายนั้นๆเป็นโรค ไข้เลือดออกหรือไม่ และกำลังอยู่ใน ระดับความรุนแรงขึ้นใด	33 แอมทริบิวท์
รวม		412 แอมทริบิวท์

ในงานวิจัยนี้ประกอบด้วยเอกสารที่เกี่ยวข้องกับงานวิจัยโรคไข้เลือดออกทั้งหมด 6 เอกสารและแอมทริบิวท์ที่เกี่ยวข้องทั้งหมด 412 แอมทริบิวท์ แต่การทำงานวิจัยแต่ละโจทย์ ผู้วิจัยคัดเลือกแอมทริบิวท์ที่ใช้ในแต่ละการทดลองให้เหมาะสมกับความต้องการในแต่ละโจทย์

2. ขั้นตอนวิธีการค้นหาคำความรู้

ก่อนการทดลองผู้วิจัยได้ประมวลผลข้อมูลตามขั้นตอนของการค้นหาคำความรู้ดังจะอธิบายขั้นตอนได้ดังนี้

2.1 ขั้นตอนการเตรียมข้อมูลก่อนทดลอง

2.1.1 ขั้นตอนการทำความสะอาดข้อมูล ผู้วิจัยได้จัดการข้อมูลที่ขาดหายและเติมค่าขาดหายด้วยค่าประมาณจากแอมทริบิวท์นั้น ซึ่งขั้นตอนนี้ได้รับความร่วมมือจากเจ้าหน้าที่ของโรงพยาบาลศิริราชซึ่งรับผิดชอบข้อมูลและได้ความร่วมมือจากแพทย์ผู้เชี่ยวชาญในการตรวจสอบค่าที่ขาดหายและการประมาณค่าข้อมูล

2.1.2 ขั้นตอนการรวมข้อมูล ผู้วิจัยได้รับความร่วมมือจากเจ้าหน้าที่โรงพยาบาลศิริราชได้รวบรวมข้อมูลจากโรงพยาบาลจังหวัดสงขลาและโรงพยาบาลจังหวัดขอนแก่น

2.1.3 ขั้นตอนการคัดเลือกข้อมูล เป็นขั้นตอนการคัดเลือกแอมทริบิวท์ที่ได้รับความร่วมมือจากแพทย์ในการคัดเลือกแอมทริบิวท์ที่ไม่เกี่ยวข้อง หรือไม่ส่งผลต่อการวิจัยออก เพื่อลดจำนวนปัจจัยที่ต้องการพิจารณาและเพื่อเพิ่มความน่าเชื่อถือให้แก่ผลลัพธ์การทำเหมืองข้อมูล

2.1.4 ขั้นตอนการแปลงข้อมูล เป็นขั้นตอนการหาตัวแทนของข้อมูล เนื่องจากข้อมูลที่ได้รับมาเป็นข้อมูลที่ถูกรวบรวมไว้เป็นช่วงเวลาซึ่งส่งผลให้ผู้ปวยรายหนึ่งมีค่าข้อมูลมากกว่า 1 รายการ แต่ค่าข้อมูลที่ใช้ในการทดลองควรเป็นตัวเลขเพียงตัวเดียวแทนข้อมูลของผู้ปวยรายนั้น

2.2 ขั้นตอนการทำเหมืองข้อมูล หลังจากได้ข้อมูลที่พร้อมประมวลผลแล้ว นำลักษณะเด่นที่สร้างขึ้นไปใช้ในการทดลองตามขั้นตอนวิธีการทำเหมืองข้อมูลด้วยขั้นตอนวิธีที่ต้องการทดลอง

ฟังก์ชันในการทำเหมืองข้อมูลที่ทดลอง 3 รูปแบบดังนี้

2.2.1 การจำแนกประเภทข้อมูล: การเลือกฟังก์ชันการจำแนกประเภทข้อมูลหรือ Classification โดยใช้ตัวแบบ J48 ซึ่งเป็นวิธีการต้นไม้ตัดสินใจคือ แบ่งแยกเป็นแผนภาพต้นไม้เพื่อให้ระบบประมวลพิจารณาแอททริบิวต์ที่ส่งผลต่อการเกิดโรคไขเลือดออก

2.2.2 การจัดกลุ่มข้อมูล: ขั้นตอนวิธีที่ใช้ในงานวิจัยคือขั้นตอนวิธีเคมีน เนื่องจากเป็นวิธีที่ไม่ซับซ้อนและสามารถประมวลผลเพื่อสังเกตและจัดกลุ่มลักษณะเด่นที่มีความสัมพันธ์กัน

2.2.3 การหาความสัมพันธ์ในข้อมูล: ขั้นตอนวิธีการที่ใช้คือขั้นตอนวิธี Apriori ซึ่งเป็นวิธีที่ได้รับความนิยม

2.3 ขั้นตอนการประเมินคุณภาพของผลลัพธ์ที่ได้จากการทดลองประมวลผล หลังจากสรุปผลที่ได้จากการทำงานของขั้นตอนวิธีเป็นกฎและความสัมพันธ์ที่ได้จากการทดลองแล้ว ผู้วิจัยวัดค่า sensitivity ค่า specificity และค่า accuracy ตามสมการ (18) (19) และ (20) ตามลำดับ

2.4 การนำเสนอความรู้แก่ผู้ใช้ ผู้วิจัยนำเสนอผลการทดลอง และรายงานแก่ผู้เชี่ยวชาญ ณ โรงพยาบาลศิริราช เพื่อตรวจสอบความถูกต้องของผลลัพธ์ที่ได้ และปรับปรุงค่าของคุณลักษณะเด่นที่เลือกใช้ให้เหมาะสมกับการทดลองมากที่สุด

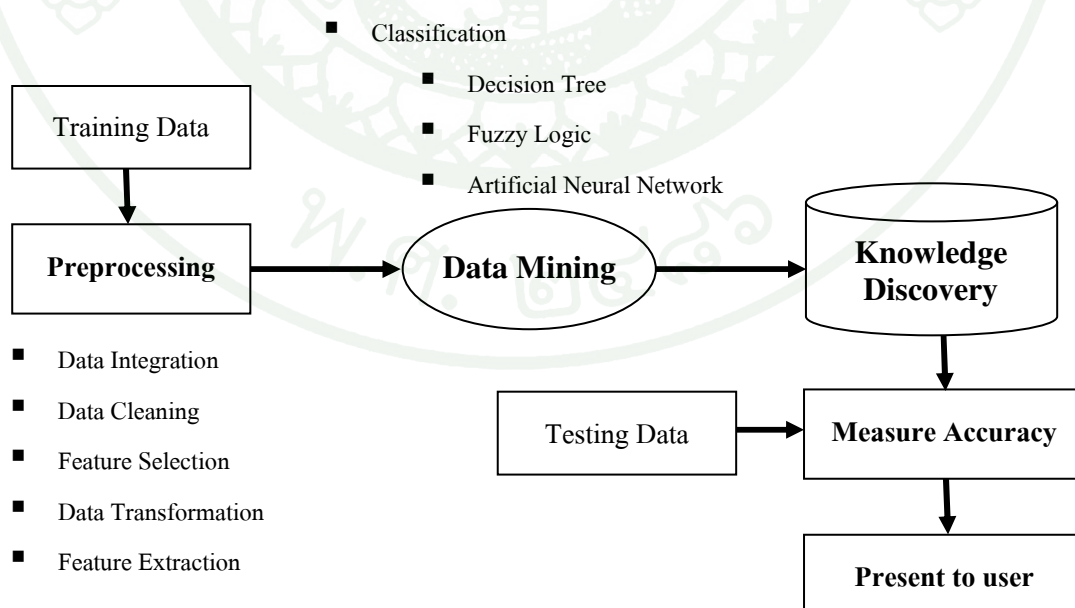
3. วิธีการทดลอง

ชุดข้อมูลและแอททริบิวต์ของการทำเหมืองข้อมูลโรคไขเลือดถูกออกแบบตามวัตถุประสงค์ของโรงพยาบาลวิจัย ดังนั้นจำนวนแอททริบิวต์และคลาสในแต่ละการทดลองอาจแตกต่างกัน

3.1 โจทย์ที่ 1: ต้องการค้นหาแอททริบิวต์ที่มีผลช่วยในการจำแนกประเภทของโรค ไข้เลือดออก แหล่งข้อมูลจากโรงพยาบาลศรีนครินทร์ จังหวัดขอนแก่น จำนวน 524 คน ผลเฉลยของข้อมูลผู้ป่วยโรคไข้เลือดออกประกอบด้วยคลาส 4 คลาสคือไข้เด็งกี (DF) ไข้เลือดออกระดับ 1 (DHF I) ไข้เลือดออกระดับ 2 (DHF II) และไข้เลือดออกระดับ 3 (DHF III) ชุดข้อมูลโจทย์ที่ 1 ประกอบด้วยคลาส DF 240 คน คลาส DHF I 116 คน คลาส DHF II 118 คน และ คลาส DHF III 50 คน และแหล่งข้อมูลจากโรงพยาบาลสงขลานครินทร์ จังหวัดสงขลา จำนวน 477 คน ประกอบด้วยคลาส DF 248 คน คลาส DHF I 106 คน คลาส DHF II 111 คน และคลาส DHF III 12 คน รวมจำนวนผู้ป่วยในการทดลองทั้งหมด 1001 คน ข้อมูลชุดที่ทดลองไม่มีข้อมูลของผู้ป่วยคลาส DHF IV เนื่องจากไม่มีผู้ป่วยคนใดที่มีอาการรุนแรงถึงระดับ DHF IV

ชุดของแอททริบิวต์มีทั้งหมด 48 แอททริบิวต์ ประกอบด้วยแอททริบิวต์ชนิดตัวเลขจำนวน 25 แอททริบิวต์ ชนิดคุณภาพจำนวน 22 แอททริบิวต์ และผลเฉลย 1 แอททริบิวต์ ค่าข้อมูลที่เป็น Temporal pattern จะเลือกข้อมูลในช่วง 4 วันแรกที่เข้ารับรักษาเท่านั้น

ขั้นตอนการทำเหมืองข้อมูลเริ่มจากการเตรียมข้อมูลและเลือกขั้นตอนวิธีการทำเหมืองข้อมูลโรคไข้เลือดออก และนำผลลัพธ์เสนอต่อผู้เชี่ยวชาญเมื่อได้ผลลัพธ์ที่มีประสิทธิภาพ



ภาพที่ 9 แผนภาพการทำเหมืองข้อมูล

3.1.1 ขั้นตอนการเตรียมข้อมูลก่อนการทดลอง

(1) การรวบรวมข้อมูล ขั้นตอนนี้ได้รับความร่วมมือจากทางโรงพยาบาลศิริราช ในการรวบรวมข้อมูลดิบจากการกรอกข้อมูลจากแบบฟอร์มของ 2 โรงพยาบาลคือโรงพยาบาลศรีนครินทร์และโรงพยาบาลสงขลานครินทร์

(2) การทำความสะอาดข้อมูล เป็นขั้นตอนวิธีการกำจัดข้อมูลที่เกิดจากการบันทึกผิดพลาด (noise) ด้วยการพิจารณาการรวมกลุ่มของข้อมูลเพื่อหาข้อมูลที่มีค่ามากหรือน้อยผิดปกติ (outlier detection) และแก้ปัญหาข้อมูลไม่ครบถ้วน (missing value) ในกรณีค่าของแอททริบิวต์เป็นค่าเชิงปริมาณค่าที่หายไปสามารถแทนด้วยค่าเฉลี่ยของแอททริบิวต์นั้น หรือในกรณีที่ค่าข้อมูลขาดหายเป็นค่าข้อมูลของแอททริบิวต์ที่สำคัญในการพิจารณาการจำแนกข้อมูล ข้อมูลของผู้ป่วยรายนั้นจะถูกตัดออกจากชุดข้อมูลที่พิจารณาก่อน

(3) การคัดเลือกลักษณะเด่น การทำงานได้รับความร่วมมือจากผู้เชี่ยวชาญจากโรงพยาบาลศิริราช อาจารย์ประพัฒน์ สุริยผล ในการคัดกรองแอททริบิวต์ที่ไม่มีความสำคัญในการจำแนกประเภทของโรคใช้เพื่อคัดออกเพื่อลดจำนวนแอททริบิวต์ก่อนการค้นหาค่าความรู้

(4) การแปลงข้อมูล เนื่องจากแหล่งข้อมูลมีมากกว่าหนึ่งแหล่ง ดังนั้นขั้นตอนการแปลงข้อมูลให้อยู่ในรูปแบบและความหมายเดียวกันจึงมีความสำคัญ ยกตัวอย่างเช่น แหล่งข้อมูลแรกกำหนดค่าของแอททริบิวต์ที่มีค่าเท่ากับ 1 หมายถึง ไม่มีอาการของโรค (No) แต่ในแหล่งข้อมูลที่ 2 กำหนดค่าของเลข 1 หมายถึง มีอาการของโรค (Yes) ซึ่งในส่วนนี้ทางผู้วิจัยขอขอบคุณความร่วมมือในการตรวจสอบความหมายดังกล่าวจากเจ้าหน้าที่ดูแลข้อมูลของโรงพยาบาลศิริราชเพื่อกำหนดความหมายของข้อมูลที่ต้องการและปรับให้ข้อมูลอยู่ในรูปแบบเดียวกัน แอททริบิวต์ที่แปลงค่าโดยตรงจากฐานข้อมูลแสดงในตารางที่ 5

ตารางที่ 5 แอททริบิวต์ที่ได้จากเอกสารการกรอกข้อมูลผู้ป่วย

ชื่อแอททริบิวต์ (Attribute Name)	ความหมาย
je	ผู้ป่วยได้รับวัคซีน JE vaccine หรือไม่ Yes หมายถึง ผู้ป่วยได้รับวัคซีน JE No หมายถึง ผู้ป่วยไม่ได้รับวัคซีน JE Unk หมายถึง ไม่ทราบว่าผู้ป่วยมีการฉีดวัคซีนหรือไม่

ตารางที่ 5 (ต่อ)

ชื่อแอททริบิวต์ (Attribute Name)	ความหมาย
leakage	ค่าแอททริบิวต์สรุปการพบอาการรั่วของพลาสติกในเลือดของผู้ป่วย Yes หมายถึง พบอาการรั่วของพลาสติกในเลือดของผู้ป่วย No หมายถึง ไม่พบอาการรั่วของพลาสติกในเลือดของผู้ป่วย
shock	ค่าแอททริบิวต์สรุปอาการช็อกของผู้ป่วย Yes หมายถึง พบอาการช็อกของผู้ป่วย No หมายถึง ไม่พบอาการช็อกของผู้ป่วย
dx	ผลการวินิจฉัยของโรคไข้เลือดออก วินิจฉัยโดยแพทย์มีผลการวินิจฉัยออกมาทั้งหมด 6 คลาส แต่ในงานวิจัยนี้ประกอบด้วย 4 คลาสคือ 1. DF หมายถึง ไข้เดงกี 2. DHF I หมายถึง เป็นไข้เลือดออกชนิดร้ายแรง ระดับที่ 1 3. DHF II หมายถึง เป็นไข้เลือดออกชนิดร้ายแรง ระดับที่ 2 4. DHF III หมายถึง เป็นไข้เลือดออกชนิดร้ายแรง ระดับที่ 3

(5) การสกัดลักษณะเด่นของข้อมูล (Feature Extraction) เป็นขั้นตอนการสร้างลักษณะเด่นของข้อมูลหรือสร้างแอททริบิวต์เพิ่มเติมเพื่อเป็นตัวแทนความหมายข้อมูลที่ต้องการ และยังส่งผลทำให้ผลลัพธ์ของการทำเหมืองข้อมูลมีคุณภาพดีขึ้น

ในงานวิจัยนี้ทางผู้วิจัยหาตัวแทนข้อมูล ด้วยการหาค่าเฉลี่ย, ค่าน้อยสุด, ค่ามากที่สุด และลักษณะแนวโน้มของข้อมูล เพื่อเป็นตัวแทนของข้อมูล โดยในขั้นตอนนี้ได้รับความร่วมมือจากแพทย์ในการตรวจสอบค่าลักษณะเด่นที่สร้างใหม่ว่าสอดคล้องกับงานทดลองหรือไม่ ดังนั้นได้ลักษณะเด่นที่เลือกใช้ในการทดลองทั้งหมดจำนวน 44 แอททริบิวต์ ประกอบด้วยแอททริบิวต์ชนิดคุณภาพจำนวน 19 แอททริบิวต์ และชนิดตัวเลขจำนวน 25 แอททริบิวต์ ดังแสดงในตารางที่ 6 และตารางที่ 7

ตารางที่ 6 แอททริบิวต์ชนิดคุณภาพที่สร้างเพื่อใช้ในการทดลอง

ชื่อแอททริบิวต์	ความหมาย
Bleeding	ผู้ป่วยมีอาการเลือดออกจากการสังเกตด้วยตาหรือไม่ การสรุปค่าพิจารณาจากสองเอกสารคือเอกสารที่บันทึกเมื่อตอนพบแพทย์ครั้งแรกและเอกสารที่บันทึกเมื่อตอนรักษาตัว Yes หมายถึง พบอาการเลือดออก No หมายถึง ไม่พบอาการเลือดออก
uri	ผู้ป่วยมีอาการติดเชื้อที่ระบบทางเดินหายใจตั้งแต่เข้าพบแพทย์ครั้งแรกหรือไม่ Yes หมายถึง ผู้ป่วยมีอาการติดเชื้อ No หมายถึง ผู้ป่วยไม่มีอาการติดเชื้อ Unk หมายถึง ผู้ป่วยไม่ได้ตรวจหรือตรวจแล้วแต่ไม่ทราบตำแหน่งที่มีการติดเชื้อ
pulse_pre_min_found	แอททริบิวต์สรุปการพบค่าของความดันที่น้อยที่สุดที่มีค่าต่างกันน้อยกว่า 20 ในแต่ละวันที่เจ้าหน้าที่พิจารณาแล้วและกรอกค่าลงในแบบฟอร์ม พิจารณาเฉพาะค่า 4 วันแรกเท่านั้น Yes หมายถึง พบค่าความต่างของความดันต่ำกว่า 20 มิลลิเมตรปรอท No หมายถึง ไม่พบค่าความต่างของความดันต่ำกว่า 20 มิลลิเมตรปรอท
itching_found	แอททริบิวต์สรุปอาการคันคันของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการคันคันในผู้ป่วย No หมายถึง ไม่พบอาการคันคันในผู้ป่วย

ตารางที่ 6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
quantity_max_found	แอททริบิวต์สรุปจำนวนจุดเลือดมากที่สุด ค่าของแอททริบิวต์เป็นช่วงของจำนวนจุดและเลือกตามช่วงที่พบมากที่สุดในช่วงรักษาตัว 4 วันแรกที่รักษาตัว R1 หมายถึง ช่วงจำนวนจุดเลือด 0 – 10 จุด R2 หมายถึง ช่วงจำนวนจุดเลือด 11 – 20 จุด R3 หมายถึง ช่วงจำนวนจุดเลือดมากกว่า 20 จุด
rash_found	แอททริบิวต์สรุปอาการผื่นขึ้นของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว ถ้ามีการตรวจแล้วพบว่ามีอาการผื่นขึ้นที่ตำแหน่งใดๆก็ตามจะสรุปคำว่ามีอาการผื่นขึ้น Yes หมายถึง พบอาการผื่นขึ้นในผู้ป่วย No หมายถึง ไม่พบอาการผื่นขึ้นในผู้ป่วย
bruising_found	แอททริบิวต์สรุปอาการฟกช้ำหลังการเจาะเลือดของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการฟกช้ำหลังการเจาะเลือดในผู้ป่วย No หมายถึง ไม่พบอาการฟกช้ำหลังการเจาะเลือดในผู้ป่วย
diarrhea_found	แอททริบิวต์สรุปอาการท้องเสียของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการท้องเสียในผู้ป่วย No หมายถึง ไม่พบอาการท้องเสียในผู้ป่วย

ตารางที่ 6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
uri_found	แอททริบิวต์สรุปอาการติดเชื้อที่ระบบทางเดินหายใจของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการติดเชื้อที่ระบบทางเดินหายใจของผู้ป่วย No หมายถึง ไม่พบอาการติดเชื้อที่ระบบทางเดินหายใจของผู้ป่วย Unk หมายถึง ผู้ป่วยไม่ได้ตรวจหรือตรวจแล้วแต่ไม่ทราบ
abdominal_found	แอททริบิวต์สรุปอาการปวดท้องของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการปวดท้องของผู้ป่วย No หมายถึง ไม่พบอาการปวดท้องของผู้ป่วย
dyspnea_found	แอททริบิวต์สรุปอาการหายใจติดขัดของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการหายใจติดขัดของผู้ป่วย No หมายถึง ไม่พบอาการหายใจติดขัดของผู้ป่วย
ascites_found	แอททริบิวต์สรุปอาการมีน้ำรั่วในช่องท้องของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการมีน้ำรั่วในช่องท้องของผู้ป่วย No หมายถึง ไม่พบอาการมีน้ำรั่วในช่องท้องของผู้ป่วย
jaundice_found	แอททริบิวต์สรุปอาการดีซ่าน ตัวเหลืองของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการดีซ่าน ตัวเหลืองของผู้ป่วย No หมายถึง ไม่พบอาการดีซ่าน ตัวเหลืองของผู้ป่วย

ตารางที่ 6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
liver_tenderness	แอททริบิวต์สรุปอาการเจ็บเมื่อกดที่บริเวณตับของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบอาการเจ็บเมื่อกดที่บริเวณตับของผู้ป่วย No หมายถึง ไม่พบอาการเจ็บเมื่อกดที่บริเวณตับของผู้ป่วย
liver_found	แอททริบิวต์สรุปอาการตับโตของผู้ป่วย พิจารณาจากค่าตับโตที่วัดได้ช่วงรักษาตัว 4 วันแรกที่รักษาตัว Yes หมายถึง พบค่าของตับโตของผู้ป่วย มีค่ามากกว่า 0 No หมายถึง ไม่พบอาการตับโตของผู้ป่วย
lymph_found	แอททริบิวต์สรุปอาการต่อมน้ำเหลืองโตของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว ถ้ามีการตรวจแล้วพบว่ามีอาการต่อมน้ำเหลืองโตที่ตำแหน่งใดๆก็ตามจะสรุปว่ามีอาการต่อมน้ำเหลืองโต Yes หมายถึง พบอาการอาการต่อมน้ำเหลืองโตในผู้ป่วย No หมายถึง ไม่พบอาการอาการต่อมน้ำเหลืองโตในผู้ป่วย
injected_found	แอททริบิวต์สรุปอาการตาอักเสบของผู้ป่วย พิจารณาจากช่วงรักษาตัว 4 วันแรกที่รักษาตัว ถ้ามีการตรวจแล้วพบว่ามีอาการตาอักเสบที่ตำแหน่งใดๆก็ตามจะสรุปว่ามีอาการตาอักเสบ Yes หมายถึง พบอาการอาการตาอักเสบในผู้ป่วย No หมายถึง ไม่พบอาการอาการตาอักเสบในผู้ป่วย

ตารางที่ 6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
atyp_found	แอททริบิวต์สรุปการพบความผิดปกติของเม็ดเลือดในผู้ป่วย Yes หมายถึง พบความผิดปกติของเม็ดเลือดในผู้ป่วย No หมายถึง ไม่พบความผิดปกติของเม็ดเลือดในผู้ป่วย เนื่องจากแอททริบิวต์ Atypical lymphocyte เป็นแอททริบิวต์ที่สนใจ เป็นพิเศษว่าแอททริบิวต์นี้เป็นตัวบ่งบอกว่ามีอยู่ในผู้ป่วยไข้เลือดออก ทุกคนหรือไม่
Effusion_Result	ค่าแอททริบิวต์อาการร้าวของน้ำในปอด Positive หมายถึง ผู้ป่วยมีอาการร้าวของน้ำในปอด Negative หมายถึง ผู้ป่วยไม่มีอาการร้าวของน้ำในปอด พิจารณาจากข้อมูลช่วงเวลาที่รักษา ถ้ามีอาการจะสรุปผลว่ามีค่าเป็น Positive

ตารางที่ 7 แอททริบิวต์ชนิดตัวเลขที่สร้างเพื่อใช้ในการทดลอง

ชื่อแอททริบิวต์	ความหมาย
temp_max	อุณหภูมิมากที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (องศาเซลเซียส) เนื่องจากข้อมูลในเอกสารเป็นแบบ Temporal data ดังนั้นจึงได้สรุปค่า แอททริบิวต์โดยเลือกเอาค่ามากที่สุดภายใน 4 วันแรกที่รักษาตัวเพื่อดู ความสัมพันธ์ว่าอุณหภูมิมีผลต่อการวินิจฉัยโรคหรือไม่
temp_min	อุณหภูมิน้อยที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (องศาเซลเซียส) พิจารณาค่ามากที่สุดภายใน 4 วันแรกที่รักษาตัวเท่านั้น

ตารางที่ 7 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
sbp_minus_dbp_avg	ค่าความต่างระหว่างค่าสูง-ต่ำของความดันโดยเฉลี่ย (mm.Hg: มิลลิเมตรปรอท) แอททริบิวต์ถูกสร้างจากการพิจารณาค่าความดันตัวหน้า ลบกับ ตัวหลัง เพื่อเป็นข้อมูลในแต่ละวันของผู้ป่วยคนหนึ่ง หลังจากนั้นนำมาคิดค่าเฉลี่ยเพื่อดูความสัมพันธ์ว่าอุณหภูมิมีผลต่อการวินิจฉัยโรคหรือไม่
hct_max	ค่าความเข้มข้นของเลือดมากที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่ามากที่สุดภายใน 4 วันแรกที่รักษาตัวเท่านั้น
hct_min	ค่าความเข้มข้นของเลือดน้อยที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่ามากที่สุดภายใน 4 วันแรกที่รักษาตัวเท่านั้น
liver_size_avg	ค่าขนาดของตับโดยเฉลี่ยค่าเฉลี่ยของค่าของตับที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่ามากที่สุดภายใน 4 วันแรกที่รักษาตัวเท่านั้น
hct_max_d05	ค่าความเข้มข้นของเลือดมากที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
hct_min_d05	ค่าความเข้มข้นของเลือดน้อยที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
hct_avg_d05	ค่าความเข้มข้นเฉลี่ยของเลือด (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวันและสรุปผลเป็นค่าเฉลี่ย
wbc_max	ค่าเม็ดเลือดขาว (x1000 cells) มากที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
wbc_min	ค่าเม็ดเลือดขาว (x1000 cells) น้อยที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
wbc_avg	ค่าเม็ดเลือดขาว (x1000 cells) เฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
platelet_max	ค่าปริมาณเกล็ดเลือด (x1000 cells) มากที่สุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน

ตารางที่ 7 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
platelet_min	ค่าปริมาณเกล็ดเลือด (x1000 cells) น้อยสุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
platelet_avg	ค่าปริมาณเกล็ดเลือด (x1000 cells) เฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
protein_avg	ค่าโปรตีนในเลือดเฉลี่ยของแอททริบิวต์ที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_albumin_avg	ค่าของ abumin โดยเฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
globurin_avg	ค่าของ protine-ablumin โดยเฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย (%) พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
ratio_abumin_avg	ค่าสัดส่วนเฉลี่ยของ abumin หารด้วย globurin แต่คิดย่อยเป็นเรคคอร์ด แล้วนำมาหาค่าเฉลี่ยอีกครั้ง พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_ast_max	ค่า AST มากสุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_ast_min	ค่า AST น้อยสุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_ast_avg	ค่า AST เฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_alt_max	ค่า ALT มากสุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน
lft_alt_min	ค่า ALT น้อยสุดที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน

ตารางที่ 7 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
lft_alt_avg	ค่า ALT เฉลี่ยที่วัดได้ในแต่ละวันของผู้ป่วย พิจารณาค่าข้อมูลของผู้ป่วยจากข้อมูลการรักษาแต่ละวัน

3.1.2 ขั้นตอนการทำเหมืองข้อมูล

สำหรับโจทย์ข้อ 1 ผู้วิจัยเลือกใช้ขั้นตอนวิธี 3 วิธีคือ

3.1.2.1 ขั้นตอนวิธีต้นไม้ตัดสินใจ เนื่องจากการทำงานของขั้นตอนวิธีนี้จะแสดงผลการเลือกแอททริบิวต์ที่สามารถในการจำแนกสูงมาเป็นโหนดในการสร้างต้นไม้ตัดสินใจ โหนดที่อยู่บนสุดของต้นไม้มีค่าความสามารถในการจำแนกสูงกว่าโหนดด้านล่าง และผลลัพธ์ของวิธีการต้นไม้ตัดสินใจยังสามารถอธิบายกฎการจำแนกตามกิ่งต่างๆของต้นไม้ตัดสินใจที่ได้ เครื่องมือของขั้นตอนวิธีนี้คือโปรแกรมเวก้า เลือกตัวแบบ J48 ซึ่งเป็นวิธีที่พัฒนาจาก C4.5 และสามารถจำแนกได้ทั้งข้อมูลชนิดตัวเลขและข้อมูลชนิดคุณภาพ

3.1.2.2 ขั้นตอนวิธีฟิชเชอร์ลอจิก หลังจากได้ผลลัพธ์จากต้นไม้ตัดสินใจแล้ว ทางผู้วิจัยพบว่าแอททริบิวต์ชนิดตัวเลขบางแอททริบิวต์ที่ถูกนำมาเป็นโหนดในต้นไม้ มีการนำตัวเลขค่าหนึ่งมาเป็นค่าที่จำแนกคลาสต่างๆออกจากกัน ยกตัวอย่างเช่น ถ้ามีแอททริบิวต์ที่บ่งบอกอุณหภูมิของผู้ป่วยซึ่งเป็นชนิดตัวเลขปรากฏอยู่ในต้นไม้ และมีค่าตัวเลข 36.5 เป็นค่าที่ใช้จำแนกบอกว่าผู้ป่วยมีไข้หรือไม่ไข้ ปัญหาที่เกิดขึ้นมีสองประเด็นด้วยกัน ประเด็นแรกคือ ค่าที่ต้นไม้เลือกมาเพื่อตัดสินใจคลาสนั้นอาจไม่ใกล้เคียงกับความเป็นจริงเพราะว่าค่าตัวเลขที่ถูกเลือกเป็นค่าการประมาณการจากข้อมูลจริงเท่านั้น และประเด็นที่สองคือ ถ้ามีข้อมูลผู้ป่วยที่มีค่าอุณหภูมิเท่ากับ 36.6 ต้นไม้ตัดสินใจจะตัดสินใจว่าผู้ป่วยรายนั้นมีไข้ในทันที ซึ่งในความเป็นจริงแล้วผู้ป่วยคนดังกล่าวอาจไม่มีไข้หรือมีไข้ต่ำๆเท่านั้น ดังนั้นทางผู้วิจัยจึงได้เลือกวิธีการฟิชเชอร์ลอจิกเข้ามาช่วยในการวิเคราะห์เพื่อยืนยันผลและพยายามปรับปรุงค่าประสิทธิภาพในการจำแนกให้มีประสิทธิภาพมากขึ้น

วิธีการฟัซซี่ลอจิกเริ่มต้นด้วยการนำค่าปกติของแอททริบิวต์ที่นำมาทดลองมาสร้างฟังก์ชันความเป็นสมาชิกรูปแบบต่างๆ ขั้นตอนการทำงานของวิธีการฟัซซี่ลอจิกมีดังนี้

(1) สร้างฟังก์ชันความเป็นสมาชิกให้แก่ตัวแปรนำเข้าหรือแอททริบิวต์ที่เลือกใช้ด้วยค่าปกติของแต่ละแอททริบิวต์ และสร้างฟังก์ชันความเป็นสมาชิกให้แก่ตัวแปรออก หรือคลาสให้มีจำนวนเท่ากับคลาสที่ต้องการพิจารณา

(2) ใส่กฎที่ใช้ในการพิจารณาค่าระดับ (degree) ของคลาสแต่ละคลาสโดยกฎมีรูปแบบเป็นคำสั่งเงื่อนไข ถ้า...แล้ว (If ... then)

(3) สร้างฟังก์ชันในการใส่ข้อมูลให้ระบบฟัซซี่ลอจิกทำงานตามขั้นตอนคือ Fuzzification, Rule Inference, Rule Composition และ Defuzzification

ในการทดลองนี้เลือกรูปแบบเกาส์เซียน สำหรับค่าแอททริบิวต์เชิงคุณภาพคือมีค่าใช่ หรือ ไม่ใช่ ข้อมูลจะถูกแปลงให้อยู่ในรูปแบบของตัวเลขที่เป็นตัวแทนข้อมูลกล่าวคือ ถ้าใช่มีค่าเท่ากับ 1 ถ้าไม่ใช่มีค่าเท่ากับ 0 เป็นต้น

3.1.2.3 การหาค่าสหสัมพันธ์ ค่าสหสัมพันธ์หรือค่า correlation เป็นการทดลองเพื่อยืนยันว่าแอททริบิวต์ที่พบในต้นไม้ตัดสินใจมีความสัมพันธ์กับคลาสหรือไม่ ถ้ามีความสัมพันธ์เป็นความสัมพันธ์ในทิศทางไหนและมีระดับความสัมพันธ์เท่าใด

3.1.3 ขั้นตอนการวัดประสิทธิภาพหรือขั้นตอนการประเมินคุณภาพของผลลัพธ์

ผู้วิจัยได้วางแผนการประเมินคุณภาพโดยใช้วิธีการ 10 cross validation ซึ่งเป็นวิธีการแบ่งข้อมูลออกเป็น 10 ชุด และการทดลองทั้งหมด 10 รอบ โดยในแต่ละรอบใช้ข้อมูล 9 ชุดในการเรียนรู้ และข้อมูล 1 ชุดในการทดสอบตัวแบบที่สร้างขึ้น รวมทั้งคำนวณค่าประสิทธิภาพของผลลัพธ์ด้วยการวัดค่าความไวในการตรวจจับผู้ป่วยที่เป็นโรคได้ถูกต้อง (sensitivity) การวัดค่าความสามารถในการคัดกรองผู้ป่วยที่ไม่เป็นโรค (specificity) และการวัดค่าความถูกต้องที่บ่งบอกถึงประสิทธิภาพรวมของตัวแบบที่สร้างขึ้น (accuracy) นอกจากนี้ทางผู้วิจัยเลือกวิธีการหาค่าสหสัมพันธ์เพื่อตรวจสอบความสัมพันธ์ระหว่างแอททริบิวต์ที่พบในต้นไม้กับคลาสของข้อมูลด้วย

3.1.4 การนำเสนอองค์ความรู้แก่ผู้เชี่ยวชาญ

เมื่อผู้วิจัยได้ผลลัพธ์ที่มีค่าความสามารถในการจำแนกประเภทของไข้เลือดออกที่พอใจ ผู้วิจัยจะรายงานผลของแอททริบิวต์สำคัญและรูปแบบกฎแก่ผู้เชี่ยวชาญเพื่อตรวจสอบและวิเคราะห์ผลลัพธ์ว่ามีความถูกต้องและสามารถนำไปใช้ได้จริง

3.2 โจทย์ที่ 2: รูปแบบของแอททริบิวต์ที่ส่งผลให้สามารถจำแนกคลาสแต่ละคลาสผลลัพธ์สามารถอธิบายได้จากต้นไม้ตัดสินใจจากโจทย์ที่ 1 รูปแบบของแอททริบิวต์สำหรับแต่ละคลาส สามารถอธิบายได้ด้วยกฎที่ปรากฏในแต่ละกิ่งของต้นไม้

3.3 โจทย์ที่ 3: แพทย์ต้องการทราบวันที่ผู้ป่วยมีอาการไข้สูงเพื่อหาวิธีป้องกัน เนื่องจากภายใน 24 ชั่วโมงระหว่างนั้นผู้ป่วยมักมีปริมาณเกล็ดเลือดต่ำที่สุด มีการรั่วของพลาสมาสูงที่สุด และอาจเกิดการช็อกแต่เมื่อผ่านช่วงนี้ไปแล้ว ผู้ป่วยจะฟื้นตัวอย่างรวดเร็ว

สำหรับโจทย์นี้ทางผู้วิจัยได้เลือกขั้นตอนวิธีดังนี้

3.3.1 ขั้นตอนวิธีต้นไม้ตัดสินใจ ปัญหาที่เกิดขึ้นคือเมื่อนำข้อมูลสร้างต้นไม้ตัดสินใจด้วยแอททริบิวต์ที่ใช้งานทั้งหมด 38 แอททริบิวต์ เกิดปัญหาต้นไม้ตัดสินใจจำข้อมูลมากเกินไป (overfitting tree)

3.3.2 การหาค่าสหสัมพันธ์ เพื่อต้องการแอททริบิวต์ที่มีความสำคัญในการทำนาย day0 ซึ่งในปัจจุบันยังไม่สามารถระบุอย่างชัดเจน วิธีการนี้ใช้เพื่อหาแอททริบิวต์ที่มีค่าสหสัมพันธ์ที่มีความสัมพันธ์กับคลาสมากกว่า 0.5 เพื่อให้เป็นแอททริบิวต์เข้าของขั้นตอนวิธีฟิชเชอร์ลอจิก

3.3.3 ขั้นตอนวิธีฟิชเชอร์ลอจิก แอททริบิวต์ที่นำไปสร้างตัวแปรเข้าในวิธีฟิชเชอร์ลอจิกมีทั้งหมด 7 ตัวคือ จำนวนเม็ดเลือดขาว (wbc), จำนวนเกล็ดเลือด (platelet), อาการช็อก (shock), การรั่วของพลาสมาในเลือด (leakage), อาการเลือดออก (Bleeding), ค่าเอนไซม์ ALT ในตับ และผลสรุปอาการรั่วของน้ำในปอด (Effusion Result) แต่ก่อนนำมาสร้างเป็นตัวแปรเข้าจำเป็นต้องสร้างต้นไม้ตัดสินใจจากชุดแอททริบิวต์ที่เลือกก่อน เพื่อดูกฎจากต้นไม้

หลังจากที่ได้ผลลัพธ์จากต้นไม้ตัดสินใจ ผู้วิจัยได้สร้างชุดแอททริบิวต์ขึ้นมาหนึ่งชุดเพื่อเป็นแอททริบิวต์ในการทำนายเบื้องต้น กระบวนการทำงานของฟัซซี่ลอจิกดำเนินการตามขั้นตอนดังนี้

(1) สร้างฟังก์ชันความเป็นสมาชิกของแอททริบิวต์ที่เลือกมาทดลอง รวมถึงสร้างฟังก์ชันความเป็นสมาชิกของคลาส ชนิดของฟังก์ชันความเป็นสมาชิกที่เลือกคือฟังก์ชันความเป็นสมาชิกรูปแบบเกาส์เซียน (Guassian membership function) โดยค่าที่นำมาสร้างเป็นค่าปกติของข้อมูลแต่ละแอททริบิวต์

(2) สร้างกฎฟัซซี่เพื่อค้นหาคำตอบหรือค่าดีกรีของข้อมูลที่เป็นข้อมูลเข้า กฎที่สร้างในส่วนนี้ผู้วิจัยนำมาจากผลลัพธ์ต้นไม้ตัดสินใจเพื่อพิจารณารูปแบบของกฎเบื้องต้นก่อนการใส่กฎ จะใส่ในรูปแบบของคำสั่งเงื่อนไข ถ้า...แล้ว (If ... then) ยกตัวอย่างเช่น ถ้าอุณหภูมิสูงกว่า 37.5 แล้วผู้ป่วยรายนั้นถือว่าไข้สูง

(3) สร้างฟังก์ชันในการวัดค่าประสิทธิภาพของผลลัพธ์ที่ได้จากวิธีการฟัซซี่ลอจิกโดยอิงจากสูตรการคำนวณค่า sensitivity, specificity และค่า accuracy

จากวัตถุประสงค์ของโจทย์ข้อนี้คือการทำนาย ผู้วิจัยจึงได้ทดลองขั้นตอนวิธีโครงข่ายประสาทเทียมเข้ามามีส่วนหาคำตอบ เนื่องจากขั้นตอนวิธีโครงข่ายประสาทเทียมได้รับความนิยมในการนำมาสร้างระบบการทำนายข้อมูล อีกทั้งขั้นตอนวิธีนี้ใช้แนวการเรียนรู้เหมือนสมองมนุษย์เหมาะกับการจัดการกับข้อมูลที่ซับซ้อน

3.4 โจทย์ที่ 4: ผู้เชี่ยวชาญต้องการทราบว่าอาการตับโตเป็นปัจจัยที่ช่วยในการจำแนกประเภทของไข้เลือดออกหรือไม่ อาการตับโตของผู้ป่วยโรคไข้เลือดออกนั้นสามารถดูได้จากค่าขนาดของตับซึ่งเป็นข้อมูลที่บันทึกโดยแพทย์ ถ้าข้อมูลค่านี้มีค่ามากกว่า 0 แสดงว่าผู้ป่วยมีอาการตับโตเกิดขึ้น

สำหรับการหาบทสรุปของโจทย์การทดลองนี้ ผู้วิจัยได้พิจารณาจาก 3 วิธีดังนี้

3.4.1 ผลลัพธ์ต้นไม้ตัดสินใจ วิเคราะห์จากผลลัพธ์จากโจทย์ที่ 1

3.4.2 ค่าสหสัมพันธ์ เพื่อพิจารณาความสัมพันธ์ของแอททริบิวต์อาการตับโตกับคลาสระดับความรุนแรงของผู้ป่วยโรคไข้เลือดออก

3.4.3 พิจารณาจากอัตราส่วนข้อมูลดิบ พิจารณาโดยแยกพิจารณาเป็นคลาส DF หรือไข้แดงก็และคลาส DHF หรือไข้เลือดออก ข้อมูลของผู้ป่วยประกอบด้วยคลาส DF จำนวน 488 คน และคลาส DHF จำนวน 513 คน

3.5 โจทย์ที่ 5: บุคคลใดได้รับวัคซีนป้องกันโรคไข้สมองอักเสบ (Japanese Encephalitis: JE) แล้ว สามารถเป็นโรคไข้เลือดออกได้อีกหรือไม่

สำหรับโจทย์ที่ 5 ผู้วิจัยเลือกวิธีการทดลองดังนี้

3.5.1 ขั้นตอนวิธีต้นไม้ตัดสินใจ โดยข้อมูลที่ใช้ในการทดลองนี้ตัดข้อมูลของผู้ป่วยที่มีค่าของแอททริบิวต์ je ที่มีค่าเป็น Unknown ซึ่งหมายความว่าไม่ทราบว่าการฉีดวัคซีน JE หรือไม่ ออกไปจากชุดข้อมูล ดังนั้นชุดข้อมูลในการทดลองประกอบด้วยข้อมูลผู้ป่วย 714 คน และจำนวนแอททริบิวต์ทั้งหมด 48 แอททริบิวต์

3.5.2 การคำนวณค่าสหสัมพันธ์ หลังจากได้ผลลัพธ์ต้นไม้ตัดสินใจแล้วผู้วิจัยได้เลือกใช้วิธีการนี้เพื่อวัดระดับความสัมพันธ์ของแอททริบิวต์ je ที่สนใจ และต้องการทราบว่าแอททริบิวต์ je ส่งผลต่อการไม่เกิดโรคไข้เลือดออกด้วยหรือไม่

3.5.3 การหาคุณภาพความสัมพันธ์ ค่าของข้อมูลแต่ละแอททริบิวต์ที่ใช้ในขั้นตอนวิธีนี้ จำเป็นต้องเป็นข้อมูลเชิงคุณภาพ ดังนั้นผู้วิจัยจึงนำแอททริบิวต์ที่มีค่าเป็นตัวเลขมาปรับโดยการใส่ค่าขอบเขตปกติช่วยในการแปลงข้อมูลคือ ต่ำ ปกติ และสูง แต่แอททริบิวต์ใดที่ไม่มีค่าปกติในการเปรียบเทียบทางผู้วิจัยจะตัดทิ้งไม่พิจารณาในการทดลองนี้ ยกตัวอย่างเช่น แอททริบิวต์ค่าความต่างของความดันเลือด (sbp_dbp) เป็นต้น การทดลองโจทย์นี้แยกพิจารณาเป็นคลาส DF และ DHF และผู้ป่วยคนใดที่มีค่าแอททริบิวต์ je เป็น Unknown ผู้ป่วยรายนั้นจะถูกตัดออกจากการพิจารณาในการทดลองเช่นเดียวกัน

ข้อมูลที่ใช้ในการทดลองประกอบด้วยข้อมูลผู้ป่วยคลาส DF จำนวน 364 คน และข้อมูลผู้ป่วยคลาส DHF จำนวน 349 คน แอททริบิวต์ที่ใช้ในการทดลองมีทั้งหมด 31 แอททริบิวต์ดังต่อไปนี้

je	bruising_found	liver_tenderness	atyp_found
Bleeding	diarrhea_found	liver_found	protien
uri	uri_found	lymph_found	albumin
pulse_pre_min_found	abdominal_found	injected_found	globurin
quantity_max_found	dyspnea_found	hct_d05	AST
rash_found	ascites_found	wbc	ALT
itching_found	jaundice_found	platelet	EffusionResult
leakage	shock	dx	

ขั้นตอนวิธีการหาความสัมพันธ์ในการทดลองคือ วิธีการ Apriori ค่าพารามิเตอร์ 2 ค่าที่ต้องกำหนด คือค่าความถี่ขั้นต่ำ (Minimum Support) และค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) โปรแกรมเวก้าสามารถกำหนดค่าความถี่ขั้นต่ำที่พารามิเตอร์ lowerBoundMinSupport และค่าความเชื่อมั่นขั้นต่ำที่พารามิเตอร์ minMetric นอกจากนั้นยังสามารถปรับค่าจำนวนกฎเพื่อหาความสัมพันธ์ที่มากขึ้น ค่าพารามิเตอร์ที่กำหนดคือ ค่าความถี่ขั้นต่ำมีค่าเท่ากับ 0.1 (ค่าตั้งต้น) และ 0.05 และกำหนดค่าความน่าเชื่อถือขั้นต่ำเท่ากับ 0.9 จำนวนกฎตั้งค่าเท่ากับ 10, 100 และ 1,000 กฎ

3.6 โจทย์ที่ 6: สำหรับผู้ป่วยโรคไข้เลือดออก ค่าเอนไซม์ Aspartate aminotransferase (AST) ที่บ่งบอกการทำงานของตับมีค่าสูงกว่าค่าของเอนไซม์ Alanine aminotransferase (ALT) เสมอหรือไม่

การทดลองนี้ผู้วิจัยเลือกใช้ 2 วิธีดังนี้

3.6.1 การหาความสัมพันธ์ สำหรับการหาความสัมพันธ์ ผู้วิจัยสร้างแอททริบิวต์ใหม่ชื่อว่า AST_ALT ซึ่งเป็นแอททริบิวต์สรุปผลค่า AST มากกว่าค่า ALT หรือไม่ ถ้ามากกว่าค่าแอททริบิวต์เป็น Yes แต่ถ้าไม่มากกว่าค่าแอททริบิวต์เป็น No ดังนั้นชุดข้อมูลที่ใช้โจทย์ที่ 6 ประกอบด้วย 32 แอททริบิวต์ดังต่อไปนี้

je	bruising_found	liver_tenderness	atyp_found
Bleeding	diarrhea_found	liver_found	protien
uri	uri_found	lymph_found	albumin

pulse_pre_min_found	abdominal_found	injected_found	globulin
quantity_max_found	dyspnea_found	hct_d05	AST
rash_found	ascites_found	wbc	ALT
itching_found	jaundice_found	platelet	EffusionResult
AST_ALT	leakage	shock	dx

พิจารณาแยกตามคลาส จำนวนผู้ป่วยคลาส DF 487 คน และจำนวนผู้ป่วยคลาส DHF 513 คน การทดลองมีรายละเอียดดังนี้ พิจารณาจากกฎความสัมพันธ์ด้วยวิธี Apriori กำหนดค่าความถี่ขั้นต่ำคือ 0.1 และค่าความเชื่อมั่นขั้นต่ำเท่ากับ 0.9 จำนวนกฎที่พิจารณาคือ 10 กฎ และ 100 กฎ

3.6.2 การพิจารณาสัดส่วนของข้อมูลจริง พิจารณาจากสัดส่วนข้อมูลจริงของผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่าค่าเอนไซม์ ALT จากข้อมูลในฐานข้อมูล

3.7 โจทย์ที่ 7: กลุ่มผู้ป่วยที่เป็นโรคไข้เลือดออกมีค่าความผิดปกติของเม็ดเลือดขาว (Atypical Lymphocyte) ทุกรายหรือไม่

ผู้วิจัยแบ่งวิธีการพิจารณาคำตอบออกเป็น 2 วิธีดังนี้

3.7.1 พิจารณาจากการหากฎความสัมพันธ์ การพิจารณาโจทย์นี้สามารถดูผลลัพธ์จาก โจทย์ที่ 5 และโจทย์ที่ 6 มาประกอบการพิจารณาได้

3.7.2 พิจารณาจากสัดส่วนในฐานข้อมูลโดยพิจารณาแยกตามคลาส DF และ DHF

3.8 โจทย์ที่ 8: บั๊จยี่ที่สามารจจำแนกผู้ป่วยที่เป็นโรค DHF I ออกจาก DHF II และ DHF III สามารจค้นพบได้หรือไม่ ถ้าได้ อะไรบ้าง

เนื่องจากทางารแพทยัพพว่าข้อมูลทางพันธุกรรมบางส่วน ถ้านาคลาส DHF I ไปรวมกับคลาส DF และนาคลาส DHF II รวมกับ DHF III ให้ความสำคัญ (significant) บางอย่าง แต่ถ้

จัดเป็นคลาส DHF กลับไม่มีความสำคัญ ดังนั้นจึงอยากทราบว่าปัจจัยหรือแอททริบิวต์ใดมีความสำคัญที่สามารถจำแนก DHF I ออกจาก DHF II และ DHF III ได้

การพิจารณาคำตอบในโจทย์ที่ 8 ได้แบ่งวิธีการพิจารณาออกเป็น 3 วิธีดังนี้

3.8.1 การแบ่งกลุ่มข้อมูลด้วยวิธีการเคมีน (K-Mean Clustering) ซึ่งเป็นขั้นตอนวิธีแบบไม่มีผู้สอน เพื่อดูธรรมชาติของข้อมูลว่าสามารถแบ่งได้ตามสมมติฐานนั้นจริงหรือไม่ กำหนดค่า $k = 2$ เพื่อให้เท่ากับจำนวนคลาสในสมมติฐานที่เราต้องการทดสอบ ชุดแอททริบิวต์ที่ใช้ในการทดลองแยกพิจารณา 3 ชุดคือ ชุดแรกพิจารณาจากข้อมูลแอททริบิวต์ทั้งหมดโดยไม่ได้แยกตามชนิดของแอททริบิวต์ ชุดที่สองพิจารณาแอททริบิวต์ข้อมูลทางกายภาพ (Clinical Attributes) และชุดที่สามพิจารณาแอททริบิวต์จากผลตรวจห้องปฏิบัติการ (Laboratory Attributes)

3.8.2 ขั้นตอนวิธีต้นไม้ตัดสินใจ เพื่อต้องการทราบปัจจัยหรือแอททริบิวต์ที่มีความสำคัญและรูปแบบของการจำแนกคลาส ข้อมูลในการทดลองคือข้อมูลผู้ป่วยโรคไข้เลือดออกจำนวน 1001 คน แอททริบิวต์จำนวน 48 แอททริบิวต์ คลาสแบ่งใหม่ออกเป็น 2 คลาสคือ Group1 (Non-Severe) และ Group 2 (Severe) โดยที่ Group1 มีจำนวนผู้ป่วย 710 คนและใน Group2 มีจำนวนผู้ป่วย 291 คน

3.8.3 การหาค่าสหสัมพันธ์ ค่าสหสัมพันธ์หรือค่า correlation เป็นการทดลองเพื่อยืนยันว่าแอททริบิวต์ที่พบในต้นไม้ตัดสินใจมีความสัมพันธ์กับคลาสหรือไม่ ถ้ามีความสัมพันธ์เป็นความสัมพันธ์ในทิศทางไหนและมีระดับความสัมพันธ์เท่าใด

4. การวัดประสิทธิภาพ

งานวิจัยนี้วัดประสิทธิภาพด้วยค่าความไว ค่าความจำเพาะ และค่าความถูกต้อง ตามสมการหมายเลข (18) (19) และ (20) ตามลำดับ และในส่วนการเปรียบเทียบผลการทดลองบางโจทย์ที่สนใจ ผู้วิจัยได้คิดสัดส่วนค่า False negative ซึ่งเป็นค่าที่ระบบทายว่าไม่เป็นโรคแต่จริงๆแล้วผู้ป่วยรายนั้นเป็นโรค ซึ่งทางการแพทย์ให้ความสำคัญกับค่านี้มากกว่าค่าอื่นๆ รวมทั้งค่า False Positive ซึ่งเป็นค่าที่ระบบบอกว่าเป็นโรคแต่ผู้ป่วยจริงไม่ได้เป็นโรค ซึ่งค่านี้อาจมีความสำคัญทางด้านการได้รับการรักษาอย่างทั่วถึงและคุณภาพของการรักษาโรค

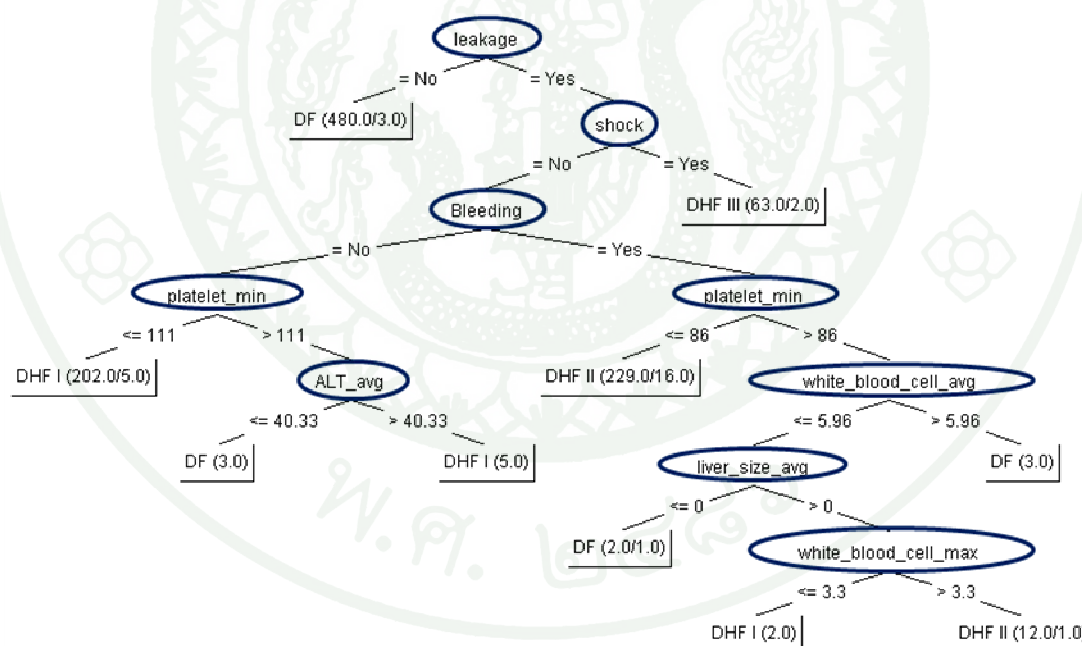
ผลและวิจารณ์

ผล

ผลลัพธ์จากการทดลองแยกสรุปตามโจทย์ปัญหา 8 ข้อซึ่งบางการทดลองสามารถใช้ผลการทดลองก่อนหน้านี้อธิบายแทนได้ ดังแสดงรายละเอียดต่อไปนี้

1. ผลการทดลองโจทย์ที่ 1 ค้นหาปัจจัยที่ช่วยจำแนกระดับของโรคไข้เลือดออก

ผลลัพธ์การทดลองโจทย์ที่ 1 แสดงผลในรูปของต้นไม้ตัดสินใจโดยแอททริบิวต์ที่อยู่ด้านบนของต้นไม้มีอำนาจการจำแนกได้ดีกว่าแอททริบิวต์ที่อยู่ด้านล่างของต้นไม้



ภาพที่ 10 ต้นไม้ตัดสินใจแสดงปัจจัยที่ช่วยในการจำแนกระดับของโรคไข้เลือดออก

จากผลลัพธ์ต้นไม้ตัดสินใจพบแอททริบิวต์ที่น่าสนใจที่สามารถจำแนกคลาสของผู้ป่วยได้แก่ อาการรั่วของพลาสมาในเลือด (leakage) อาการช็อก (shock) อาการเลือดออก (Bleeding) จำนวนเกล็ดเลือดน้อยที่สุดที่วัดได้ (platelet_min) ค่าเฉลี่ยของเอนไซม์ Alanine aminotransferase

ในตับ (ALT_avg) ค่าเฉลี่ยของจำนวนเม็ดเลือดขาว (wbc_avg) ค่าขนาดของอาการตับโต (liver_found) และค่ามากที่สุดของจำนวนเม็ดเลือดขาว (wbc_max)

หลังจากได้ผลลัพธ์จากต้นไม้ตัดสินใจ ผู้วิจัยได้นำเอาทริบิวต์และกฎของต้นไม้ตัดสินใจ มาทดลองด้วยวิธีการฟัซซี่ลอจิกเพื่อทดลองปรับปรุงการจำแนกของแอททริบิวต์ที่เป็นตัวเลขและ ประเมินค่าประสิทธิภาพของการจำแนกประเภทของไข้เลือดออกได้ดังตารางที่ 8

ตารางที่ 8 ค่าความถูกต้องของการทดลองโจทย์ที่ 1

Method	Class	Sensitivity (%)	Specificity (%)	Accuracy (%)	Overall Accuracy (%)
Decision Tree	DF	97.34	98.56	97.95	96.70
	DHF I	89.64	98.44	96.46	
	DHF II	95.63	96.84	96.56	
	DHF III	98.39	99.55	99.48	
FuzzyLogic	DF	97.75	99.39	98.57	97.94
	DHF I	91.89	98.44	96.98	
	DHF II	96.51	97.25	97.08	
	DHF III	98.39	99.78	99.69	

จากผลการทดลองตารางที่ 8 พบว่าได้ค่าความถูกต้องของวิธีต้นไม้ตัดสินใจเท่ากับ 96.70% ต่อมาเมื่อนำผลลัพธ์มาพัฒนาด้วยวิธีการฟัซซี่ลอจิกพบว่าค่าประสิทธิภาพในการจำแนกมีค่าสูงขึ้น ประมาณ 1% มีค่าเท่ากับ 97.94%

2. ผลการทดลองโจทย์ที่ 2 รูปแบบการเกิดของระดับโรคไข้เลือดออก

การทดลองโจทย์ที่ 2 ต้องการทราบรูปแบบของแอททริบิวต์สำหรับคลาสที่พิจารณา ดังนั้น คำตอบของโจทย์นี้สามารถสรุปผลได้จากผลลัพธ์ต้นไม้ตัดสินใจในการทดลองโจทย์ที่ 1 สามารถอธิบายได้ด้วยกฎทั้งหมด 10 กฎดังนี้

คลาส DF จำนวน 4 กฎดังนี้

- (1) ถ้าผู้ป่วยไม่พบอาการรั่วของพลาสมาในเลือด ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DF
- (2) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก ไม่มีอาการเลือดออก จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 111,000 cells/ μ l และค่าเฉลี่ยของเอนไซม์ลอทานินมีค่าน้อยกว่าหรือเท่ากับ 40.33 U/L แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DF
- (3) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก มีอาการเลือดออกด้วยการสังเกตได้ จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 86,000 cells/ μ l และค่าเฉลี่ยของจำนวนเม็ดเลือดขาวมีค่ามากกว่า 5,960 cells/ μ l แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DF
- (4) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก มีอาการเลือดออกด้วยการสังเกตได้ จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 86,000 cells/ μ l ค่าเฉลี่ยของจำนวนเม็ดเลือดขาวมีค่าน้อยกว่าหรือเท่ากับ 5,960 cells/ μ l และไม่มีอาการดับโต ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DF

คลาส DHF I จำนวน 3 กฎดังนี้

- (1) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก ไม่มีอาการเลือดออก และจำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่าน้อยกว่าหรือเท่ากับ 111,000 cells/ μ l แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF I
- (2) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อกไม่มีอาการเลือดออก จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 111,000 cells/ μ l และค่าเฉลี่ยของเอนไซม์ลอทานินมีค่ามากกว่า 40.33 U/L แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF I
- (3) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อกมีอาการเลือดออกด้วยการสังเกตได้จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 86,000 cells/ μ l ค่าเฉลี่ยของจำนวนเม็ดเลือดขาวมีค่าน้อยกว่าหรือเท่ากับ 5,960 cells/ μ l ไม่มีอาการดับโตและเมื่อจำนวนเม็ดเลือดขาวมากที่สุดมีค่าน้อยกว่าหรือเท่ากับ 3,300 cells/ μ l แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF I

คลาส DHF II จำนวน 3 กฎดังนี้

- (1) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก มีอาการเลือดออกด้วยการสังเกตได้ และจำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่าน้อยกว่าหรือเท่ากับ 86,000 cells/ μ l แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF II

(2) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด ไม่มีอาการช็อก มีอาการเลือดออกด้วยการสังเกตได้ จำนวนเกล็ดเลือดน้อยสุดที่วัดได้มีค่ามากกว่า 86,000 cells/ μ l ค่าเฉลี่ยของจำนวนเม็ดเลือดขาวมีค่าน้อยกว่าหรือเท่ากับ 5,960 cells/ μ l มีอาการตับโตและเมื่อจำนวนเม็ดเลือดขาวมากที่สุดมีค่ามากกว่า 3,300 cells/ μ l แล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF II

คลาส DHF III จำนวน 1 กฎดังนี้

(1) ถ้าผู้ป่วยมีอาการรั่วของพลาสมาในเลือด และมีอาการช็อกแล้ว ผู้ป่วยรายนั้นจะถูกจำแนกเป็นคลาส DHF III

3. ผลการทดลองโงทย์ที่ 3 ทำนายวัน day0 ของผู้ป่วยโรคไข้เลือดออก

เนื่องจากต้นไม้ตัดสินใจที่ได้จากการทดลองการทำนายข้อมูล day0 พบว่าเกิดปัญหา Overfitting Tree หรือปัญหาต้นไม้ที่จดจำข้อมูลฝึกมากเกินไป ดังนั้นผู้วิจัยจึงได้เลือกชุดของปัจจัยหรือแอททริบิวต์ที่มีความสัมพันธ์กับคลาสและค่าสหสัมพันธ์มากกว่า 0.5 เพื่อเป็นตัวแปรนำเข้าของขั้นตอนฟิชชีลอจิก แอททริบิวต์ที่เลือกมาทดลองโงทย์ที่ 3 ได้แก่

- (1) bleeding: ค่าสรุปอาการเลือดออกว่ามีอาการหรือไม่
- (2) wbc: ค่าปริมาณเม็ดเลือดขาวที่ตรวจพบ
- (3) platelet: ค่าเกล็ดเลือดของผู้ป่วย
- (4) alt: ค่าเอนไซม์ Alanine Aminotransferase
- (5) leakage: ค่าสรุปอาการรั่วของพลาสมาในเลือด
- (6) shock: ค่าสรุปอาการช็อก
- (7) EffusionResult: ค่าสรุปอาการรั่วของน้ำในช่องท้อง

หลังจากได้ชุดของแอททริบิวต์แล้ว ทางผู้วิจัยได้นำชุดแอททริบิวต์ดังกล่าวไปทดลองด้วยวิธีการต้นไม้ตัดสินใจอีกครั้งเพื่อหาคุณภาพของความสัมพันธ์ของแอททริบิวต์เพื่อนำไปใช้ในวิธีการฟิชชีลอจิก เมื่อได้คุณภาพต้นไม้ตัดสินใจแล้ว ผู้วิจัยดำเนินการตามขั้นตอนการทำฟิชชีลอจิก ซึ่งผลการทดลองจากการทำนายวัน day0 ของผู้ป่วยโรคไข้เลือดออกจำนวน 968 รายการให้ค่าความถูกต้องสูงขึ้นเป็น 65.45%

จากตารางที่ 9 พบว่าค่าความถูกต้องที่ได้จาก 2 ขั้นตอนวิธียังมีค่าไม่สูงมาก สาเหตุอาจเกิดจากแอททริบิวต์ที่เลือกมาใช้งานยังไม่มีอำนาจในการทำนายมากเท่าที่ควรรวมถึงจำนวนของข้อมูล day0 ในแต่ละคลาสจากฐานข้อมูลมีจำนวนมากน้อยไม่สมดุลกัน และในข้อมูลผู้ป่วยบางคนมีข้อมูลในแอททริบิวต์ที่นำมาใช้งานไม่ครบถ้วน

ตารางที่ 9 ค่าความถูกต้องของการทดลองโจทย์ที่ 3

Method	Class	Sensitivity (%)	Specificity (%)	Accuracy (%)	Overall Accuracy (%)
Decision Tree	day0	61.61	57.92	59.82	64.80
	day -1	51.56	54.66	53.43	
	day -2	23.08	85.74	71.11	
	day -3	10.00	96.91	88.79	
Fuzzy Logic	day0	50.86	75.13	64.88	65.45
	day -1	71.11	45.37	54.75	
	day -2	3.21	96.92	81.82	
	day -3	2.00	99.67	94.63	

หลังจากได้ผลลัพธ์ที่แสดงถึงแอททริบิวต์ที่อาจไม่มีอำนาจในการทำนายมากพอ ทางผู้วิจัยจึงได้ทดลองโจทย์นี้ด้วยขั้นตอนวิธีโครงข่ายประสาทเทียมโดยใช้จำนวนแอททริบิวต์ทั้งหมด 37 แอททริบิวต์ และคลาสที่เป็นผลเฉลยของวัน day0 และวันอื่นๆ โดยในการทดลองนั้นได้มีการตั้งค่าการทดลองที่แตกต่างกัน เพื่อดูว่าค่าใดที่ให้ผลลัพธ์ที่ดีที่สุด ค่าที่กำหนดในการทดลองได้แก่จำนวนรอบในการทำงานคือ 300 และ 1000 รอบ ค่าสมรรถนะของเป้าหมาย (Performance goal) หรือค่าความผิดพลาดน้อยสุดที่ยอมรับได้ คือ 0.05 และ 0.005 เป็นค่าที่ตั้งเพื่อเป็นเงื่อนไขในการหยุดการเรียนรู้ และอีกค่าหนึ่งคือ ค่าอัตราการเรียนรู้ ยิ่งกำหนดค่าสูงเท่าใดระบบจะยิ่งฉลาดมากเท่านั้น ค่าที่กำหนดในการทดลองคือ 0.3, 0.5 และ 0.8

ผลการทดลองได้ค่าความถูกต้องมากที่สุดเท่ากับ 69.80% เมื่อกำหนดจำนวนรอบไว้ที่ 300 รอบ ค่าเป้าหมายเท่ากับ 0.05 และมีอัตราการเรียนรู้เท่ากับ 0.5

4. ผลการทดลองไจท์ที่ 4 อาการตับโตมีส่วนช่วยในการจำแนกระดับของโรคไข้เลือดออกหรือไม่

อาการตับโตสามารถดูได้จากค่าของแอททริบิวต์ขนาดตับที่โตขึ้นในฐานข้อมูล กล่าวคือ ถ้ามีค่ามากกว่า 0 จากผลการทดลองของไจท์ที่ 1 พบว่ามีแอททริบิวต์ที่แสดงถึงอาการตับโตปรากฏอยู่ในต้นไม้ตัดสินใจโดยถ้าผู้ป่วยคนใดไม่มีอาการตับโตจะถูกจำแนกเป็นคลาส DF และถ้ามีอาการตับโตจะถูกจำแนกเป็นคลาส DHF หลังจากนั้นพิจารณาค่าสหสัมพันธ์พบว่ามีความสัมพันธ์กับการเกิดโรคในระดับต่ำด้วยค่า r เท่ากับ 0.4177 และจากการพิจารณาสัดส่วนข้อมูลผู้ป่วยคลาส DF จำนวน 488 คนและคลาส DHF จำนวน 513 คนที่มีอาการตับโตของไจท์ที่ 4 พิจารณาแยกตามคลาส DF, DHF I, DHF II และ DHF III ได้ผลการพิจารณาดังนี้

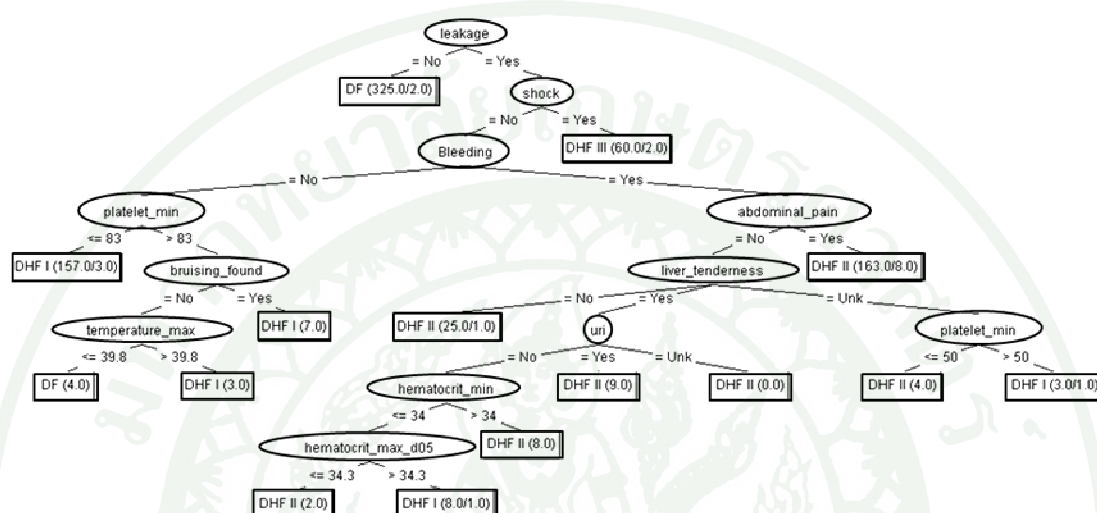
ตารางที่ 10 สัดส่วนการพิจารณาจำนวนตับโตในแต่ละระดับของโรคไข้เลือดออก

	DF	DHF I	DHF II	DHF III
จำนวนทั้งหมด(คน)	488	222	229	62
พบอาการ (คน)	333	180	206	59
พบอาการ (%)	68.24%	81.08%	89.96%	95.16%

จากผลการพิจารณาจากผลลัพธ์ทั้ง 3 การทดลอง พบว่า อาการตับโตมีอำนาจในการจำแนกประเภทของโรคไข้เลือดออก คือ สามารถจำแนก DF ออกจาก DHF ได้ และเมื่อพิจารณาตามระดับความรุนแรงยังพบอีกว่า อัตราส่วนของผู้ป่วยไข้แดงที่มีอาการตับโตคิดเป็น 68.24% ส่วนผู้ป่วยไข้เลือดออกความรุนแรงระดับ 1, 2 และ 3 มีสัดส่วนเท่ากับ 81.08, 89.96 และ 95.16% ตามลำดับ ซึ่งจากการสังเกตเห็นว่าเมื่อผู้ป่วยมีอาการตับโตเกิดขึ้น ระดับความรุนแรงของโรคไข้เลือดออกก็จะรุนแรงขึ้นเช่นเดียวกัน

แต่จากผลลัพธ์ต้นไม้ตัดสินใจไจท์ที่ 1 เมื่อพิจารณาจำนวนผู้ป่วยที่ถูกจำแนกด้วยค่าข้อมูลอาการตับโตพบว่ามีความน้อยมากโดยคิดเป็น 1.6% ของผู้ป่วยทั้งหมด ดังนั้นแพทย์ต้องการทราบเพิ่มเติมว่าถ้าใช้อาการตับโตเป็นปัจจัยหลักในการจำแนก สามารถทำได้หรือไม่ เนื่องจากอาการตับโตเป็นปัจจัยที่ไม่ซับซ้อนและง่ายต่อการตรวจพบ รวมถึงไม่เสียค่าใช้จ่ายเพิ่มเติมในการวินิจฉัย

ดังนั้นผู้วิจัยจึงทดลองเพิ่มในส่วนของการทดลองเพิ่มเติม जोทย์ที่ 4 โดยวิเคราะห์เฉพาะกลุ่มผู้ป่วยที่มีอาการตับโต เมื่อพิจารณาเฉพาะผู้ป่วยที่มีอาการตับโตจำนวน 778 คน ได้ผลการทดลองดังภาพที่ 11



ภาพที่ 11 ต้นไม้ตัดสินใจเฉพาะผู้ป่วยที่มีอาการตับโตใน जोทย์ที่ 4 (เพิ่มเติม)

คลาส DF

คลาส DF ส่วนใหญ่ถูกจำแนกด้วยแอททริบิวต์ leakage ถ้า leakage มีค่าเท่ากับ No สามารถแยก DF 325 คน ผิด 2 คน

คลาส DHF I

คลาส DHF I ส่วนใหญ่ถูกจำแนกด้วยแอททริบิวต์ leakage, shock, Bleeding, platelet_min ถ้า Bleeding มีค่าเท่ากับ No และ platelet_min น้อยกว่าหรือเท่ากับ 83,000 cells/ μ l (ไมโครลิตร) สามารถแยก DHF I ได้จำนวน 157 คน ผิด 3 คน

คลาส DHF II

คลาส DHF II ส่วนใหญ่ถูกจำแนกด้วยแอททริบิวต์ leakage, shock, Bleeding, abdominal_pain ถ้า Bleeding มีค่าเท่ากับ Yes และ abdominal_pain มีค่าเท่ากับ Yes สามารถแยก DHF I จำนวน 163 คน ผิด 8 คน

คลาส DHF III

คลาส DHF III ส่วนใหญ่ถูกจำแนกด้วยแอททริบิวต์ leakage, shock ถ้า leakage มีค่าเท่ากับ Yes และแอททริบิวต์ shock มีค่าเท่ากับ Yes สามารถแยก DHF III จำนวน 60 คน ผิด 2 คน

ส่วนแอททริบิวต์อื่นๆ ในต้นไม้ที่ไม่กล่าวถึงเพราะแอททริบิวต์นั้นสามารถจำแนกผู้ป่วยได้จำนวนน้อย

หลังจากการทดลองข้างต้นพบว่าในกลุ่มผู้ป่วยที่มีอาการตับโต แอททริบิวต์ที่มีอำนาจการจำแนกสูงคือ leakage และ shock ซึ่งเป็นแอททริบิวต์สำคัญที่เหมือนกับผลการทดลองโทยที่ 1

หลังจากการทดลองข้างต้น ผู้วิจัยทดลองตัดแอททริบิวต์ที่ต้องใช้เวลาและใช้ค่าใช้จ่ายในการวินิจฉัยออกจากการพิจารณา ได้แก่ แอททริบิวต์ leakage, shock และ Effusion_Result เหตุผลที่ต้องตัดแอททริบิวต์ดังกล่าวออกไป เนื่องจากแอททริบิวต์ leakage คือค่าของข้อมูลต้องรอเวลาให้แสดงอาการ shock คือค่าข้อมูลที่เกิดอาการช็อก ซึ่งถ้าเกิดแล้วแพทย์ก็ไม่มีเวลาจำเป็นต้องการวินิจฉัยโรคเพื่อวินิจฉัยล่วงหน้าแล้วนั่นเอง ส่วนแอททริบิวต์ Effusion_Result นั้นเป็นค่าที่ได้มาจากการฉายรังสีซึ่งมีค่าใช้จ่ายค่อนข้างสูง ด้วยเหตุผลทั้งหมดจึงได้ตัดแอททริบิวต์นั้นออกจากการพิจารณาในการทดลองเพิ่มของโทยที่ 4

จากผลการทดลองพบว่าต้นไม้ตัดสินใจของการทดลองที่ตัดแอททริบิวต์ leakage, shock และ Effusion_Result ออก พบว่าต้นไม้มีขนาดใหญ่ ประกอบด้วยกฎในการจำแนก 76 กฎและได้ค่าความถูกต้องเท่ากับ 80.4% อธิบายผลการทดลองแยกตามคลาสได้ดังนี้

คลาส DF มีจำนวนกฎทั้งหมด 29 กฎในการจำแนกคลาส กฎที่ดีที่สุดของการจำแนกมีดังนี้

- Bleeding = No และ platelet_min > 83 และ wbc_max < 12.6 จำแนก DF ได้จำนวน 142 คน ผิด 8 คน คิดเป็น 42.64% ของจำนวน DF ทั้งหมด

- Bleeding = Yes และ platelet_min > 59 และ abdominal_pain = No และ injected_found = No จำแนก DF ได้จำนวน 54 คน ผิด 6 คน คิดเป็น 16.22% ของจำนวน DF ทั้งหมด

- Bleeding = Yes และ platelet_min > 59 และ abdominal_pain = Yes และ

hematocrit_max <= 41 และ lymph_found = No และ rash_found = No จำแนก DF ได้ 21 คน ผิด 1 คน คิดเป็น 6.3% ของจำนวน DF ทั้งหมด

- และกฎอื่นๆอีก 26 กฎ คิดสัดส่วนข้อมูลพบว่ามีแต่ละกฎจำแนกผู้ป่วยจำนวนไม่เกิน 5% ของจำนวน DF ทั้งหมด

คลาส DF มีจำนวนกฎที่ใช้ในการจำแนกทั้งหมด 29 กฎ กฎที่ดีที่สุดของคลาส DF สามารถจำแนกผู้ป่วยได้ 42.64% ของผู้ป่วยคลาส DF ทั้งหมด ส่วนกฎข้ออื่นๆอีก 28 กฎ สามารถจำแนกผู้ป่วยได้ไม่ถึง 20% ของผู้ป่วยคลาส DF

คลาส DHF I มีจำนวนกฎทั้งหมด 16 กฎในการจำแนกคลาส กฎที่ดีที่สุดในการจำแนกคลาสมีดังนี้

- Bleeding = No และ platelet_min <= 83 และ pulse_pre_min = No และ hctMax <= 46 และ platelet_min <= 39 และ temperature_min < 36.6 จำแนก DHF I จำนวน 59 ผิด 4 คน คิดเป็น 32.77% ของผู้ป่วยคลาส DHF I ทั้งหมด

- Bleeding = No และ platelet_min <= 83 และ pulse_pre_min = No และ hctMax > 46 จำแนก DHF I จำนวน 34 คน ผิด 2 คน คิดเป็น 18.9% ของผู้ป่วยคลาส DHF I ทั้งหมด

- Bleeding = No และ platelet_min <= 83 และ pulse_pre_min = No และ hctMax <= 46 และ platelet_min > 39 และ hematocrit_min > 31 และ lymph_found = Yes และ hct_min <= 39 และ hct_avg_d05 <= 34.67 จำแนก DHF I จำนวน 14 คน ผิด 1 คน คิดเป็น 7.78% ของผู้ป่วยคลาส DHF I ทั้งหมด

- Bleeding = No และ platelet_min <= 83 และ pulse_pre_min = No และ hctMax <= 46 และ platelet_min > 39 และ hematocrit_min > 31 และ lymph_found = No และ hct_max > 37 และ je = Yes และ wbc_min > 3 จำแนกคลาส DHF I จำนวน 14 คน ผิด 1 คน คิดเป็น 7.78% ของผู้ป่วยคลาส DHF I ทั้งหมด

- Bleeding = No และ platelet_min $\leq$ 83 และ pulse_pre_min = No และ hctMax $\leq$ 46 และ platelet_min $>$ 39 และ hct_min $\leq$ 31 จำแนกคลาส DHF I จำนวน 10 คน คิดเป็น 5.56% ของผู้ป่วยคลาส DHF I ทั้งหมด

- และกลุ่มอื่นๆ อีก 11 กลุ่ม คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกลุ่มนั้นไม่เกิน 5% ของจำนวน DHF I ทั้งหมด

คลาส DHF I มีจำนวนกลุ่มที่ใช้ในการจำแนกทั้งหมด 16 กลุ่ม กลุ่มที่ดีที่สุดของคลาส DHF I สามารถจำแนกผู้ป่วยได้ 32.77% ของผู้ป่วยคลาส DHF I ทั้งหมด ส่วนกลุ่มข้ออื่นๆอีก 15 กลุ่ม สามารถจำแนกผู้ป่วยได้ไม่ถึง 20% ของผู้ป่วยคลาส DHF I

คลาส DHF II มีจำนวนกลุ่มทั้งหมด 22 กลุ่ม กลุ่มที่ดีที่สุดในการจำแนกคลาสนี้มีดังนี้

- Bleeding = No และ platelet_min $\leq$ 59 และ pulse_pre_min = No และ ascites_found = No และ sbp_minus_dbp $>$ 30.25 และ liver_tenderness = Yes และ abdominal_pain = Yes จำแนกคลาส DHF II จำนวน 64 คน ผิด 6 คน คิดเป็น 31.07% ของผู้ป่วยคลาส DHF II ทั้งหมด

- Bleeding = No และ platelet_min $\leq$ 59 และ pulse_pre_min = No และ ascites_found = Yes และ Globulin $>$ 2.6 และ lft_albumin_avg $>$ 2.57 จำแนกคลาส DHF II จำนวน 45 คน ผิด 1 คน คิดเป็น 21.85% ของผู้ป่วยคลาส DHF II ทั้งหมด

- และกลุ่มอื่นๆ อีก 20 กลุ่ม คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกลุ่มนั้นไม่เกิน 5% ของจำนวน DHF II ทั้งหมด

คลาส DHF II มีจำนวนกลุ่มในการจำแนกทั้งหมด 22 กลุ่ม กลุ่มที่ดีที่สุดของคลาส DHF II สามารถจำแนกผู้ป่วยได้ 31.07% ของผู้ป่วยคลาส DHF II กลุ่มต่อมาสามารถจำแนกได้ 21.85% ของผู้ป่วยคลาส DHF II ส่วนกลุ่มข้ออื่นๆอีก 20 กลุ่มสามารถจำแนกผู้ป่วยได้ต่ำกว่า 20% ของผู้ป่วยคลาส DHF II แต่เมื่อสังเกตที่รูปแบบของกลุ่มพบว่าในแต่ละกลุ่มต้องใช้เงื่อนไขของแอททริบิวต์จำนวนมาก ซึ่งถ้านำไปใช้งานจริงอาจเกิดปัญหาเงื่อนไขที่ใช้ในการจำแนกไม่ครบถ้วน ทำให้แพทย์ไม่สามารถนำกลุ่มดังกล่าวไปใช้จำแนกข้อมูลคลาส DHF II ได้จริง

คลาส DHF III มีจำนวนกฎทั้งหมด 9 กฎในการจำแนกคลาส กฎที่ดีที่สุดในการจำแนกคลาสมีดังนี้

- Bleeding = Yes และ platelet_min $\leq$ 59 และ pulse_pre_min = Yes และ ascites_found = No และ temperature_min $<$ 36.5 จำแนกคลาส DHF III จำนวน 26 คน ผิด 3 คน คิดเป็น 44.07% ของผู้ป่วยคลาส DHF III ทั้งหมด

- และกฎอื่นๆ อีก 8 กฎ คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกฎนั้นน้อยกว่า 10% ของจำนวน DHF III ทั้งหมด

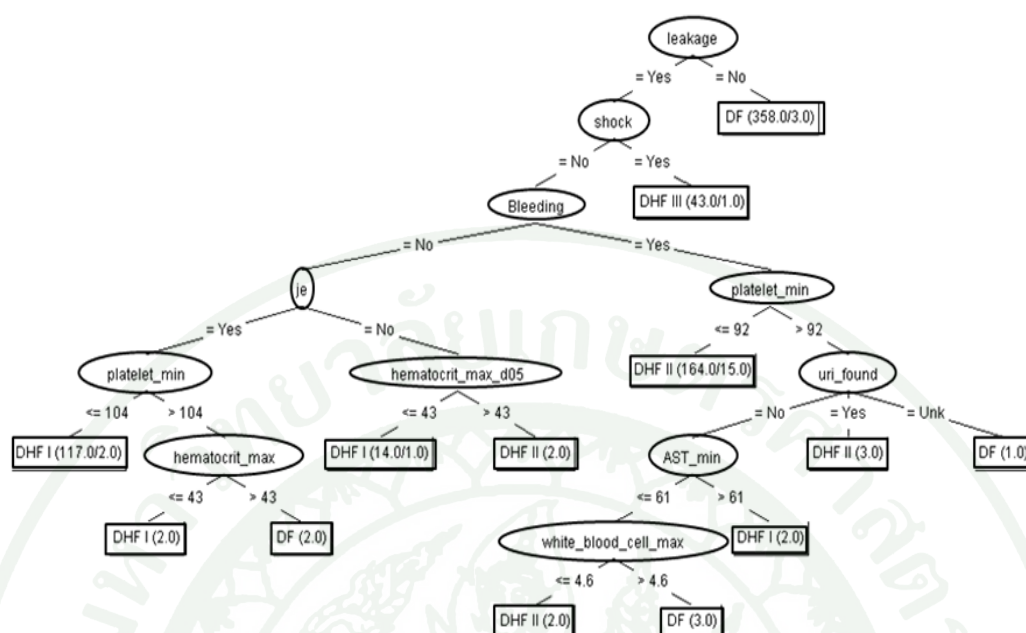
คลาส DHF III มีจำนวนกฎในการจำแนกทั้งหมด 9 กฎ กฎที่ดีที่สุดของคลาส DHF III สามารถจำแนกผู้ป่วยได้ 44.07% ของผู้ป่วยคลาส DHF III นอกจากนั้นอีก 8 กฎจำแนกผู้ป่วยได้น้อยกว่า 10% ของผู้ป่วยคลาส DHF III ทั้งหมด

จากผลการทดลองเพิ่มเติมโจทย์ที่ 4 พบว่า ถ้าตัดแอททริบิวต์ leakage, shock และ Effusion_Result ออกแล้ว กฎจากต้นไม้ตัดสินใจที่ได้สามารถในการจำแนกผู้ป่วยได้สูงที่สุดให้ความถูกต้องไม่ถึง 50% ของผู้ป่วยคลาสในคลาสที่จำแนกได้ และเมื่อพิจารณารูปแบบของกฎพบว่ากฎแต่ละข้อประกอบด้วยเงื่อนไขแอททริบิวต์ที่ใช้ในการจำแนกคลาสจำนวนมาก กล่าวคือถ้าต้องการจำแนกคลาสใดๆ จำเป็นต้องมีข้อมูลตามเงื่อนไขของกฎให้ครบถ้วน นอกจากนี้การที่กฎมีเงื่อนไขจำนวนมากทำให้โอกาสที่เงื่อนไขจะเป็นจริงเป็นไปได้ยากส่งผลให้กฎดังกล่าวไม่เกิดประโยชน์ใดในการจำแนกคลาส

5. ผลการทดลองโจทย์ที่ 5 วัคซีนโรคไข้สมองอักเสบมีส่วนลดการเกิดโรคไข้เลือดออกหรือไม่

การทดลองโจทย์ที่ 5 เพื่อหาบทสรุปที่ว่าบุคคลใดที่ได้รับวัคซีน JE แล้วยังสามารถเป็นโรคไข้เลือดออกได้อยู่หรือไม่ ทางผู้วิจัยได้เลือกขั้นตอนวิธีต้นไม้ตัดสินใจ ผลการทดลองดังภาพที่ 12

จากผลการทดลองต้นไม้ตัดสินใจค่าความถูกต้อง 96.1% พบว่ามีแอททริบิวต์ je ปรากฏอยู่ในต้นไม้ ถ้าพิจารณาจากจำนวนในกิ่งการตัดสินใจของแอททริบิวต์ je พบว่าถ้าแอททริบิวต์นี้มีค่าเป็น No ข้อมูลที่อยู่ในกิ่งประกอบด้วยข้อมูลผู้ป่วยคลาส DHF I และ DHF II และถ้าค่าแอททริบิวต์นี้มีค่าเป็น Yes ข้อมูลที่อยู่ในกิ่งนี้ประกอบด้วยข้อมูลคลาส DF และ DHF I



ภาพที่ 12 ต้นไม้ตัดสินใจของ 5 คัดผู้ป่วยที่มีค่าแอทริบิวต์ je เป็น Unknown ออก

เมื่อคำนวณค่าสหสัมพันธ์ของแอทริบิวต์ je กับระดับความรุนแรงของโรคไข้เลือดออก พบว่าการได้รับวัคซีน JE ไม่มีความสัมพันธ์กับการเกิดคลาส

ต่อมาผลลัพธ์ของวิธีการของกฎความสัมพันธ์ผลการทดลอง (รายละเอียดภาคผนวก ค) อธิบายแยกตามคลาสดังนี้

คลาส DF หรือผู้ป่วยไข้แดงก็ เมื่อพิจารณากฎความสัมพันธ์จำนวน 10, 100 และ 1,000 กฎในการทดลองที่มีค่า minimum support เท่ากับ 0.1 และ 0.05 ไม่พบแอทริบิวต์ je ปรากฏทางด้านซ้ายของกฎ ดังนั้นสามารถสรุปได้ว่าแอทริบิวต์ je ไม่มีความสัมพันธ์กับคลาส DF หรือกล่าวคือการได้รับวัคซีน JE ไม่ได้ส่งผลต่อการเกิดโรคของผู้ป่วยไข้แดงก็

คลาส DHF หรือผู้ป่วยไข้เลือดออก เมื่อพิจารณากฎความสัมพันธ์จำนวน 10, 100, 1000 กฎในการทดลองที่มีค่า minimum support เท่ากับ 0.1 และ 0.5 พบว่าเมื่อกำหนดจำนวนกฎเท่ากับ 10 และ 100 กฎ ไม่พบแอทริบิวต์ je ปรากฏทางด้านซ้ายของกฎ แต่เมื่อกำหนดจำนวนกฎเท่ากับ 1,000 กฎ พบแอทริบิวต์ je ปรากฏทางด้านซ้ายของกฎในกฎข้อที่ 48 เป็นต้นไป ด้วยค่าความเชื่อมั่นเท่ากับ 1 หรือ 100% และจากกฎข้อที่ 49 “je=Yes 308 ==> dx=DHF 308” จากกฎนี้

สามารถตีความหมายได้ว่าแอททริบิวต์ *je* ที่มีค่าเป็น Yes มีความสัมพันธ์กับคลาส DHF จำนวน 308 กล่าวคือ ผู้ป่วยโรคไข้เลือดออกจำนวน 308 คน เป็นผู้ป่วยที่ได้รับวัคซีน JE แล้ว คิดเป็น 88.25% ของคลาสผู้ป่วยไข้เลือดออกทั้งหมด ดังนั้นสามารถสรุปได้ว่าคนที่ได้รับวัคซีน JE ยังสามารถป่วยเป็นโรคไข้เลือดออกได้

ดังนั้นจากผลการทดลองทั้งสามวิธีคือ ขั้นตอนวิธีต้นไม้ตัดสินใจ การคำนวณค่าสหสัมพันธ์ และขั้นตอนวิธีการหาความสัมพันธ์ด้วยวิธี *Apriori* พบว่า การได้รับวัคซีนโรคไข้สมองอักเสบไม่ได้ป้องกันการเกิดโรคไข้เลือดออก บุคคลที่ได้รับวัคซีนแล้วอาจมีโอกาสป่วยเป็นโรคไข้เลือดออกได้ แต่ถ้าบุคคลใดที่ได้รับวัคซีนแล้วป่วยเป็นโรคไข้เลือดออก บุคคลนั้นจะมีระดับของการติดเชื้อรุนแรงน้อยกว่าบุคคลที่ไม่ได้รับวัคซีน

นอกจากนี้แพทย์ต้องการทราบว่า ถ้าวิเคราะห์โดยยึดข้อมูลการได้รับวัคซีนไข้สมองอักเสบหรือวัคซีน JE เป็นหลัก จะสามารถจำแนกคลาสระดับความรุนแรงของโรคไข้เลือดออกได้หรือไม่ ดังนั้นผู้วิจัยทดลองเพิ่มเติมของโจทย์ที่ 5 โดยวิเคราะห์เฉพาะกลุ่มผู้ป่วยโรคไข้เลือดออกที่ได้รับวัคซีน JE และตัดแอททริบิวต์ *leakage*, *shock* และ *Effusion_Result* ออกด้วยเหตุผลเดียวกับการทดลองเพิ่มเติมโจทย์ที่ 4

ผลการทดลองได้ต้นไม้ตัดสินใจขนาดใหญ่ประกอบด้วยกฎทั้งหมด 63 กฎ มีความถูกต้องในการจำแนกคลาสเท่ากับ 79.5% สามารถอธิบายตามคลาสได้ดังนี้

คลาส DF มีจำนวนกฎทั้งหมด 22 กฎ กฎที่ดีที่สุดดังนี้

- ถ้า *platelet_min* > 67 และ *Bleeding* = No และ *AST_min* ≤ 101 และ *platelet_min* > 103 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DF จำนวน 99 คน คิดเป็น 30.75% ของผู้ป่วยคลาส DF

- ถ้า *platelet_min* > 67 และ *Bleeding* = Yes และ *abdominal_pain* = No แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DF จำนวน 79 คน ผิด 9 คน คิดเป็น 24.53% ของผู้ป่วยคลาส DF

- ถ้า *platelet_min* > 67 และ *Bleeding* = No และ *ascites_found* = No และ *abdominal_pain* = No และ *liver_tenderness* = Yes และ *protien_avg* > 5.98 และ *sbp_minus_dbp*

< 40.71 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DF จำนวน 23 คน ผิด 7 คน คิดเป็น 7.14% ของผู้ป่วยคลาส DF

- และกฎอื่นๆอีก 19 กฎ คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกฎนั้นน้อยกว่า 7% ของจำนวน DF ทั้งหมด

คลาส DF มีจำนวนกฎที่ใช้ในการจำแนกทั้งหมด 22 กฎ กฎที่ดีที่สุดของคลาส DF สามารถจำแนกผู้ป่วยได้ 30.75% และ 24.53% ของผู้ป่วยคลาส DF ทั้งหมด ส่วนกฎข้ออื่นๆอีก 20 กฎ สามารถจำแนกผู้ป่วยได้ไม่ถึง 10% ของผู้ป่วยคลาส DF

คลาส DHF I มีจำนวนกฎทั้งหมด 19 กฎ กฎที่ดีที่สุดดังนี้

- ถ้า platelet_min $\leq$ 67 และ Bleeding = No และ ascites_found = No และ abdominal_pain = Yes และ platelet_max $>$ 41 และ ALT_max $>$ 40 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF I จำนวน 49 คน ผิด 4 คน คิดเป็น 37.7% ของผู้ป่วยคลาส DHF I
- ถ้า platelet_min $\leq$ 67 และ Bleeding = No และ ascites_found = Yes แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF I จำนวน 22 คน ผิด 2 คน คิดเป็น 16.92% ของผู้ป่วยคลาส DHF I
- และกฎอื่นๆอีก 17 กฎ คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกฎน้อยกว่า 7% ของจำนวน DHF I ทั้งหมด

คลาส DHF I มีจำนวนกฎที่ใช้ในการจำแนกทั้งหมด 19 กฎ กฎที่ดีที่สุดของคลาส DHF I สามารถจำแนกผู้ป่วยได้ 37.7% และ 16.92% ของผู้ป่วยคลาส DHF I ทั้งหมด ส่วนกฎข้ออื่นๆอีก 17 กฎ สามารถจำแนกผู้ป่วยได้ไม่ถึง 7% ของผู้ป่วยคลาส DHF I

คลาส DHF II มีจำนวนกฎทั้งหมด 16 กฎ กฎที่ดีที่สุดดังนี้

- ถ้า platelet_min $\leq$ 67 และ Bleeding = Yes และ albumin $>$ 2.65 และ ascites_found = No และ platelet_min $\leq$ 28 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF II จำนวน 36 คน ผิด 1 คน คิดเป็น 26.08% ของผู้ป่วยคลาส DHF II

- ถ้า platelet_min ≤ 67 และ Bleeding = Yes และ albumin > 2.65 และ ascites_found = No และ platelet_min > 28 และ platelet_avg > 45.67 และ AST_max > 72 และ abdominal_pain = Yes และ platelet_min ≤ 57 และ ALT_max ≤ 231 และ temperature_min ≤ 36.5 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF II จำนวน 31 คน ผิด 1 คน คิดเป็น 22.46% ของผู้ป่วยคลาส DHF II

- ถ้า platelet_min ≤ 67 และ Bleeding = Yes และ albumin > 2.65 และ ascites_found = Yes และ Globulin > 2.55 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF II จำนวน 23 คน ผิด 1 คน คิดเป็น 16.67% ของผู้ป่วยคลาส DHF II

- และกฎอื่นๆอีก 13 กฎ คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกฎนั้นน้อยกว่า 7% ของจำนวน DHF II

คลาส DHF II มีจำนวนกฎที่ใช้ในการจำแนกทั้งหมด 16 กฎ กฎที่ดีที่สุดของคลาส DHF II สามารถจำแนกผู้ป่วยได้ร้อยละ 26.08, 22.46 และ 16.67 ของผู้ป่วยคลาส DHF II ส่วนกฎอื่นๆอีก 13 กฎสามารถจำแนกผู้ป่วยได้ต่ำกว่า 7% แต่เมื่อสังเกตที่รูปแบบของกฎพบว่าในแต่ละกฎต้องใช้เงื่อนไขของแอททริบิวต์จำนวนมาก (ยกตัวอย่างเช่นกฎที่ดีที่สุดลำดับที่ 2 ของคลาส DHF II) ซึ่งถ้านำไปใช้งานจริงอาจเกิดปัญหาเงื่อนไขที่ใช้ในการจำแนกไม่ครบถ้วน ทำให้แพทย์ไม่สามารถนำกฎดังกล่าวไปใช้จำแนกข้อมูลคลาส DHF II ได้จริง

คลาส DHF III มีจำนวนกฎทั้งหมด 6 กฎ กฎที่ดีที่สุดดังนี้

- ถ้า platelet_min ≤ 67 และ Bleeding = Yes และ albumin ≤ 2.65 และ quantity_max_found = Ra1 แล้ว ผู้ป่วยถูกจำแนกเป็นคลาส DHF III จำนวน 9 คน คิดเป็น 27.27% ของผู้ป่วยคลาส DHF III

- และกฎอื่นๆอีก 5 กฎ คิดสัดส่วนข้อมูลพบว่ามีข้อมูลอยู่ในแต่ละกฎนั้นน้อยกว่า 5 คน

คลาส DHF III มีจำนวนกฎที่ใช้ในการจำแนกทั้งหมด 6 กฎ กฎที่ดีที่สุดของคลาส DHF III สามารถจำแนกผู้ป่วยได้ 27.27% นอกนั้นอีก 5 กฎจำแนกผู้ป่วยได้จำนวนน้อยกว่า 5 คนในผู้ป่วยคลาส DHF III

จากการทดลองเพิ่มเติมโจทย์ที่ 5 พบว่า กฎที่สามารถในการจำแนกคลาสผู้ป่วยกลุ่มที่ได้รับวัคซีน JE ได้สูงที่สุดมีความถูกต้องต่ำกว่า 40% ของคลาสที่จำแนกได้จากกฎนั้น และเมื่อพิจารณารูปแบบของกฎพบว่ากฎแต่ละข้อประกอบด้วยเงื่อนไขแอททริบิวต์ที่ใช้ในการจำแนกคลาสจำนวนมาก กล่าวคือถ้าต้องการจำแนกคลาส DF หรือคลาสอื่น จำเป็นต้องมีข้อมูลตามเงื่อนไขของกฎให้ครบถ้วน นอกจากนี้การที่กฎมีเงื่อนไขจำนวนมากทำให้โอกาสที่เงื่อนไขจะเป็นจริงเป็นไปได้ยากทำให้กฎดังกล่าวไม่เกิดประโยชน์ใดในการจำแนกคลาส

6. ผลการทดลองโจทย์ที่ 6 ผู้ป่วยโรคไข้เลือดออกมีค่าเอนไซม์ AST มากกว่า ALT เสมอหรือไม่

ผลการหาความสัมพันธ์โจทย์ที่ 6 (รายละเอียดภาคผนวก ก) อธิบายผลตามคลาสได้ดังนี้

คลาส DF หรือผู้ป่วยไข้เดงกี เมื่อพิจารณาความสัมพันธ์จำนวน 10 และ 100 กฎไม่พบแอททริบิวต์ AST_ALT ปรากฏทางด้านซ้ายของกฎ

คลาส DHF หรือผู้ป่วยไข้เลือดออก เมื่อพิจารณาผลลัพธ์กฎความสัมพันธ์จำนวน 10 กฎพบว่ามีแอททริบิวต์ AST_ALT ปรากฏอยู่ทางด้านซ้ายของกฎข้อที่ 8 และข้อที่ 10 ส่วนผลลัพธ์ 100 กฎพบว่ามีแอททริบิวต์ AST_ALT ปรากฏอยู่ทางด้านซ้ายของกฎจำนวน 50 กฎ ซึ่งมีค่าความน่าเชื่อมั่นเท่ากับ 1 จำนวน 30 กฎ และค่าความเชื่อมั่นเท่ากับ 0.99 จำนวน 20 กฎ

ผลการพิจารณาสัดส่วนจำนวนจากข้อมูลจริงในฐานข้อมูลได้รายละเอียดดังตารางที่ 11

ตารางที่ 11 จำนวนผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่าเอนไซม์ ALT

	DF	DHF I	DHF II	DHF III	DHF
จำนวนผู้ป่วยทั้งหมด	487	222	229	62	513
พบ AST > ALT (คน)	411	213	226	62	501
พบ AST > ALT (%)	84.39	95.95	98.69	100	97.66

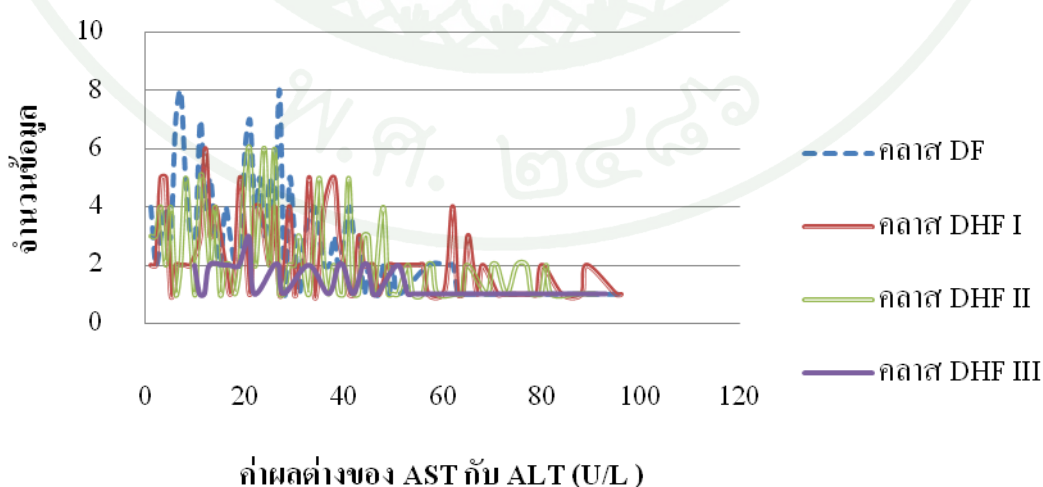
สรุปผลการพิจารณาร่วมกันระหว่างกฎความสัมพันธ์และสัดส่วนของข้อมูลผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่าเอนไซม์ ALT สามารถสรุปผลแยกตามคลาส DF และ DHF ได้ดังนี้

คลาส DF สัดส่วนจำนวนผู้ป่วยที่มีค่าของเอนไซม์ AST มากกว่า ALT คิดเป็น 84.39% และจากกฎความสัมพันธ์ที่ไม่ปรากฏเอทริบิวท์ AST_ALT ทางด้านซ้ายของกฎทำให้สรุปได้ว่า ในผู้ป่วยโรคไข้แดงก็ส่วนใหญ่มีค่าเอนไซม์ AST สูงกว่า ALT แต่ค่าของเอนไซม์ AST มากกว่า ALT นั้นไม่ได้สัมพันธ์กับการเกิดคลาส DF

คลาส DHF สัดส่วนจำนวนผู้ป่วยที่มีค่าของเอนไซม์ AST มากกว่า ALT คิดเป็น 97.66% และจากกฎความสัมพันธ์ที่ปรากฏเอทริบิวท์ AST_ALT ทางด้านซ้ายจำนวน 50 กฎ สรุปได้ว่า ในผู้ป่วยโรคไข้เลือดออกส่วนใหญ่มีค่าเอนไซม์ AST สูงกว่า ALT และยังระดับความรุนแรงของโรคไข้เลือดออกรุนแรงมากขึ้น อัตราส่วนจำนวนผู้ป่วยที่มีค่าเอนไซม์ AST สูงกว่า ALT ก็มีสูงขึ้นด้วยเช่นกัน

นอกจากนี้แพทย์ต้องการทราบว่า ค่าผลต่างของค่า AST มากกว่า ALT สามารถหาได้หรือไม่ ดังนั้นผู้วิจัยได้ออกแบบการทดลองเพิ่มเติมให้วิเคราะห์ค่าผลต่างของ AST และ ALT ในแต่ละคลาสและสังเกตจากการกระจายตัวของข้อมูลว่ามีรูปแบบที่แตกต่างกันหรือไม่

การทดลองเพิ่มเติม โจทย์ที่ 6 ผู้วิจัยเลือกข้อมูลคลาส DF, DHF I และ DHF II คลาสละ 150 คน และคลาส DHF III ทั้งหมด 53 คน นำมาสร้างกราฟการกระจายตัวได้ผลดังนี้



ภาพที่ 13 กราฟการกระจายตัวของผลต่างของ AST มากกว่า ALT

จากกราฟแสดงการกระจายของข้อมูลผลต่างค่า AST และ ALT ในแต่ละคลาสจากกลุ่มตัวอย่างพบว่า ค่าผลต่างของ AST และ ALT ที่มีการกระจายตัวตั้งแต่ช่วง 0 ถึง 90 พบในความถี่ที่ใกล้เคียงกัน แสดงให้เห็นว่าไม่มีรูปแบบที่แน่นอนสำหรับคลาสใดๆ ดังนั้นจึงไม่สามารถหาค่าผลต่างที่จะนำมาเป็นค่า threshold หรือค่าที่จะเป็นเกณฑ์ในการจำแนกประเภทของโรคไข้เลือดออก จากผลการกระจายตัวในข้อมูลชุดนี้ได้

7. ผลการทดลองโถยที่ 7 ผู้ป่วยโรคไข้เลือดออกมีความผิดปกติของเม็ดเลือดขาวทุกรายหรือไม่

ผลทดลองกฎความสัมพันธ์โถยที่ 6 สามารถอธิบายแยกตามคลาสได้ดังนี้

คลาส DF เมื่อพิจารณากฎความสัมพันธ์จำนวน 10 และ 100 กฎ ไม่พบแอททริบิวต์ atypical lymphocyte ปรากฏทางด้านซ้ายของกฎ

คลาส DHF เมื่อพิจารณาผลลัพธ์กฎความสัมพันธ์จำนวน 10 กฎ พบว่ามีแอททริบิวต์ atypical lymphocyte ปรากฏอยู่ทางด้านซ้ายของกฎข้อที่ 6 และข้อที่ 9 ส่วนผลลัพธ์ 100 กฎพบว่ามีแอททริบิวต์ atypical lymphocyte ปรากฏอยู่ทางด้านซ้ายของกฎจำนวน 39 กฎ

ผลการพิจารณาสัดส่วนจำนวนผู้ป่วยโรคไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาวรายละเอียดดังตารางที่ 12

ตารางที่ 12 สรุปรูปจำนวนผู้ป่วยไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาว

	DF	DHF
จำนวนผู้ป่วยทั้งหมด	487	513
พบอาการ (คน)	454	503
พบอาการ (%)	93.22	98.05

เมื่อนำผลการพิจารณาทั้งสองผลการทดลองระหว่างกฎความสัมพันธ์และสัดส่วนของข้อมูลผู้ป่วยไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาว สามารถสรุปผลแยกตามคลาส DF และ DHF ได้ดังนี้

คลาส DF สัดส่วนจำนวนผู้ป่วยไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาว คิดเป็น 93.22% และจากกฎความสัมพันธ์ไม่ปรากฏเอทริบิวท์ ความผิดปกติของเม็ดเลือดขาว ทางด้านซ้ายของกฎสรุปได้ว่าความผิดปกติของเม็ดเลือดขาวไม่มีความสัมพันธ์กับการเกิดคลาส DF เป็นเพียงอาการที่แสดงออกในผู้ป่วยคลาส DF เท่านั้น

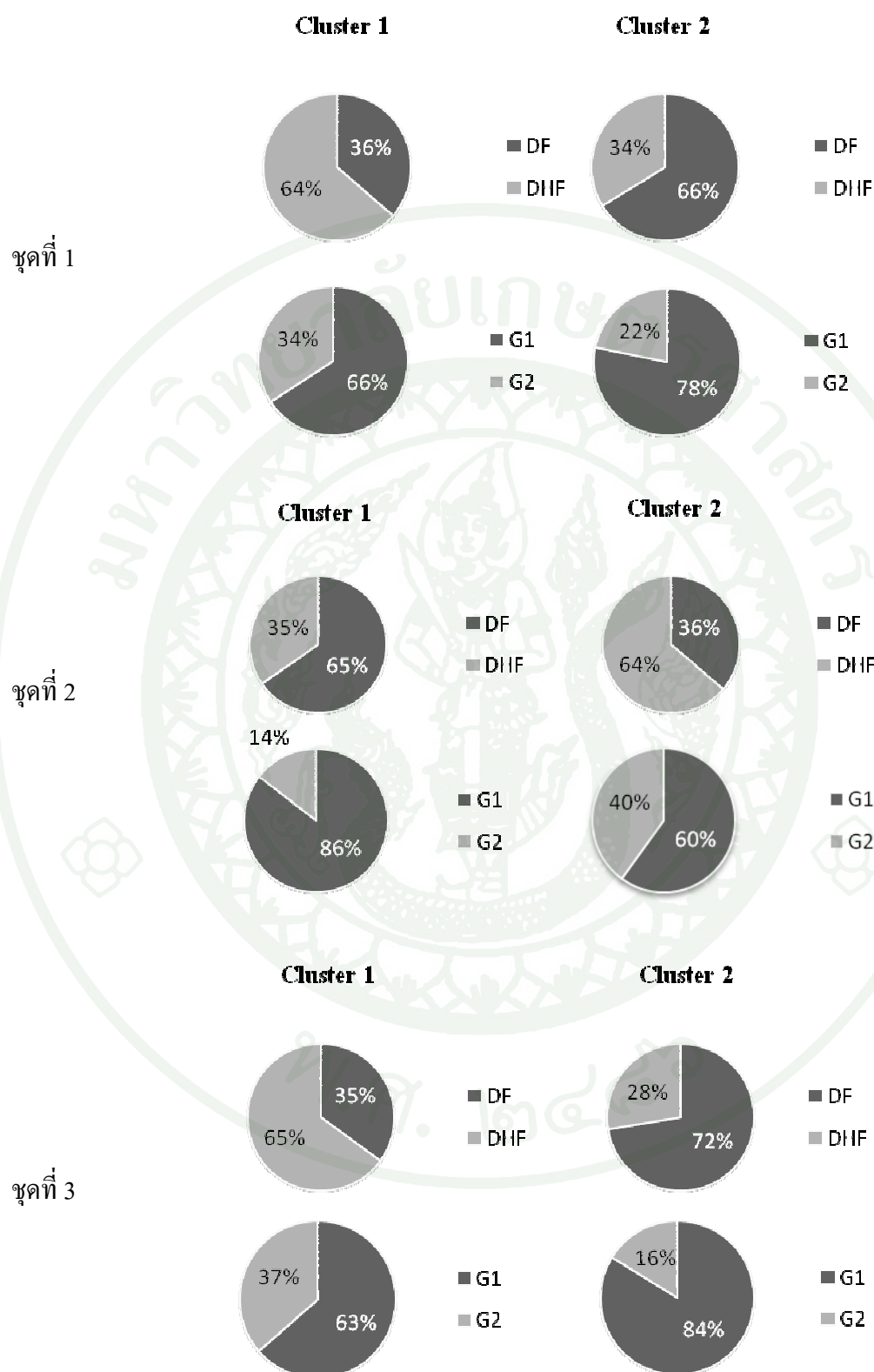
คลาส DHF สัดส่วนจำนวนผู้ป่วยไข้เลือดออกที่มีความผิดปกติของเม็ดเลือดขาว คิดเป็น 98.05% และจากกฎความสัมพันธ์ปรากฏเอทริบิวท์ atypical lymphocyte ทางด้านซ้ายจำนวน 39 กฎ จึงบอกได้ว่าความผิดปกติของเม็ดเลือดขาวมีความสัมพันธ์กับคลาส DHF คือ ถ้าพบความผิดปกติของเม็ดเลือดขาวในผู้ป่วยไข้แดงก็คนใด โอกาสที่ผู้ป่วยคนนั้นจะป่วยเป็นโรคไข้เลือดมีสูง

สรุปผลจากโจทย์ที่ 7 พบว่าผู้ป่วยโรคไข้แดงก็และผู้ป่วยโรคไข้เลือดออกส่วนใหญ่มีความผิดปกติของเม็ดเลือดขาว ดังนั้นบอกได้ว่าผู้ป่วยที่ติดเชื้อไข้แดงก็ (คลาส DF และคลาส DHF) มีโอกาสพบอาการความผิดปกติของเม็ดเลือดขาว แต่ไม่จำเป็นต้องมีความผิดปกตินี้ในผู้ป่วยโรคไข้เลือดออกทุกคน

8. ผลการทดลองโจทย์ที่ 8 ค้นหาปัจจัยเพิ่มเติมที่สามารถช่วยแยกผู้ป่วยคลาส DHF ระดับ 1 (DHF I) ออกจาก DHF ระดับ 2 (DHF II) ได้หรือไม่

จากผลการทดลองการแบ่งกลุ่มโดยดูจากธรรมชาติของข้อมูลพบว่า ข้อมูลผู้ป่วยโรคไข้เลือดออกในการทดลองสามารถแบ่งพิจารณาเป็น 2 กลุ่มคือ คลัสเตอร์ที่ 1 (Cluster1) และ คลัสเตอร์ที่ 2 (Cluster2) วัดประสิทธิภาพโดยการพิจารณาเทียบการแบ่งคลาส 2 แบบคือ แบ่งคลาสออกเป็น DF และ DHF กับอีกแบบคือ แบ่งคลาสออกเป็น G1 (Group1) และ G2 (Group2) โดยใน Group1 ประกอบด้วยคลาส DF และคลาส DHF I และใน Group2 ประกอบด้วยคลาส DHF II และคลาส DHF III

ผลการทดลองการจัดกลุ่มเพื่อดูผลการจัดกลุ่มด้วยวิธีการเคมีน ทางผู้วิจัยได้ทำการวัดค่าประสิทธิภาพในการแบ่งกลุ่มของข้อมูลโรคไข้เลือดออกด้วยค่า Entropy และค่า Purity ตามสมการ (13) และ (15) ได้ผลดังนี้



ภาพที่ 14 แผนภูมิแสดงอัตราส่วนของข้อมูลแต่ละคลัสเตอร์

ตารางที่ 13 ประสิทธิภาพของการแบ่งกลุ่มด้วยขั้นตอนวิธีเคมีน

ชุดแอททริบิวต์	คลาสที่พิจารณา	Entropy	Purity
ชุดที่ 1	DF / DHF	0.90	0.68
	G1/G2	0.83	0.71
ชุดที่ 2	DF / DHF	0.94	0.64
	G1/G2	0.81	0.71
ชุดที่ 3	DF / DHF	0.9	0.68
	G1/G2	0.83	0.71

จากผลการทดลองของแอททริบิวต์ชุดที่ 1 เมื่อพิจารณาเปรียบเทียบคลาสได้ผลดังนี้

เปรียบเทียบคลาส DF และ DHF พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส DF จำนวน 214 คน คิดเป็น 36% และคลาส DHF จำนวน 373 คนคิดเป็น 64% ซึ่งเห็นว่าคลาส DHF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส DF จำนวน 274 คน คิดเป็น 66% และคลาส DHF จำนวน 140 คนคิดเป็น 34% ซึ่งเห็นว่าคลาส DF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

เปรียบเทียบคลาส G1 และ G2 พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส G1 จำนวน 386 คน คิดเป็น 66% และคลาส G2 จำนวน 201 คนคิดเป็น 34% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส G1 จำนวน 324 คน คิดเป็น 78% และคลาส G2 จำนวน 90 คนคิดเป็น 22% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

จากผลการทดลองของแอททริบิวต์ชุดที่ 2 เมื่อพิจารณาเปรียบเทียบคลาสได้ผลดังนี้

เปรียบเทียบคลาส DF และ DHF พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส DF จำนวน 280 คน คิดเป็น 65% และคลาส DHF จำนวน 149 คนคิดเป็น 35% ซึ่งเห็นว่าคลาส DF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส DF จำนวน 208 คน คิดเป็น 36% และคลาส DHF จำนวน 364 คนคิดเป็น 64% ซึ่งเห็นว่าคลาส DHF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

เปรียบเทียบคลาส G1 และ G2 พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส G1 จำนวน 367 คน คิดเป็น 86% และคลาส G2 จำนวน 62 คนคิดเป็น 14% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส G1 จำนวน 343 คน คิดเป็น 60% และคลาส G2 จำนวน 229 คนคิดเป็น 40% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

จากผลการทดลองของแอททริบิวต์ชุดที่ 3 เมื่อพิจารณาเปรียบเทียบคลาสได้ผลดังนี้

เปรียบเทียบคลาส DF และ DHF พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส DF จำนวน 223 คน คิดเป็น 35% และคลาส DHF จำนวน 412 คนคิดเป็น 65% ซึ่งเห็นว่าคลาส DHF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส DF จำนวน 265 คน คิดเป็น 72% และคลาส DHF จำนวน 101 คนคิดเป็น 28% ซึ่งเห็นว่าคลาส DF เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

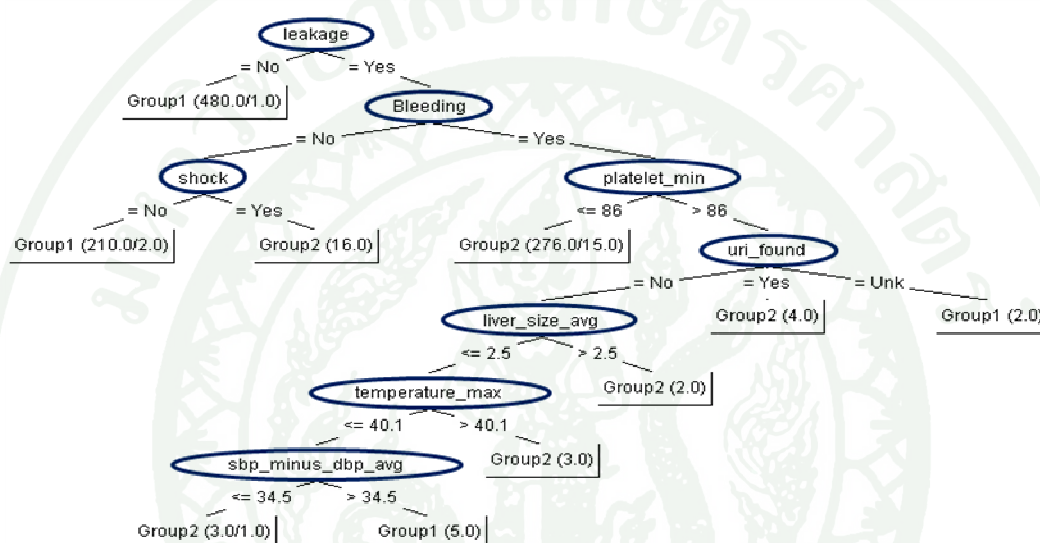
เปรียบเทียบคลาส G1 และ G2 พบว่า ในคลัสเตอร์ที่ 1 ประกอบด้วยคลาส G1 จำนวน 403 คน คิดเป็น 63% และคลาส G2 จำนวน 232 คนคิดเป็น 37% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ในคลัสเตอร์ที่ 1 สำหรับคลัสเตอร์ที่ 2 ประกอบด้วยคลาส G1 จำนวน 307 คน คิดเป็น 84% และคลาส G2 จำนวน 59 คนคิดเป็น 16% ซึ่งเห็นว่าคลาส G1 เป็นคลาสใหญ่ในคลัสเตอร์ที่ 2

สรุปผลจากการทดลองการแบ่งกลุ่มด้วยวิธีเคมีน พบว่าเมื่อพิจารณาเปรียบเทียบคลาสระหว่าง DF/DHF กับ G1/G2 พบว่าการเปรียบเทียบคลาสเป็น G1/G2 ให้ค่าเพียรดีสูงกว่า และค่าความปะปนของข้อมูลน้อยกว่า แต่ถ้าพิจารณาจากสัดส่วนของคลาสมากที่สุดในแต่ละคลัสเตอร์พบว่าคลาสที่เป็นคลาสใหญ่ของทั้งสองคลัสเตอร์เป็นคลาสเดียวกัน ดังนั้นการแบ่งกลุ่มไม่สามารถบอกได้ว่า DHF I รวมกับ DF แล้วได้ความสำคัญทางการแบ่งข้อมูล

จากการทดลองการแบ่งกลุ่มที่ไม่ประสบผลสำเร็จ ผู้วิจัยจึงเลือกใช้วิธีการต้นไม้ตัดสินใจอีกครั้ง โดยในการทดลองครั้งที่ 8 ได้แบ่งคลาสใหม่ โดยแบ่งคลาสออกเป็น Group1 (คลาส DF และ DHF I) และ Group2 (คลาส DHF II และ DHF III) ผลการทดลองแสดงในตารางที่ 14

ตารางที่ 14 ค่าความถูกต้องของต้นไม้ตัดสินใจโทยที่ 8 (%)

Class	Sensitivity	Specificity	Accuracy	Overall Accuracy
Group1	96.90	96.91	96.90	96.40
Group2	96.91	96.90	96.90	



ภาพที่ 15 ต้นไม้ตัดสินใจ คลาส Group1 และ Group2

ผลการทดลองแสดงให้เห็นแอททริบิวต์ที่สำคัญในการจำแนกคลาส DHF I ในคลาส Group1 ออกจาก DHF II ในคลาส Group 2 โดยมีแอททริบิวต์ที่น่าสนใจได้แก่ อาการรั่วของพลาสมาในเลือด (leakage) อาการเลือดออกของผู้ป่วย (Bleeding) อาการช็อก (shock) จำนวนเกล็ดเลือดน้อยที่สุดที่พบ (platelet_min) อาการติดเชื้อระบบทางเดินหายใจ (uri_found) ขนาดตับโต (liver_size_avg) อุณหภูมิมากที่สุดที่วัดได้ (temperature_max) และค่าความต่างระหว่างค่าสูง-ต่ำของความดัน (sbp_minus_dbp_avg) หลังจากนั้นผู้วิจัยคำนวณค่า correlation หรือค่าสหสัมพันธ์ของแอททริบิวต์กับคลาสของข้อมูลที่ปรากฏในต้นไม้ พบว่าทุกตัวสัมพันธ์กับคลาส ยกเว้นแอททริบิวต์ uri_found ที่ไม่มีความสัมพันธ์กับคลาสในชุดข้อมูล

วิจารณ์

การทดลองโทยที่ 1 ใช้ขั้นตอนวิธีต้นไม้ตัดสินใจเพื่อหาโครงสร้างต้นไม้ที่สามารถบ่งบอกถึงปัจจัยที่มีอำนาจและความสำคัญในการจำแนกประเภทของผู้ป่วยไข้เลือดออก ถ้าปัจจัยอยู่ส่วนต้นของต้นไม้ ปัจจัยนั้นมีความสำคัญมากกว่าปัจจัยที่อยู่ด้านล่างของต้นไม้ตัดสินใจ ดังนั้นผู้ที่ทดลองจึงได้หาค่าสหสัมพันธ์ซึ่งบ่งบอกถึงความสัมพันธ์ของปัจจัยของต้นไม้กับคลาสของผู้ป่วย ได้ผลดังนี้

ตารางที่ 15 ค่าสหสัมพันธ์ระหว่างปัจจัยและคลาสที่ปรากฏในผลการทดลองโทยที่ 1

ชื่อปัจจัย / ชื่อแอททริบิวต์	ค่าสหสัมพันธ์	ชนิดและระดับความสัมพันธ์
leakage	0.8456	เชิงบวก สูง
shock	0.5377	เชิงบวก ปานกลาง
Bleeding	0.3719	เชิงบวก ต่ำ
platelet_min	-0.5853	เชิงลบ ปานกลาง
ALT_avg	0.1941	เชิงบวก ต่ำมาก
wbc_avg	0.0842	เชิงบวก ต่ำมาก
lymp_found	-0.0905	เชิงลบ ต่ำมาก
wbc_min	-0.0785	เชิงลบ ต่ำมาก

จากตารางที่ 15 พบว่าปัจจัยอาการร้าวของพลาสมาในเลือดและอาการช็อกมีความสัมพันธ์กับคลาสในเชิงบวกระดับสูงและปานกลางตามลำดับ กล่าวคือถ้ามีการร้าวของพลาสมาหรือผู้ป่วยมีอาการช็อกส่งผลให้ความรุนแรงของโรคของผู้ป่วยรายนั้นสูงขึ้นด้วย ในทางเดียวกันถ้าผู้ป่วยมีอาการเลือดออกภายในส่วนใดส่วนหนึ่งของร่างกาย ก็อาจพบว่าอาการของโรคจะรุนแรงกว่าผู้ป่วยที่ไม่มีอาการเลือดออก เป็นต้น ปัจจัยที่น่าสนใจต่อมาคือ จำนวนเกล็ดเลือดน้อยที่สุดที่พบในผู้ป่วย จากค่าสหสัมพันธ์พบว่าจำนวนเกล็ดเลือดมีความสัมพันธ์กับคลาสของผู้ป่วยในเชิงลบระดับปานกลาง กล่าวคือถ้าจำนวนเกล็ดเลือดมีจำนวนน้อยลง ความรุนแรงของโรคในผู้ป่วยรายนั้นๆจะเพิ่มขึ้น และอีกปัจจัยหนึ่งที่น่าสนใจคือขนาดของตับโตขึ้นของผู้ป่วยจากตารางความสัมพันธ์พบว่ามีความสัมพันธ์กับคลาสในเชิงบวกระดับต่ำ กล่าวคือ ถ้าขนาดของตับโตขึ้น ระดับความ

รุนแรงของโรคก็เพิ่มขึ้นด้วยเช่นกัน ส่วนปัจจัยอื่นๆมีความสัมพันธ์กับคลาสของผู้ป่วยเช่นกัน แต่มีระดับต่ำมาก (รายละเอียดความสัมพันธ์ของแอททริบิวต์เพิ่มเติมดูได้จากตารางภาคผนวก ง)

นอกจากนี้ทางผู้ทดลองได้เปรียบเทียบเกณฑ์ที่ได้จากต้นไม้และเกณฑ์ขององค์การอนามัยโลก (World Health Organization: WHO) ด้วยค่า False Negative และค่า False Positive ของแต่ละคลาส เนื่องจากทางการแพทย์ค่า False Negative (FN) มีความสำคัญในด้านการรักษาและการป้องกัน ค่า False Negative คือค่าที่ระบบทำนายเป็นว่าไม่เป็นคลาสที่สนใจ แต่จริงๆแล้วผลเฉลยเป็นคลาสที่สนใจ

อีกค่าหนึ่งคือค่า False Positive (FP) คือค่าที่ระบบทำนายว่าเป็นคลาสที่สนใจแต่จริงๆแล้วผู้ป่วยไม่ได้เป็นคลาสนั้น ค่านี้มีความสำคัญทางด้านธุรกิจมากกว่าด้านการรักษา แต่ในบางกรณีอาจส่งผลกับการรักษา ยกตัวอย่างเช่น ถ้าผู้ป่วยที่ป่วยเป็นโรคไข้แดงก็แต่แพทย์วินิจฉัยว่าเป็นไข้เลือดออกในปริมาณมาก โรงพยาบาลอาจมีผู้ป่วยเพิ่มขึ้นจนทำให้ไม่มีห้องพักหรือปัจจัยที่จะรักษาผู้ป่วยที่ป่วยเป็นโรคไข้เลือดออกในระดับรุนแรง ค่าดังกล่าวแสดงในตารางดังนี้

ตารางที่ 16 ค่า False Negative และ False Positive เทียบกับเกณฑ์ของผลการทดลอง โจทย์ที่ 1

Class	จำนวน	จำนวนผู้ป่วย		FN (%)	จำนวนFP	FP (%)
	ผู้ป่วยตาม คลาสจริง	จากเกณฑ์ ต้นไม้ตัดสินใจ	จำนวนFN			
DF	488	482	13	2.66	7	1.43
DHF I	222	211	23	10.36	12	5.41
DHF II	229	243	10	4.37	24	10.48
DHF III	62	65	1	1.61	4	6.45

จากตารางที่ 16 พบว่าในคลาส DF ค่า False Negative ของงานวิจัยมีค่ามากกว่าค่าที่ได้จากการใช้เกณฑ์ขององค์การอนามัยโลก ในตารางที่ 17 แต่ในส่วน of คลาสอื่นๆ ค่า False Negative มีค่าน้อยกว่าค่าที่ได้จากการใช้เกณฑ์ขององค์การอนามัยโลกอย่างมาก และเมื่อพิจารณาจากค่า False Positive พบว่าเกณฑ์ของจากต้นไม้ให้ค่าน้อยกว่าเกณฑ์ขององค์การอนามัยโลกซึ่งถ้าพิจารณาผลจากเกณฑ์ขององค์การอนามัยโลกพบว่า มีค่า False Positive น้อยลงเมื่อระดับความรุนแรงเพิ่มขึ้น

ตารางที่ 17 ค่า False Negative และ False Positive เทียบกับเกณฑ์ขององค์การอนามัยโลก

Class	จำนวนผู้ป่วย ตามคลาสจริง	จำนวนจาก เกณฑ์ WHO	จำนวน FN	FN (%)	จำนวนFP	FP (%)
DF	488	819	8	1.64	339	69.47
DHF I	222	137	156	70.27	71	31.98
DHF II	229	41	198	86.46	10	4.37
DHF III	62	0	57	91.94	0	0

พบว่าค่า False Negative ที่ได้จากงานวิจัยมีค่าไม่เกิน 10% ซึ่งแตกต่างกับค่าที่ได้จากเกณฑ์ขององค์การอนามัยโลก ที่มีค่าสูงถึง 70 ถึง 90% จากนั้นเมื่อพิจารณาจากข้อมูลจำนวนคนของผลการจำแนกพบว่า เกณฑ์ต้นไม้ตัดสินใจให้ค่า False Positive น้อยกว่าค่าที่ได้จากการเปรียบเทียบจากเกณฑ์ขององค์การอนามัยโลก ดังนั้นสามารถกล่าวได้ว่าเกณฑ์ที่ได้จากการทดลองต้นไม้ตัดสินใจในโจทย์ที่ 1 มีประสิทธิภาพดีกว่าเกณฑ์ขององค์การอนามัยโลก

การทำนายวันที่ไข้ลดในโจทย์ที่ 3 ด้วยขั้นตอนวิธีโครงข่ายประสาทเทียมพบว่าผลการทดลองที่ได้มีค่าความถูกต้องสูงที่สุดแต่ก็ยังไม่สูงเท่าที่ต้องการ ทั้งนี้อาจเกิดจากจำนวนข้อมูลในส่วน day0 เพราะผู้ป่วยส่วนใหญ่มักเข้ามารักษาเมื่ออาการหนักในระดับหนึ่งแล้ว ดังนั้นข้อมูลวันก่อนหน้าวัน day0 ที่นำมาใช้ทำนายจึงมีปริมาณไม่เพียงพอต่อการเรียนรู้ของขั้นตอนวิธี อีกทั้งในปัจจุบันทางการแพทย์ยังไม่สามารถบอกได้ถึงปัจจัยที่สามารถสังเกตวัน day0 ได้อย่างใกล้เคียงจึงทำให้ผลลัพธ์ของวิธีการที่ต้องใช้ปัจจัยในการตัดสินใจยังไม่เป็นที่น่าพอใจเท่าที่ควร

จากผลการทดลองโจทย์ที่ 4 พบว่าอาการตับโตมีความสัมพันธ์กับการเกิดโรคไข้เลือดออกอยู่ในระดับต่ำ และถ้าดูจากผลในต้นไม้แล้วพบว่า อาการตับโตมีอำนาจในการช่วยจำแนกระดับของไข้เลือดออกได้ด้วย และเมื่อวิเคราะห์เฉพาะผู้ป่วยไข้เลือดออกที่มีอาการตับโตพบว่าปัจจัยที่ช่วยในการจำแนกเพิ่มเติมดังแสดงในผลการทดลองโจทย์ที่ 4 และจากการวิเคราะห์โดยยึดอาการตับโตเป็นหลักและตัดเอาทริบิวท์ leakage, shock และ Effusion_Result ออกไปพบว่า กฎที่ได้จากต้นไม้ตัดสินใจประกอบด้วยเงื่อนไขจำนวนมากต่อการจำแนกกฎแต่ละข้อ ดังนั้นโอกาสที่เงื่อนไขจะเป็นจริงเป็นไปได้ยาก ทำให้กฎที่ได้ไม่สามารถนำไปใช้ได้จริง

โจทย์ที่ 5 ผู้ทดลองต้องการพิจารณาเอทรีบิวท์ je หรือเอทรีบิวท์ที่บอกว่าผู้ป่วยได้รับวัคซีนโรคไข้มองอักเสบหรือไม่ จากการทดลองผู้ทดลองค้นไม่ตัดสินใจผลลัพธ์ของโจทย์ที่ 5 มีเอทรีบิวท์ je ปรากฏในค้นไม่ แต่ปรากฏอยู่ในระดับล่างของค้นไม่ ดังนั้นผู้ทดลองจึงได้หาค่าสหสัมพันธ์ของเอทรีบิวท์ที่ปรากฏเทียบกับคลาส ได้ผลดังนี้

ตารางที่ 18 ค่าสหสัมพันธ์ระหว่างคลาสและปัจจัยที่ปรากฏในผลการทดลองโจทย์ที่ 5

ชื่อปัจจัย / ชื่อเอทรีบิวท์	ค่าสหสัมพันธ์	ชนิดและระดับความสัมพันธ์
leakage	0.845613	เชิงบวก สูง
shock	0.537681	เชิงบวก ปานกลาง
Bleeding	0.371913	เชิงบวก ต่ำ-ปานกลาง
je	-0.05606	ไม่มีความสัมพันธ์
platelet_min	-0.58527	เชิงลบ ปานกลาง
hct_max	0.326651	เชิงบวก ต่ำ
hct_max_d05	0.264204	เชิงบวก ต่ำมาก
uri_found	-0.06759	เชิงลบ ต่ำมาก
AST_max	0.239272	เชิงบวก ต่ำมาก
wbc_max	0.130681	เชิงบวก ต่ำมาก

จากค่าสหสัมพันธ์พบว่าเอทรีบิวท์หลักที่มีความสำคัญในการจำแนกผู้ป่วยโรคไข้มองอักเสบยังคงเป็น leakage, shock, bleeding และ platelet_min ส่วนเอทรีบิวท์ uri_found, AST_min และ wbc_max ที่ใช้ในการจำแนกผู้ป่วย 19 คนจากผู้ป่วยทั้งหมด 713 คน ก็ถือว่ามีความสำคัญอยู่ในระดับต่ำ เอทรีบิวท์ที่มีค่าความสัมพันธ์ต่ำมากและอาจกล่าวได้ว่าไม่มีความสัมพันธ์ได้แก่ uri_found และเอทรีบิวท์ที่ไม่มีความสัมพันธ์กับการเกิดคลาสเลยคือ je ดังนั้นจากการทดลองสามารถบอกได้ว่าเอทรีบิวท์ je ไม่เกี่ยวข้องกับการเกิดโรคไข้มองอักเสบจริง กล่าวคือผู้ป่วยที่ได้รับวัคซีนโรคไข้มองอักเสบแล้วยังคงสามารถเป็นโรคไข้มองอักเสบในภายหลังได้ และเมื่อวิเคราะห์เฉพาะผู้ป่วยที่ได้รับวัคซีน JE พบว่ากฎของค้นไม่ตัดสินใจที่ใช้จำแนกคลาสในแต่ละกฎมีความถูกต้องต่ำ โดยกฎที่ดีที่สุดในการจำแนกสามารถจำแนกได้ถูกต้องเพียง 40% ของผู้ป่วยทั้งหมดในคลาสนั้น ดังนั้นสรุปได้ว่ากฎที่ได้จากการทดลองเพิ่มเติมอาจไม่สามารถนำไปใช้งานจริงได้

โจทย์ที่ 6 จากผลการพิจารณาอัตราส่วนผู้ป่วยที่มีค่า AST มากกว่า ALT ตามระดับความรุนแรงของโรคไข้เลือดออกพบว่า เมื่อระดับความรุนแรงของโรคไข้เลือดออกเพิ่มขึ้น อัตราส่วนของผู้ป่วยที่มีค่า AST มากกว่า ALT มีค่ามากขึ้นด้วย สามารถสังเกตได้จากอัตราส่วนของคลาส DHF III พบว่าอัตราส่วนผู้ป่วยที่มีค่า AST มากกว่า ALT มีค่าเท่ากับ 100% หรือกล่าวคือผู้ป่วยคลาส DHF III ทุกคนมีค่า AST มากกว่า ALT นั่นเอง นอกจากนี้แพทย์ต้องการทราบว่าค่า threshold ของผลค่า AST มากกว่า ALT ซึ่งจากการวิเคราะห์การกระจายตัวของข้อมูลพบว่ายังไม่สามารถหาค่านั้นได้ในข้อมูลชุดที่ทดลอง

การพิจารณาค่าความผิดปกติของเม็ดเลือดขาวชนิด atypical lymphocyte ในโจทย์ที่ 7 พบว่า ผู้ป่วยโรคไข้เดงกีและผู้ป่วยโรคไข้เลือดออกส่วนใหญ่มีความผิดปกติของเม็ดเลือดขาว หรือกล่าวได้ว่าผู้ป่วยที่มีการติดเชื้อไข้เดงกี (ผู้ป่วยคลาส DF และคลาส DHF) มีโอกาสพบค่าความผิดปกติของเม็ดเลือดขาวชนิด atypical lymphocyte สูง แต่ไม่จำเป็นต้องมีอาการนี้ในผู้ป่วยทุกรายที่มีการติดเชื้อไข้เดงกี

โจทย์ที่ 8 ผลการแบ่งกลุ่มด้วยวิธีการเคมีนไม่สามารถแบ่งกลุ่มข้อมูลตามความต้องการที่ทางผู้วิจัยสนใจได้ ดังนั้นผู้วิจัยจึงเลือกใช้ขั้นตอนวิธีต้นไม้ตัดสินใจอีกครั้ง ผลการทดลองที่ได้พบปัจจัยที่น่าสนใจที่สามารถจำแนก DHF I กับ DHF II ได้ดีกว่าโจทย์ที่ 1 เปรียบเทียบได้จากค่า False Negative ของคลาส DHF I และ DHF II ของโจทย์ที่ 1 และ โจทย์ที่ 8 พบว่าค่า False Negative ของโจทย์ 8 มีค่าน้อยกว่าค่าของโจทย์ที่ 1 ดังตารางที่ 19 จึงกล่าวได้ว่าปัจจัยที่ได้จากผลลัพธ์ต้นไม้ของโจทย์ที่ 8 สามารถนำไปเป็นปัจจัยในการจำแนกคลาส DHF I ออกจากคลาส DHF อื่นๆได้

ตารางที่ 19 อัตราส่วนค่า False Negative เปรียบเทียบ โจทย์ที่ 1 และ โจทย์ที่ 8

โจทย์ที่สนใจ	DF	DHF I	DHF II	DHF III
โจทย์ที่ 1	13	23	10	1
โจทย์ที่ 8	1	14	4	0
เปอร์เซ็นต์ที่ลดลง (%)	92.31	39.13	60.0	100

นอกจากนี้ทางผู้วิจัยได้เปรียบเทียบผลกับงานของ Ibrahim จากเกณฑ์ที่ค้นพบมาแบ่งคลาสของข้อมูลออกเป็น Low Risk กับ High Risk เปรียบเทียบ Low Risk ได้กับคลาส Group1 และ

เปรียบเทียบ High Risk ได้กับคลาส Group2 ผลการเปรียบเทียบแสดงในค่า False Negative และค่า False Positive ดังตารางที่ 20 จากตารางที่ 20 พบว่าค่าของ False Negative (FN) และ False Positive (FP) ของเกณฑ์โทยี่ 8 ให้ค่าสองค่าน้อยกว่าค่าที่ได้จากเกณฑ์ของ Ibrahim ดังนั้นแสดงว่าในฐานข้อมูลไข้เลือดออกชุดนี้ เกณฑ์ของโทยี่ 8 สามารถแยกแยะกลุ่มความเสี่ยงได้มีประสิทธิภาพดีกว่าเกณฑ์ของ Ibrahim นอกจากนั้นได้คิดค่าสหสัมพันธ์ของปัจจัยของต้นไม้ได้ผลดังตารางที่ 21 พบว่าแอททริบิวต์ส่วนใหญ่มีความสัมพันธ์กับคลาสทั้งหมดแต่อาจอยู่ในระดับเล็กน้อยไม่เท่ากัน ยกเว้นแอททริบิวต์ uri_found ที่ไม่มีความสัมพันธ์กับการเกิดคลาส

ตารางที่ 20 อัตราส่วนค่า False Negative เทียบ โทยี่ 8 กับเกณฑ์ของ Ibrahim

เกณฑ์การวัด	โทยี่ 8				Ibrahim			
	FN	FN (%)	FP	FP (%)	FN	FN (%)	FP	FP (%)
Group1 (Low Risk)	16	2.25	3	0.42	220	30.99	42	5.92
Group2 (High Risk)	3	1.03	16	5.50	42	14.43	220	75.60

ตารางที่ 21 ค่าสหสัมพันธ์ระหว่างคลาสและปัจจัยที่ปรากฏในผลการทดลองโทยี่ 8

ชื่อปัจจัย / ชื่อแอททริบิวต์	ค่าสหสัมพันธ์	ชนิดและระดับความสัมพันธ์
leakage	0.6101	เชิงบวก ปานกลาง
Bleeding	0.5655	เชิงบวก ปานกลาง
shock	0.3884	เชิงบวก ต่ำ
platelet_min	-0.4413	เชิงลบ ต่ำ
uri_found	-0.0374	ไม่มีความสัมพันธ์
liversize_avg	0.3393	เชิงบวก ต่ำ
temperature_max	0.0838	เชิงบวก ต่ำมาก
sbp_minus_dbp_avg	-0.0609	ไม่มีความสัมพันธ์

สรุปและข้อเสนอแนะ

สรุป

วัตถุประสงค์ของงานวิจัยคือ นำเสนอวิธีการสกัดองค์ความรู้จากข้อมูลด้วยการทำเหมืองข้อมูลทางการแพทย์ วิธีการที่เลือกใช้ขึ้นอยู่กับวัตถุประสงค์ของการทดลอง วิธีการจำแนกประเภทข้อมูลที่ใช้ในการวิจัยคือ ขั้นตอนวิธีต้นไม้ตัดสินใจร่วมกับขั้นตอนวิธีฟัซซี่ลอจิก วิธีการจัดกลุ่มข้อมูลที่ใช้ในการทดลองคือ ขั้นตอนวิธีเคมีน วิธีการหาความสัมพันธ์ระหว่างแอททริบิวต์ที่ใช้คือ ขั้นตอนวิธีกฎความสัมพันธ์ และการคำนวณค่าสหสัมพันธ์

จากงานวิจัยการจำแนกประเภทของโรคไข้เลือดออก ผู้ทดลองเลือกใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ ผลการทดลองพบว่าปัจจัยที่ส่งผลต่อการจำแนกโรคไข้เลือดออกที่น่าสนใจคือ อาการร้าวของพลาสมาในเลือด อาการช็อก ซึ่งทั้งสองปัจจัยที่กล่าวข้างต้นมีความสัมพันธ์กับคลาสในระดับสูงและระดับปานกลางตามลำดับ ปัจจัยอื่นๆที่น่าสนใจต่อมาคือ จำนวนเกล็ดเลือดน้อยที่สุดที่พบในผู้ป่วย ซึ่งมีความสัมพันธ์ในเชิงลบระดับปานกลาง ส่วนปัจจัยอื่นๆอันได้แก่ ค่าเฉลี่ยของเอนไซม์ Alanine aminotransferase ในตับ ค่าเฉลี่ยของจำนวนเม็ดเลือดขาว อาการต่อน้ำเหลืองโต และค่าน้อยสุดของจำนวนเม็ดเลือดขาว มีความสัมพันธ์กับคลาสแต่ในระดับต่ำมาก ต่อมาได้มีการนำวิธีการฟัซซี่ลอจิกมาใช้ในการพิจารณาปัจจัยที่มีค่าเป็นตัวเลขที่ปรากฏในผลการทดลอง และจากผลการทดลองที่ได้ ทำให้ผู้ทดลองพบรูปแบบการเกิดโรคว่าควรพิจารณาปัจจัยใดและค่าของปัจจัยที่สนใจเพื่อพิจารณาคาสของผู้ป่วยคนดังกล่าว

ส่วนงานวิจัยการทำนายวัน day0 ทางผู้วิจัยได้ทดลองใช้ขั้นตอนวิธีโครงข่ายประสาทเทียมพบว่า ค่าการทำนายวัน day0 มีค่าความถูกต้องเท่ากับ 69.80% ซึ่งเห็นว่าค่าที่ได้มีค่าไม่สูงมาก สาเหตุเนื่องมาจากข้อมูลที่ไม่เพียงพอต่อการเรียนรู้ เพราะปกติของผู้ป่วยที่เข้ามารับการรักษาส่วนใหญ่เข้ามาเมื่อมีอาการหนัก ซึ่งก็เป็นวันที่ใกล้กับ day0 หรือวัน day0 ทำให้ข้อมูลวันก่อนหน้า day0 มีไม่เพียงพอสำหรับการเรียนรู้ ดังนั้นค่าความถูกต้องของวันก่อนหน้า day0 จึงมีค่าน้อยมาก

จากการวิเคราะห์โจทย์ย่อยอื่นๆพบว่า อาการตับโตเป็นปัจจัยที่ช่วยในการจำแนกผู้ป่วยโรคไข้เลือดออกได้ สังเกตจากผลลัพธ์ต้นไม้ตัดสินใจที่อาการตับโตสามารถจำแนกผู้ป่วยได้กับผู้ป่วยโรคไข้เลือดออกได้ อาการตับโตมีความสัมพันธ์กับการเกิดโรคในทางบวก กล่าวคือ เมื่อ

ระดับของไข้เลือดออกเพิ่มขึ้น อัตราส่วนของผู้ป่วยที่มีอาการตับโตก็เพิ่มขึ้นด้วยเช่นเดียวกัน และเมื่อวิเคราะห์เรื่องการยึดเกณฑ์การจำแนกคลาสด้วยอาการตับโตพบว่า เกณฑ์ที่ดีที่สุดที่ได้จากต้นไม้ตัดสินใจให้ค่าความถูกต้องในการจำแนกไม่ถึง 50% ซึ่งหมายถึงกฎที่ได้อาจไม่สามารถนำไปใช้ได้จริงนั่นเอง สำหรับเรื่องของวัคซีนโรคไข้สมองอักเสบ บุคคลใดที่ได้รับวัคซีนโรคไข้สมองอักเสบแล้ว บุคคลนั้นก็ยังจะสามารถป่วยเป็นโรคไข้เลือดออกได้ แต่ถ้าผู้ป่วยจริง ระดับความรุนแรงของผู้ป่วยที่ได้รับวัคซีนแล้วจะน้อยกว่าผู้ป่วยที่ไม่ได้รับวัคซีนมาก่อนเลย และเมื่อวิเคราะห์การจำแนกโดยยึดจากข้อมูลผู้ป่วยที่ได้รับวัคซีน JE พบว่าเกณฑ์ที่ดีที่สุด ต้นไม้ตัดสินใจให้ความถูกต้องไม่เกิน 40% ซึ่งแสดงให้เห็นถึงความไม่มีประสิทธิภาพของเกณฑ์ที่ได้เมื่อตัดแอมพริบที่ที่ต้องใช้เวลาและค่าใช้จ่ายสูงออกไป และในผู้ป่วยที่ติดเชื้อโรคไข้แดงก็ยังพบอีกว่าอัตราส่วนผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่า ALT มีค่าเท่ากับ 84.39% และอัตราส่วนผู้ป่วยที่มีค่าเอนไซม์ AST มากกว่า ALT ในผู้ป่วยโรคไข้เลือดออกมีค่าเท่ากับ 97.66% และเมื่อวิเคราะห์แยกตามระดับความรุนแรงของโรคไข้เลือดออกพบว่าในผู้ป่วยโรคไข้เลือดออกที่มีระดับความรุนแรงสูง จะมีอัตราส่วนผู้ป่วยที่มีค่า AST มากกว่า ALT สูงขึ้นด้วยเช่นเดียวกัน สืบเนื่องจากผู้ป่วยโรคไข้เลือดออกระดับสามทุกคนมีค่า AST มากกว่า ALT จากนั้นผู้วิจัยต้องการศึกษาว่าผู้ป่วยไข้เลือดออกมีความผิดปกติของเม็ดเลือดขาวทุกรายหรือไม่ จากการวิเคราะห์พบว่า ทั้งผู้ป่วยโรคไข้แดงก็และผู้ป่วยโรคไข้เลือดออกมีอัตราส่วนผู้ป่วยที่มีค่าความผิดปกติของเม็ดเลือดขาวเท่ากับ 93.22 และ 98.05% ตามลำดับ ดังนั้นไม่สามารถกล่าวได้ว่าผู้ป่วยโรคไข้เลือดออกต้องมีค่าความผิดปกติของเม็ดเลือดขาวทุกคน ความผิดปกติของเม็ดเลือดขาวเป็นเพียงอาการที่แสดงออกในผู้ป่วยที่ติดเชื้อไข้เลือดออกเท่านั้น

จากโจทย์ 8 ผู้วิจัยจึงเลือกการแบ่งกลุ่มตามธรรมชาติด้วยขั้นตอนวิธีเคมีน ผลการทดลองพบว่าไม่สามารถแยกผู้ป่วยคลาส DHF I และ DHF II ได้เนื่องจากวิเคราะห์จากสัดส่วนการแบ่งกลุ่มพบว่าข้อมูลทั้งในกลุ่มที่ 1 และกลุ่มที่ 2 มีข้อมูลของผู้ป่วยในทุกๆคลาสจึงทำให้วิธีการนี้และคลาสใหญ่ของข้อมูลทั้งสองกลุ่มเป็นคลาสเดียวกัน ดังนั้นไม่สามารถบอกได้ว่าข้อมูล DHF I และ DHF II แตกต่างกันหรือไม่ ต่อมาจึงได้ใช้วิธีการขั้นตอนวิธีต้นไม้ตัดสินใจโดยการแบ่งคลาสใหม่เป็น Group1 และ Group2 โดยที่ Group1 ประกอบด้วยคลาส DF และ DHF I และ Group2 ประกอบด้วยคลาส DHF II และ DHF III โดย Group1 แทนกลุ่มผู้ป่วยที่มีความรุนแรงน้อย และ Group2 แทนกลุ่มผู้ป่วยที่มีความรุนแรงมาก จากผลการทดลองพบว่าปัจจัยอาการช็อก จำนวนเกล็ดเลือดน้อยที่สุด อาการตับโต อุณหภูมิของผู้ป่วย และค่าความต่างของความดันสูง-ต่ำโดยเฉลี่ยที่พบในผู้ป่วยสามารถช่วยจำแนกคนไข้ DHF I และ DHF II เพิ่มเติมได้ละเอียดมากยิ่งขึ้น ซึ่งต้นไม้ตัดสินใจนี้ให้ค่าความถูกต้องเท่ากับ 96.4%

ข้อเสนอแนะ

การทดลองในงานวิจัยใช้ค่าตัวแทนของข้อมูลในลักษณะของค่ามากที่สุด ค่าน้อยสุด และค่าเฉลี่ยใน 4 วันแรกที่ผู้ป่วยเข้ารับรักษา ดังนั้นสามารถพัฒนาต่อในส่วนของการจัดการเรื่อง temporal data เพื่อหาตัวแทนข้อมูลในลักษณะแนวโน้มเข้ามาพิจารณาด้วย อีกทั้งงานวิจัยฉบับนี้เน้นการจำแนกประเภทของโรคไข้เลือดออกและจากการทดลองให้ค่าความถูกต้องในการจำแนกเท่ากับ 97.39% แต่ทั้งนี้ก็ยังจะสามารถพัฒนาประสิทธิภาพต่อได้ด้วยการทดลองเปลี่ยนชนิดของฟังก์ชันความเป็นสมาชิกในขั้นตอนวิธีฟัซซี่ลอจิก และการเขียนโปรแกรมเพื่อหาจุดเหลื่อมล้ำที่ให้ค่าความถูกต้องมากที่สุดได้

ส่วนการทดลองการทำนาย day0 ของงานวิจัยนี้สามารถพัฒนาต่อได้ด้วยการหาขั้นตอนวิธีการคัดเลือกแอททริบิวต์ที่มีอิทธิพลต่อการทำนาย day0 หรือเก็บข้อมูลเพิ่มเติมในส่วนก่อนการเกิด day0 เพื่อให้ได้ข้อมูลที่เพียงพอต่อการเรียนรู้

เรื่องของการหาปัจจัยเพื่อจำแนก DHF I ออกจาก DHF คลาสอื่นๆ ยังสามารถพัฒนาต่อได้โดยหาวิธีการแบ่งกลุ่มข้อมูลที่ดีกว่า K-Mean เช่น Density Based Clustering (DBScan) เป็นต้น อาจทำให้ได้ปัจจัยหรือความสำคัญบางอย่างในการช่วยจำแนก DHF I ออกจาก DHF คลาสอื่นๆ

เอกสารและสิ่งอ้างอิง

นวลวรรณ สุนทรภิชัย. 2553. **ปัญญาประดิษฐ์**. โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย
กรุงเทพมหานคร

Cambridge University. 2008. **Evaluation of Clustering**. Available Source:

<http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>,
November 28, 2012.

Delen, D., G. Walker and A. Kadam. 2005. Predicting breast cancer survivability: a comparison
of three data mining methods. **Artificial Intelligence in Medicine** 34:113–127.

Faisal Tarig, Ibrahim Fatimah and Taib Mohd Nasir. 2010a. A noninvasive intelligent approach
for predicting the risk in dengue patients. **Expert Systems with Applications** 37(3):
2175-2181, ISSN 0957-4174.

_____. 2010b. Neural network diagnostic system for dengue patients risk classification. **J. Med.
Syst.**

_____. 2010c. Reexamination of risk criteria in dengue patients using the self-organizing map.
Medical and Biological Engineering and Computing. 48:293-301.

Frank, E., Paynter, G.W., I. H. Witten, C. Gutwin and C. G. Nevill-Manning. 1999. **Domain
specific keyphrase extraction**. pp. 668–673. In Proceedings of the 16th International
Joint Conference on Artificial Intelligence, Morgan Kaufmann.

Hamburg. 2008. **Evaluation of Cluster Quality**. 83-96. Available Source:

<http://dissertations.ub.rug.nl/FILES/faculties/arts/2010/t.van.de.cruys/06c6.pdf>,
November 28, 2012.

- Ho Tu Bao, Nguyen Trong Dung, Kawasaki Saori, Le Si Quang, Nguyen Dung Duc, Yokoi Hideto and Takabayashi Katsuhiko. 2003. **Mining Hepatitis Data with Temporal Abstraction**. pp. 369 – 377. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.
- Holmesd, G., Cummingham, S. J., B. D. Rue and F. Bollen. 1998. Predicting apple bruising using machine learning. **Acta Hort.** 476: 289-296.
- Ibrahim Fatimah, Taib Mohd Nasir, Abas Wan Abu Bakar Wan, Guan Chan Chong and Sulaiman. 2005. A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). **Comput. Methods Programs Biomed.** 79: 273–281.
- Kusabs, N., F. Bollen, L. Trigg, G. Holmes and S. Inglis. 1998. **Objective measurement of mushroom quality**. pp 51. In Proc New Zealand Institute of Agricultural Science and the New Zealand Society for Horticultural Science Annual Convention.
- Li, J., H. Liu, J. R. Downing, A. E.-J. Yeoh and L. Wong. 2003. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients. **Bioinformatics** 19:71–78.
- Lin Rang-Ho. 2009. An intelligent model for liver disease diagnosis. **Artificial Intelligence in Medicine** 47(1): 53-62.
- Lin Weiqiang, Orgun Mehmet A. and Williams Graham J. 2002. **An Overview of Temporal Data temporal Data Mining**. pp. 83–90. In Proceedings of the 1st Australian Data Mining Workshop (ADM'02), Canberra, Australia.
- Laxman Srivatsan and P.S. Sastry. 2006. A survey of temporal data mining. **Proceedings in Engineering Sciences** 31(2): 173-198.

N. Lavrac. 1999. Selected techniques for data mining in medicine. **Artificial Intelligence in Medicine** 16 (2): 3-23.

Ratanamahatana Chotirat (Ann), Lin Jessica, Gunopulos Dimitrios, Keogh Eamonn J., Vlachos Michail and Das Gautam. 2005. Mining Time Series Data. pp. 1069-1103. **The Data Mining and Knowledge Discovery Handbook**. Tel-Aviv University, Israel.

Sauban, M. and B. Pfahringer. 2003. **Text categorization using document profiling**. pp. 411–422. In Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases.

Tanner Lukas, Schreiber Mark, Low Jenny, G. H., Ong Adrian, Tolfvenstam Thomas, Lai Yee Ling, Ng Lee Ching, Leo Yee Sin, Puong Le Thi, Vasudevan Subhash, G., Simmons Cameron P., Hibberd Martin L. and Ooi Eng Eong. 2008. **Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness**. PLoS Negl Trop Dis.

World Health Organization (WHO). 1997. Dengue haemorrhagic fever: Diagnosis, treatment, prevention and control. **WHO Library Cataloguing in Publication Data**. ISBN 92 4 154500 3.



ภาคผนวก



ภาคผนวก ก
แบบฟอร์มการบันทึกข้อมูล



D01-01

คำชี้แจง

กรุณาเขียนตัวเลขบรรทัดลงในช่อง ☐ ช่องละหนึ่งตัวให้มีขนาดในอุปกรณ์ไม่แตะเส้นขอบ และเขียนกากบาท (X) ลงในช่อง ☐

ซึ่งเป็นช่องเพื่อใช้สำหรับกากบาทเท่านั้น ไม่ใช้ในการกรอกตัวอักษรต่าง ๆ

ตัวเลขที่โปรแกรมสุ่มกันผ่านได้ถูกต้องสูงสุดมีลักษณะดังนี้ 1 2 3 4 5 6 7 8 9 0

โปรดใช้ปากกาสีน้ำเงินเขียน ขนาด 0.5 ขึ้นไป

SNV:

สำหรับสถิติเกอร์ชื่อผู้ป่วย

1 8 ธิ.ย. 2551

1. Age 1-15 years age 1.No ☒ 2.Yes2. Fever ≥ 38.5 (Axilla) fever_inclusion 1.No ☒ 2.Yes3. Duration of fever ≤ 5 days duration_fever 1.No ☒ 2.Yes4. Tourniquet test If fever ≥ 3 days, TT $\geq 10/\text{Inch}^2$ tt1 1.No ☒ 2.YesIf fever ≤ 2 days, TT $\geq 5/\text{Inch}^2$ tt2 1.No ☒ 2.Yes6. No obvious source of infection infection 1.No ☒ 2.Yes6. Informed consent consent 1.No ☒ 2.Yes7. WBC (d4, d5) $\leq 7,000$ wbc 1.No ☒ 2.Yes ☒ 3. Not Done8. PCR screening positive pcr_screen 1.No ☒ 2.Yes ☒ 3. Not Done9. มีการระบาดในพื้นที่อาศัย outbreak 1.No ☒ 2.Yes ☒ 3. Not Done

nurse

Study Nurse..... Date

Check by nurse..... Date

Version 1.00000001

Page 1

Date: 11/11/2011 11:11:11 AM

29



D01-02

Form 1 Enroll

1. Birth date (dd/mm/yyyy)

 / /

2. Place of birth (Province)

☐ 1. Khon Kaen ☐ 2. Song Kla ☐ 3. Other.....

3. Religion

☐ 1. Buddhism ☐ 2. Christianity ☐ 3. Islam ☐ 4. Other.....

4. Current address : House No.....

Moo.....

Soi.....

Road.....

District.....

Amphur.....

Province.....

Zipcode.....

Tel.....

Comment.....

Map (draw map of address)

5. Guardian Information

Name.....

Age.....

 Year

Relationship.....

☐ 1. Father ☐ 2. Mother ☐ 3. Grandfather

☐ 4. Grandmother

☐ 5. Baby sitter ☐ 6. Relative (specify.....)

Address for contact.....

Tel.....

6. Ever received JE vaccine ? (if N/U skip to Q.7) ☐ 1. No ☐ 2. Yes ☐ 3. Unknown

If yes ,specify source and injection (s) :

Source.....

☐ 1. Health book ☐ 2. OPD ☐ 3. ตามจากญาติ
1st Injection.....
☐ 1. Yes , when
☐ 2. Unknown
2nd Injection.....
☐ 1. No ☐ 2. Yes , when
☐ 3. Unknown
3rd Injection.....
☐ 1. No ☐ 2. Yes , when
☐ 3. Unknown

7. Height.....

 cm

8. ผู้ป่วยเคยเป็นไข้เลือดออกหรือไม่

☐ 1. เคย ☐ 2. ไม่เคย



D01-03

Admission Date (dd/mm/yyyy)	12 34 5678	สำหรับสถิติการชื้อผู้ป่วย ป.8 ส.ย. 2551
Enrollment Date (dd/mm/yyyy)	12 34 5678	
Discharge Date (dd/mm/yyyy)	12 34 5678	
เกิดเหตุการณ์/ความผิดปกติที่ไม่คาดหวังหรือไม่ ขณะเข้ารับการรักษาผู้ป่วยเด็กใด		
accident	☐ 1.ไม่มี ☐ 2.มี ระบุ..... วันที่..... แก้ไขโดย.....	
Withdrawal Date (dd/mm/yyyy)	12 34 5678	
Withdrawal reason.....		
Clinical Dx at admission	☐ 1.AFI ☐ 2.DF ☐ 3.DHF ☐ 4.R/O DF ☐ 5.R/ODHF ☐ 6.Acute dengue infection ☒ 7.Other.....	
Any other febrile illness during the past 2 months? ☐ 1.No ☐ 2.Yes ☐ 3.Unknown		
1.fever	☐ 1.No ☐ 2.Yes	
Date 1 st ill (onset of fever)	12 34 5678	
fever time	12 34	
2. Headache	☐ 1.No ☐ 2.Yes ☐ 3. Unknown headache day 12 Day	
3. Myalgia	☐ 1.No ☐ 2.Yes ☐ 3. Unknown myalgia day 12 Day	
4. Bone pain	☐ 1.No ☐ 2.Yes ☐ 3. Unknown bone day 12 Day	
5. Retro-orbital pain	☐ 1.No ☐ 2.Yes ☐ 3. Unknown retro day 12 Day	
6. Flushed face	☐ 1.No ☐ 2.Yes ☐ 3. Unknown flushed day 12 Day	
7. Rash without itching	☐ 1.No ☐ 2.Yes ☐ 3. Unknown rash without day 12 Day	
8. Rash with itching	☐ 1.No ☐ 2.Yes ☐ 3. Unknown rash with day 12 Day	
9. Anorexia	☐ 1.No ☐ 2.Yes ☐ 3. Unknown anorexia day 12 Day	
10. Nausea	☐ 1.No ☐ 2.Yes ☐ 3. Unknown nausea day 12 Day	



D01-04

11. Vomiting ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		สำหรับศีกษากรณีผู้ป่วย 11 8 ส.ย. 2551
12. Abdominal pain ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		
13. Drains ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		
14. Bleeding (if N/U skip to Q.15)		
bleeding ☐ 1.No ☐ 2.Yes ☐ 3.Unknown		
14.1 Spontaneous petechiae ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		14.2 Ecchymosis ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk
14.3 Gum ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		14.4 Nose ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk
14.5 Vomiting ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		14.6 Melena/red stool ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk
14.7 Other ☐ 1.No ☐ 2.Yes ☐ 3.Unk Day ☐ 3.Unk		
15. Tourniquet test. (if N/U skip to Q.16)		
tourniquet ☐ 1.No ☐ 2.Yes ☐ 3.Unknown		tourniquet_date ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8
15.1 Quantity Data ☐ 1.0-9 ☐ 2.10-20 ☐ 3.>20		
15.2 Quality Data ☐ 1.Fine ☐ 2.Medium ☐ 3.Coarse ☐ 4.Mixed ☐ 5.Unknown		
16. URI (if N/U skip to Q.17) ☐ 1.No ☐ 2.Yes ☐ 3.Unknown		
16.1 Exudative rhinorrhea ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
16.2 Clear rhinorrhea ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
16.3 Non-productive cough ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
16.4 Productive cough ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
16.5 Feeling of sore throat ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
17. Diarrhea ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
18. Other symptoms..... ☐ 1.No ☐ 2.Yes ☐ 3. Unknown		
Study Nurse..... Date		
Check by nurse..... Date		



D02-01

Form 2 : Drug Taken Before Admission									
SNO : 1 2 3 4 5 6 7 8 9		สำหรับนักศึกษาที่เข้าเรียน ปี 8 ส.ย. 2551							
Date (dd/mm/yyyy) 1 2 / 3 4 / 5 6 7 8									
1. Antipyretic : ยาแก้ปวด 1 1.No 2 2.Yes 3 3.Unknown									
If yes : 1 1.Paracetamol 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2 2.ASA 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
3 3.Brufen 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
4 4.other..... 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2. Antibiotics : ยาปฏิชีวนะ 1 1.No 2 2.Yes 3 3.Unknown									
If yes : 1 1.Amoxycilin 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2 2.Sulfa(Bactrim) 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
3 3.Macrolide 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
4 4.Cephalosporin 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
5 5.Penicillin 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
6 6.other..... 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
3. ORS : ยารักษาท้องเสีย 1 1.No 2 2.Yes 3 3.Unknown									
4. GI Drug : ยาระบบทางเดินอาหาร 1 1.No 2 2.Yes 3 3.Unknown									
If yes : 1 1.Antiemetic 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2 2.Anta-acid 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
5. Respiratory drug : ยาระบบทางเดินหายใจ 1 1.No 2 2.Yes 3 3.Unknown									
If yes : 1 1.Anticough 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2 2.Antihistamine 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
6. Injection drug : ยาฉีดและสารน้ำ 1 1.No 2 2.Yes 3 3.Unknown									
If yes : 1 1.Antipyretic 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
2 2.Antibiotic 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
3 3.Anticonvulsion 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
4 4.Antiemetic 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
5 5.IV fluid 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4									
7.Unknown drug 1 1.No 2 2.Yes 3 3.Unknown									
If yes, Unknown.....									
Study Nurse.....					Date				
Check by nurse.....					Date				



D03-01

Form 3 : Master									
SNO: 1 2 3 4 5 6 7 8 9		สำหรับนักศึกษาที่ผู้ป่วย ๒. 8 ต.ย. 2551							
Date (dd/mm/yyyy) date 1 2 / 3 4 / 5 6 7 8									
Temp max temp_max 1 2 3									
min. temp_min 1 2 3		temp_nd 1 Not Done							
B.P.(SBP/DBP) min sbp 1 2 3 / 1 2 3		bp_nd 1 Not Done							
Pulse pressure min pulse_pre_min 1 2 3		pulse_pre_nd 1 Not Done							
Pulse rate max pulse_rate_max 1 2 3		min pulse_rate_min 1 2 3		pulse_rate_nd 1 Not Done					
Hct (ward) max hct_max 1 2 3		min hct_min 1 2 3		hct_nd 1 Not Done					
Intake(ml) 0-8 AM intake1 1 2 3 4		intake1_nd 1 Not Done							
8 AM - 4 PM intake2 1 2 3 4		intake2_nd 1 Not Done							
4 PM - 12 PM intake3 1 2 3 4		intake3_nd 1 Not Done							
Output (ml) 0-8 AM output1 1 2 3 4		output1_nd 1 Not Done							
8 AM - 4 PM output2 1 2 3 4		output2_nd 1 Not Done							
4 PM - 12 PM output3 1 2 3 4		output3_nd 1 Not Done							
note Note.....									
nurse Study Nurse..... Date Check by nurse..... Date									



D04-01

Form 4/1 : Admission_Clinical

SNO. 1 2 3 4 5 6 7 8 9

date 12 / 34 / 5678

Clinical Findings (morning of evaluation)

1. Mental Status

1. Alert 2. Drowsy 3. Lethargic

4. Restless or Agitated

5. Stuporous (awakens only with vigorous stimulation)

6. Comatose (won't wake up)

2. Tourniquet test (If N/U skip to Q.3) 1. No 2. Yes 3. Unknown

2.1. Quantity quantity 1 2

2.2. Quality quality 1. fine 2. medium 3. coarse 4. mixed 5. unknown

3. Rash (If N/U skip to Q.4) 1. No 2. Yes 3. Unknown

3.1. Confluent petechiae 1. No 2. Yes 3.2. Maculopapular 1. No 2. Yes

3.3. Convalescent 1. No 2. Yes 3.4. Other.....

4. Itching related to rash 1. No 2. Yes 3. Unknown

5. Bruising with venipuncture 1. No 2. Yes 3. Unknown

6. Bleeding (If N/U skip to Q.7) 1. No 2. Yes 3. Unknown

6.1 Spontaneous petechiae 1. No 2. Yes 3. Unknown 6.2 Ecchymosis 1. No 2. Yes 3. Unknown

6.3 Gum 1. No 2. Yes 3. Unknown 6.4 Nose 1. No 2. Yes 3. Unknown

6.5 Vomiting Blood 1. No 2. Yes 3. Unknown 6.6 Melena/red stool 1. No 2. Yes 3. Unknown

6.7 Other.....

7. Diarrhea 1. No 2. Yes 3. Unknown

8. URI (If N/U skip to Q.9) 1. No 2. Yes 3. Unknown

8.1. Exudative 1. No 2. Yes 3. Unknown 8.2. Clear rhinorrhea 1. No 2. Yes 3. Unknown

8.3. Non-productive cough 1. No 2. Yes 3. Unknown 8.4. Productive cough 1. No 2. Yes 3. Unknown

8.5. Feeling of sore throat 1. No 2. Yes 3. Unknown

9. Abdominal pain 1. No 2. Yes 3. Unknown 4. Not done 5. Not available

*10. Dyspnea 1. No 2. Yes 3. Unknown 4. Not done 5. Not available

*11. Ascites 1. No 2. Yes 3. Unknown 4. Not done 5. Not available

*12. Jaundice 1. No 2. Yes 3. Unknown 4. Not done 5. Not available

*13. Liver tenderness 1. No 2. Yes 3. Unknown 4. Not done 5. Not available

*14. Liver size (>RCM)/span (cm) liver_size 1 2 3

*15. Abdominal Circumference (cm) ab_circum 1 2 3 4

*16. Spleen size (cm > LCM) spleen 1

Note * Must be verified by physician

สำหรับสถิติเกอร์ชื่อผู้ป่วย

1. 8 ส.ย. 2554



D04-02

Date (dd/mm/yyyy)			
date2	12 / 34 / 5678		
lymph	*17. Lymph node enlargement ☐ 1.No ☒ 2.Yes ☐ 3.Unknown ☐ 4.Not done ☐ 5.Not available		
	17.1 Cervical (max. diam. cm.)	1	2
	17.2 Epitrochea (max. diam. cm.)	1	2
	17.3 Inguinal (max. diam. cm.)	1	2
	*18. Injected conjunctivae ☒ 1.No ☐ 2.Yes ☐ 3.Unknown ☐ 4.Not done ☐ 5.Not available		
	18.1 Limbus	1	2
	18.2 Bulbar	1	2
	19. Weight	123	4 kg
	20. Total Oral Intake Estimate	1234	ml
	21. Total IV Transfusion	1234	ml
	21.1 Crystalline fluid Tx amount	1234	ml
	21.2 Form crystalline fluid	☐ estimate value	
	1. 5% D/NSS (5% DAR)	1234	ml
	2. 5% D/N/2	1234	ml
	3. 0.9% NSS	1234	ml
	4. Other	1234	ml
	21.3 Dextran	1234	ml
	21.4 Haesteryl	1234	ml
	21.5 Blood	1	2
	1. Whole Blood	1234	ml
	2. Packed Red Cell	1234	ml
	3. Fresh Frozen Plasma	1234	ml
	4. Platelet	1234	unit
	5. Cryoprecipitate	1234	ml
note	Note on unusual manifestation.....		
nurse	Study Nurse..... Date Check by nurse..... Date Note * Must be verified by physician		



D05-01

Form 4/2 : Admission : Routine Laboratory

S/N: 1 2 3 4 5 6 7 8 9

Date (dd/mm/yyyy) 12 / 34 / 5678

สำหรับคลินิกโรคเรื้อรัง

พ. 8 ส.ย. 2551

1. Time Blood 12 . 34 1 Unknown

2. Specimen log number 1 2 3 4 5 6

3. CBC 1 1.No skip to Q.4 2 2.Yes 3 3.Unknown 4 4.Not Available

3.1. Hct (%) 1 2 3

3.2. WBC (X1,000) 1 2 3

3.3. Diff WBC (%) 1 1.No 2 2.Yes 3 3.Unknown 4 4.Not Available

1. PMN 1 2 3 2. Band 1 2 3 3. Lymphocyte 1 2 3 4. Atyp lymphocyte 1 2 3

5. Monocyte 1 2 3 6. Basophil 1 2 3 7. Eosinophil 1 2 3 8. Other 1 2 3

3.4 Platelet count (x1,000) 1 2 3

3.5 Plt (/OF) 1 Unknown 2 <5 3 5-10 4 11-25 5 >25

4.. LFT 1 1..No skip to Q.5 2 2.Yes 3 3.Unknown 4 4.Not Available 5 5.Not Done

4.1 Protein 1 2 3

4.2 Albumin 1 2

4.3 AST 1 2 3 4 5

4.4 ALT 1 2 3 4 5

4.5 Other.....

Note for other lab result

5. U/A 1 1.No skip to Q.6 2 2.Yes 3 3.Unk 4 4.ND

5.1 Color 1 Colorless 2 Pale yellow 3 Yellow 4 Brown 5 Pale red
6 Reddish 7 Green 8 NA 9 Unk 10 ND

5.2 Appearance 1 Clear 2 Slightly turbid 3 Turbid 4 milky 5 NA 6 Unknown 7 Not Done

5.3 pH 1 2

5.4 Spec gravity 1 Not Done 2 < 1.005 3 1.005 4 1.010 5 1.015
6 1.020 7 1.025 8 1.030 9 >1.030 10 NA/Unk

5.5 Protein 1 Negative 2 Trace 3 1+ 4 2+ 5 3+ 6 4+ 7 NA 8 Unk 9 ND

5.6 Sugar 1 Negative 2 Trace 3 1+ 4 2+ 5 3+ 6 4+ 7 NA 8 Unk 9 ND

5.7 Blood 1 Negative 2 Trace 3 1+ 4 2+ 5 3+ 6 4+ 7 NA 8 Unk 9 ND

5.8 WBC (/HPF) 1 0-1 2 1-2 3 2-3 4 3-5 5 5-10 6 10-20
7 20-30 8 30-50 9 50-100 10 TNTC 11 ND5.9 RBC (/HPF) 1 0-1 2 1-2 3 2-3 4 3-5 5 5-10 6 10-20
7 20-30 8 30-50 9 50-100 10 TNTC 11 ND



D05-02

Date (dd/mm/yyyy)			
data2	1 2 / 3 4 / 5 6 7 8		
6. Coagulogram		สำหรับสติกเกอร์ผู้ป่วย 11 8 ส.ย. 2551	
6.1 PT		1.No (skip to Q.6.2) 2.Yes 3.Unk 4.NA pt 1 2 3	
6.1.1 Test		1 2 3	
6.1.2 Control		1 2 3	
6.1.3 NA		1 2 3	
6.2 PTT		1.No (skip to Q.6.3) 2.Yes 3.Unk 4.NA ptt 1 2 3	
6.2.1 Test		1 2 3	
6.2.2 Control		1 2 3	
6.3 TT		1.No (skip to Q.7) 2.Yes 3.Unk 4.NA tt 1 2 3	
6.3.1 Test		1 2 3	
6.3.2 Control		1 2 3	
*7. Right lateral decubitus film		1.No (skip to Q.8) 2.Yes 3.Unk 4.NA effusion index (neg or digit) 1 Negative eff_2 1 2 3	
*8. Right lateral decubitus Ultrasound		1.No 2.Yes 3.Unk 4.NA effusion amount eff_amount 1 2 3 4 5	
note2	Note 		
nurse	Study Nurse..... Date Check by nurse..... Date Note * Must be verified by physician		



D06-01

Form 5 : Clinical Dx Vs Final Dx	
SNO: 1. Outcome ☐ 1.Recovered ☐ 2.Disabled ☐ 3.Died 2. Time of Subsidence of fever Date ☐ Unknown Time : ☐ Unknown 3. Leakage ☐ 1.No ☐ 2.Yes ☐ 3.Unknown 3.1. Hemoconcentration ☐ 1.No ☐ 2.Yes 3.3. Ascites ☐ 1.No ☐ 2.Yes 3.2. Pleural effusion ☐ 1.No ☐ 2.Yes 3.4. Plasma alb ☐ 1.No ☐ 2.Yes 4. Shock ☐ 1.No ☐ 2.Yes ☐ 3.Unknown If yes , ใน shock ครั้งแรก รายละเอียดเริ่มต้นตั้งแต่วันที่จนกระทั่งสิ้นสุดถึงวันที่และเวลานั้น 4.1 From : date / / ☐ Unknown Time : ☐ Unknown 4.2 To date / / ☐ Unknown Time : ☐ Unknown 4.3 Vital sign evidence of shock 1. Low SBP ☐ 1.No ☐ 2.Yes 2. Pulse Pressure < 20 ☐ 1.No ☐ 2.Yes 3. No Pressure ☐ 1.No ☐ 2.Yes 4.4 Clinical evidence of shock ☐ 1.No ☐ 2.Yes ☐ 3.Unknown 4.4.1 Mental status change ☐ 1.No ☐ 2.Yes 4.4.2. Cold extremities ☐ 1.No ☐ 2.Yes 4.4.3. Delay capillary refill > 2sec ☐ 1.No ☐ 2.Yes 5. Clinical Dx at Discharge ☐ 1.DF ☐ 2.DHF grade 1 ☐ 3.DHF grade 2 ☐ 4.DHF grade 3 ☐ 5.DHF grade 4 ☐ 6.Non Dengue Any remarkable aspects on this patient? PI..... Date.....	สำหรับศดึกเกอร์ชื่อผู้ป่วย



D07-01

Form 4_2 : Follow Up	
SNO:	สำหรับติดตามการรักษาผู้ป่วย
Follow Up ☐ มา ☐ ไม่มาเพราะ.....	
Date (dd/mm/yyyy) / /	
1. Time Blood Drawn (hr) . ☐ Unknown	
2. Specimen log number ☐ โครงการ - - ☐ นอกโครงการ.....	
3. CBC ☐ 1.No skip to Q.4 ☐ 2.Yes ☐ 3.Unknown ☐ 4.Not Available	
3.1. Hct (%) .	
3.2. WBC (X1,000) .	
3.3. Diff WBC (100%) ☐ 1.No ☐ 2.Yes ☐ 3.Unknown ☐ 4.Not Available	
1. PMN . 2. Band . 3. Lymphocyte . 4. Atyp lymphocyte .	
5. Monocyte . 6. Basophil . 7. Eosinophil . 8. Other .	
3.4 Platelet count (x1,000)	
3.5 Plt (/OF) ☐ Unknown ☐ <5 ☐ 5-10 ☐ 11-25 ☐ >25	
4.. LFT ☐ 1.No skip to Q.5 ☐ 2.Yes ☐ 3.Unknown ☐ 4.Not Available ☐ 5.Not Done	
4.1 Protein . 4.2 Albumin .	
4.3 AST 4.4 ALT	
4.5 Other.....	
Note for other lab result	
Study Nurse..... Date Check by nurse..... Date	

Form 1: Enroll: Clinical Database Management Project

Firstname Surname

Study number -

Version 4.8.11	Year of study	year [2][5][][]
	Hospital ☐ 1. Khon Kaen ☒ 2. Songkhla	hosp [2]
	Study group ☐ 1. Study ☐ 2. Routine	group [][]
	☐ 3. Control (non-dengue)	
	Patient number (XXXX)	id [][][][][]

Version 4.8.11	Patient Information	
	1. HN..... AN.....	hn [][][][][][][][]
	2. Sex ☐ 1. M ☐ 2. F	sex [][]
	3. Birth date:/...../ 25..... (dd/mm/yyyy)	bdate [][][][][2][5][][]
	4. Place of birth..... (Province)	prov [][][]
	5. Religion ☐ 1. Buddhism ☐ 2. Christianity ☐ 3. Islam ☐ 4. Other	relig [][]
Version 4.8.11	6. Current address: House No.	
	Moo..... Soi..... Road.....	house no [][][][][][][][]
	District..... Amphur.....	zipcode [][][][][][][]
	Province Zipcode	
	Tel Comment	

Map (draw map of address)

Version 4.8.11

Form 1: Enroll: Clinical Database Management Project

Firstname Surname

Study number -

Version 4.8.11	Year of study	year [2][5][][]
	Hospital ☒ 1. Khon Kaen ☐ 2. Songkhla	hosp [1]
	Study group ☐ 1. Study ☐ 2. Routine	group [][]
	☐ 3. Control (non-dengue)	
	Patient number (XXXX)	id [][][][]

Version 4.8.11	Patient Information	
	1. HN..... AN.....	hn [][][][][][][][]
	2. Sex ☐ 1. M ☐ 2. F	sex [][]
	3. Birth date:/...../ 25..... (dd/mm/yyyy)	bdate [][][][][2][5][][]
	4. Place of birth..... (Province)	prov [][][]
	5. Religion ☐ 1. Buddhism ☐ 2. Christianity ☐ 3. Islam ☐ 4. Other	relig [][]
	6. Current address: House No.	housetno
	Moo..... Soi..... Road.....	zipcode [][][][][][]
	District..... Amphur.....	
	Province Zipcode	
Tel Comment		
Map (draw map of address)		

Version 4.8.11

Form 1: Enroll: Clinical Database Management Project

Firstname Surname

Study number -

7. Guardian Information Name..... Age yrs Relationship ☐ 1. Father ☐ 4. Grandmother ☐ 2. Mother ☐ 5. Relative (specify.....) ☐ 3. Grandfather ☐ 6. Baby sitter Address for contact..... Tel.		age [][] grelat []
8. Ever received JE vaccine? (if N/U skip to Q.9) ☐ 1. Yes ☐ 2. No ☐ 3. Unknown if yes, specify source and injection(s): Source: ☐ 1. Health book ☐ 3. ตามจากญาติ ☐ 2. OPD ☐ 4. Unknown 1st injection ☐ 1. Yes, when 25..... ☐ 2. No ☐ 3. Unk 2nd injection ☐ 1. Yes, when 25..... ☐ 2. No ☐ 3. Unk 3rd injection ☐ 1. Yes, when 25..... ☐ 2. No ☐ 3. Unk		je [] jes [] je1 [] je1year 25 [][] je2 [] je2year 25 [][] je3 [] je3year 25 [][]
9. Is BCG scar visible? (if N skip to Q.10) ☐ 1. Yes ☐ 2. No ☐ 3. Unknown if yes, specify source and injection(s): Source: ☐ 1. Health book ☐ 3. ตามจากญาติ ☐ 2. OPD ☐ 4. Unknown 1st injection ☐ 1. Yes, when 25..... ☐ 2. No ☐ 3. Unk 2nd injection ☐ 1. Yes, when 25..... ☐ 2. No ☐ 3. Unk		bcg [] bcs [] bc1 [] bc1year 25 [][] bc2 [] bc2year 25 [][]
10. Height (cm)		Ht [][][]

Form 1: Enroll: Clinical Database Management Project

Firstname Surname

Study number

Enrollment Information	
Admission Date/...../25..... (dd/mm/yyyy)	Addate [][][][][2][5][][]
Enrollment Date/...../25..... (dd/mm/yyyy)	Endate [][][][][2][5][][]
Discharge Date/...../25..... (dd/mm/yyyy)	Dsdate [][][][][2][5][][]
Withdrawal Date/...../25..... (dd/mm/yyyy)	Wtdate [][][][][2][5][][]
Reason.....	
Clinical Dx at admission ☐ 1. AFI ☐ 4. R/O DF ☐ 2. DF ☐ 5. R/O DHF ☐ 3. DHF ☐ 6. Acute dengue infection ☐ 7. Other.....	dx [] otherdx _____
Admission Physician:	

CLINICAL HISTORY: (If age<4 years. Doubts in symptom = unknown)					
Symptoms before enrollment	1=Y	Duration*	2=N	3=U	* in days
Any other febrile illness during the past 2 months?	☐	ไม่ต้องเติม	☐	☐	Ill [][]
1. Fever	☐	ไม่ต้องเติม	☐	☐	s1 [][]
Date 1 st ill (onset of fever)/...../ 25..... (dd/mm/yyyy)			s1dt ____ / ____ /25 ____		
Time : (24 hour clock)			s1tm [][]:[][]		
2. Headache	☐		☐	☐	s2 [][] durs2 [][][][]
3. Myalgia	☐		☐	☐	s3 [][] durs3 [][][][]
4. Arthralgia	☐		☐	☐	s4 [][] durs4 [][][][]
5. Retro-orbital pain	☐		☐	☐	s5 [][] durs5 [][][][]
6. Flushed face	☐		☐	☐	s6 [][] durs6 [][][][]
7. Rash without itching	☐		☐	☐	s7 [][] durs7 [][][][]
8. Rash with itching	☐		☐	☐	s8 [][] durs8 [][][][]

Form 2:

Drugs Taken Before Admission

Firstname Surname

Study number -

Drug	วัน-เดือน-ปี / ปริมาณยาที่ได้รับ (mg)					รับยาจากที่ไหน (เลือกได้มากกว่า 1 ข้อ)
	D1	D2	D3	D4	D5	
1. Antipyretics : ยาลดไข้	B 1. Yes B 2. No B 3. Unknown					n
If yes, 1. Para						nn
4. ASA						nn
5. Brufen						nn
2. Antibiotics : ยาแก้อักเสบ	B 1. Yes B 2. No B 3. Unknown					n
If yes, 1. Amoxycillin, Ampicillin						nn
2. Sulfa (Bactrim)						nn
3. Macrolide						nn
4. Cephalosporin						nn
5. Penicillin						nn
3. ORS : การได้รับเกลือแร่	B 1. Yes B 2. No B 3. Unknown					n nn
4. GI drug : ยาระบบทางเดินอาหาร	B 1. Yes B 2. No B 3. Unknown					n
If yes, 1. Antiemetic	B 1. Yes	B 2. No	n			nn
2. Anta-acid	B 1. Yes	B 2. No	n			nn
5. Respiratory drug : ยาระบบทางเดินหายใจ	B 1. Yes B 2. No B 3. Unknown					n
If yes, 1. Anticough	B 1. Yes	B 2. No	n			nn
2. Antihistamine	B 1. Yes	B 2. No	n			nn
6. Injection drug : ยาฉีดและสารน้ำ	B 1. Yes B 2. No B 3. Unknown					n
If yes, 1. Antipyrexia	B 1. Yes	B 2. No	n			nn
2. Antibiotics	B 1. Yes	B 2. No	n			nn
3. Anticonvulsion	B 1. Yes	B 2. No	n			nn
4. Antiemetic	B 1. Yes	B 2. No	n			nn
5. IV fluid	B 1. Yes	B 2. No	n			nn

Study nurse..... Date/...../ 25.....	1st Data entry by Date/...../ 25.....
Check by nurse..... Date/...../ 25.....	2nd Data entry by..... Date/...../ 25.....
1st Verify by PIA Date/...../ 25.....	2nd Verify by Date/...../ 25.....
1st Audit by Date/...../ 25.....	2nd Audit by Date/...../ 25.....

Version 15/05/06

Form 3:

Re-check Pyrexia & Vital Signs

ขอตำแนฟอรมปรอทแนบด้วย

Firstname Surname

Study number

- -

[illegible]

	1 st Data entry by
Check by clerk.....	2 nd Data entry by.....
1 st Verify by PI assistant.....	2 nd Verify by.....
1 st Audit by PI.....	2 nd Audit by

Form 4/1: Admission:

Clinical Record

Firstname Surname

Study number

 -

Date (dd/mm/25yy)	ชื่อ					
Clinical Findings (morning of evaluation)		[Q2-8: 1 = Yes, 2 = No, 3 = Unknown]				
1. Mental Status	b1	[]	[]	[]	[]	[]
1 = Alert 2 = Drowsy 3 = Lethargic 4 = Restless or Agitated 5 = Stuporous (awakens only with vigorous stimulation) 6 = Comatose (won't wake up)						
2. Tourniquet test (Y/N/U if N/U skip to Q.3)	b2	[]	[]	[]	[]	[]
2.1 Quantity (1: 0-10 2: 11-20 3: >20)	b21	[]	[]	[]	[]	[]
2.2 Quality (1: fine, 2: medium, 3: coarse, 4: mixed, 5: unknown)	b22	[]	[]	[]	[]	[]
3. Rash (Y/N/U if N/U skip to Q.4)	b3	[]	[]	[]	[]	[]
3.1 Confluent petechial rash	b31	[]	[]	[]	[]	[]
3.2 Maculopapular rash	b32	[]	[]	[]	[]	[]
3.3 Convalescent rash	b33	[]	[]	[]	[]	[]
3.4 Other:	b34	[]	[]	[]	[]	[]
4. Itching related to rash (Y/N/U)	b4	[]	[]	[]	[]	[]
5. Bruising with venipuncture (Y/N/U)	b5	[]	[]	[]	[]	[]
6. Bleeding (Y/N/U if N/U skip to Q.7)	b6	[]	[]	[]	[]	[]
6.1 Spontaneous petechiae	b61	[]	[]	[]	[]	[]
6.2 Ecchymosis	b62	[]	[]	[]	[]	[]
6.3 Gum	b63	[]	[]	[]	[]	[]
6.4 Nose	b64	[]	[]	[]	[]	[]
6.5 Vomiting Blood	b65	[]	[]	[]	[]	[]
6.6 Blood in stool	b66	[]	[]	[]	[]	[]
6.7 Other	b67	[]	[]	[]	[]	[]
7. Diarrhea (1=Yes, 2=No, 3=Unknown)	b7	[]	[]	[]	[]	[]
8. URI (Y/N/U if N/U skip to Q.9)	b8	[]	[]	[]	[]	[]
8.1 Exudative rhinorrhea	b81	[]	[]	[]	[]	[]
8.2 Clear rhinorrhea	b82	[]	[]	[]	[]	[]
8.3 Non-productive cough	b83	[]	[]	[]	[]	[]
8.4 Productive cough	b84	[]	[]	[]	[]	[]
8.5 Feeling of sore throat	b85	[]	[]	[]	[]	[]

Form 4/1: Admission:

Clinical Record

Firstname Surname

Study number

 -

Date (dd/mm/25yy)	ข้อ					
9. Abdominal pain (Y/N/U)	b9	[]	[]	[]	[]	[]
10. Dyspnea (Y/N/U) *	b10	[]	[]	[]	[]	[]
11. Ascites (Y/N/U) *	b11	[]	[]	[]	[]	[]
12. Jaundice (Y/N/U) *	b12	[]	[]	[]	[]	[]
13. Liver tenderness (Y/N/U) *	b13	[]	[]	[]	[]	[]
14. Liver size (>RCM)/span (cm) *	b14					
15. Abdominal Circumference (cm) *	b15					
16. Spleen size (cm > LCM) *	b16					
17. Lymph node enlargement (Y/N/U) * (if N/U skip to Q.18)	b17	[]	[]	[]	[]	[]
17.1 Cervical (max. diam. cm)	b171					
17.2 Epitrochea (max. diam. cm)	b172					
17.3 Other: _____ (max. diam. cm)	b173					
18. Injected conjunctivae (Y/N/U) *	b18	[]	[]	[]	[]	[]
18.1 Limbus (Y/N/U) *	b181	[]	[]	[]	[]	[]
18.2 Bulbar (Y/N/U) *	b182	[]	[]	[]	[]	[]
[Q. 9-13, 17-18 1 = Yes, 2 = No, 3 = Unknown, 4 = Not done, 5 = Not available]						
* must be verified by physician (Q. 10 -18)						

Form 4/1: Admission:

Clinical Record

Firstname Surname

Study number

 -

Date (dd/mm/25yy)	ชื่อ					
19. Weight (kg)	b19					
20. Total Oral Intake Estimate (ml)	b20					
21. Total IV Transfusion (ml)	b21					
21.1 Crystalline fluid Tx amount (ml)	b211					
21.2 Form crystalline fluid (Y/N/U) if N/U skip to Q. 21.3	b212	[]	[]	[]	[]	[]
21.2.1 5% D/NSS (5%DAR) - (ml)	b2121					
21.2.2 5% D/N/2 - (ml)	b2122					
21.2.6 0.9% NSS - (ml)	b2126					
21.2.7 Other	b2127					
21.3 Dextran (ml)	b213					
21.4 Haesteryl (ml)	b214					
21.5 Blood (Y/N/U if N/U skip to Q. 22)	b215	[]	[]	[]	[]	[]
21.5.1 Whole Blood (ml)	b2151					
21.5.2 Packed Red Cell (ml)	b2152					
21.5.3 Fresh Frozen Plasma (ml)	b2153					
21.5.4 Platelet (unit)	b2154					
21.5.5 Cryoprecipitate (ml)	b2155					
[Q. 21.2, 21.5 1 = Yes, 2 = No, 3 = Unknown, 4 = Not done, 5 = Not available]						
Note on unusual manifestation						
Study Nurse						
		1 st Data entry by				
Check by nurse		2 nd Data entry by.....				
1 st Verify by PI assistant		2 nd Verify by.....				
1 st Audit by PI		2 nd Audit by.....				

Form 4/2: Admission:

Routine Laboratory

Firstname Surname

Study number -

Date (dd/mm/25yy)	ชื่อ					
1. Time Blood Drawn (hr)	c1					
2. Specimen log number	c2					
3. CBC (Y/N if N skip to Q.4)	c3	[]	[]	[]	[]	[]
3.1 Hct (%)	c31					
3.2 WBC (x 1,000)	c32					
3.3 Diff WBC (100%)	c33	[]	[]	[]	[]	[]
3.3.1 PMN	c331	%	%	%	%	%
3.3.2 Band	c332	%	%	%	%	%
3.3.3 Lymphocyte	c333	%	%	%	%	%
3.3.4 Atyp lymphocyte	c334	%	%	%	%	%
3.3.5 Monocyte	c335	%	%	%	%	%
3.3.6 Basophil	c336	%	%	%	%	%
3.3.7 Eosinophil	c337	%	%	%	%	%
3.3.8 Other	c338	%	%	%	%	%
3.4 Platelet count (x 1,000)	c34	[]	[]	[]	[]	[]
3.5 Plt (/OF)	c35	[]	[]	[]	[]	[]
(0 = Unknown, 1 = <5, 2 = 5-10, 3 = 11-25, 4 = >25)						
4. LFT (Y/N if N skip to Q.5)	c4	[]	[]	[]	[]	[]
4.1 Protein	c41					
4.2 Albumin	c42					
4.3 Cholesterol	c43					
4.4 Total Bilirubin	c44					
4.5 Direct Bilirubin	c45					
4.6 AST	c46					
4.7 ALT	c47					
4.8 Alk phos	c48					
4.9 Other, []	c49	[]	[]	[]	[]	[]
Note for other lab result						
Q. 3, 3.3, 4, 4.9: 1 = Yes, 2 = No, 3 = Unknown, 4 = Not Available						

Form 4/2: Admission:

Routine Laboratory

Firstname Surname

Study number

 -

Date (dd/mm/25yy)	ชื่อ					
5. U/A (Y/N if N skip to Q.6)	c5	[]	[]	[]	[]	[]
5.1 Color ¹	c51	[]	[]	[]	[]	[]
5.2 Appearance ²	c52	[]	[]	[]	[]	[]
5.3 pH	c53					
5.4 Spec. gravity ³	c54	[]	[]	[]	[]	[]
5.5 Protein ⁴	c55	[]	[]	[]	[]	[]
5.6 Sugar ⁴	c56	[]	[]	[]	[]	[]
5.7 Other	c57					
¹ Color 0 = Not Done 1 = Colorless 2 = Pale yellow 3 = Yellow 4 = Brown 5 = Pale red 6 = Reddish 7 = Green 8 = Not Available 9 = Unknown ² Appearance 0 = Not Done 1 = Clear 2 = Slightly turbid 3 = Turbid 4 = Milky 8 = Not Available 9 = Unknown ³ Spec. gravity 0 = Not Done 1 = < 1.005 2 = 1.005 3 = 1.010 4 = 1.015 5 = 1.020 6 = 1.025 7 = 1.030 8 = >1.030 9 = Not Available / Unknown ⁴ Protein/sugar 0 = Not Done 1 = Negative 2 = Trace 3 = 1+ 4 = 2+ 5 = 3+ 6 = 4+ 8 = Not Available 9 = Unknown						
Microscopic exam. (0: 0-1, 1: 1-2, 2: 2-3, 3: 3-5, 4: 5-10, 5: 10-20, 6: 20-30, 7: 30-50, 8: 50-100, 9: TNTC)						
5.8 WBC (/HPF)	c58	[]	[]	[]	[]	[]
5.9 RBC (/HPF)	c59	[]	[]	[]	[]	[]
5.10 Other	c510	[]	[]	[]	[]	[]
6. Coagulogram (Y/N) (if N skip to Q7.)	c6	[]	[]	[]	[]	[]
6.1 PT (Y/N if N skip to Q6.2)	c61	[]	[]	[]	[]	[]
6.1.1 Test	c611					
6.1.2 Control	c612					
6.1.3 INR	c613					
6.2 PTT (Y/N if N skip to Q6.3)	c62	[]	[]	[]	[]	[]
6.2.1 Test	c621					
6.2.2 Control	c622					
6.3 TT (Y/N if N skip to Q7)	c63	[]	[]	[]	[]	[]
6.3.1 Test	c631					
6.3.2 Control	c632					
Q. 6, 6.1, 6.2, 6.3: 1 = Yes, 2 = No, 3 = Unknown, 4 = Not Available						

Form 4/2: Admission:

Routine Laboratory

Firstname Surname

Study number

 - -

Date (dd/mm/25yy)	ข้อ					
7. Right lateral decubitus film (Y/N/U, if N/U skip to Q8.) *	c7	[]	[]	[]	[]	[]
7.1 Effusion amount by upright (cm) *	c71					
7.2 Effusion amount by supine (cm) *	c72					
7.3 Effusion index (neg or digit) *	c73					
8. Right lateral decubitus Ultrasound (Y/N/U) *	c8	[]	[]	[]	[]	[]
8.1 Effusion amount by upright (cm) *	c81					
8.2 Effusion amount by supine (cm) *	c82					
8.3 Effusion index (neg or digit) *	c83					
Study Nurse						
Q. 7, 8: 1 = Yes, 2 = No, 3 = Unknown, 4 = Not Available						

* Must be verified by physician

	1 st Data entry by
Check by nurse	2 nd Data entry by

Form 5: Clinical Dx Vs Final Dx

Firstname Surname

Study number -

1. Outcome ☐ 1. Recovered ☐ 2. Disabled ☐ 3. Died	d1 []
2. Time of Subsidence of fever Date...../...../25..... (dd/mm/25yy) Time : ☐ Unknown	d2a [][][][2][5][][] d2b [][]:[][]
3. Leakage ☐ 1. Yes ☐ 2. No ☐ 3. Unknown 3.3 Evidence of leakage 3.3.1. Hemoconcentration ☐ 1. Yes ☐ 2. No 3.3.2. Pleural effusion ☐ 1. Yes ☐ 2. No 3.3.3. Ascites ☐ 1. Yes ☐ 2. No 3.3.4. Plasma alb ☐ 1. Yes ☐ 2. No	d3 [] d331 [] d332 [] d333 [] d334 []
4. Shock ☐ 1. Yes ☐ 2. No ☐ 3. Unknown if yes, 4.1 From: Date/...../25 Time : 4.2 To: Date/...../25 Time : 4.3 Vital sign evidence of shock 4.3.1 Low SBP ☐ 1. Yes ☐ 2. No 4.3.2 Pulse Pressure<20 ☐ 1. Yes ☐ 2. No 4.3.3 No Pressure ☐ 1. Yes ☐ 2. No 4.4 Clinical evidence of shock ☐ 1. Yes ☐ 2. No ☐ 3. Unknown 4.4.1. Mental status change ☐ 1. Yes ☐ 2. No 4.4.2. Cold extremities ☐ 1. Yes ☐ 2. No 4.4.3. Delay capillary refill > 2sec ☐ 1. Yes ☐ 2. No	d4 [] d41 [][][][][2][5][][] [][]:[][] d42 [][][][][2][5][][] [][]:[][] d431 [] d432 [] d433 [] d44 [] d441 [] d442 [] d443 []
5 Clinical Dx at Discharge ☐ 1. DF ☐ 4. DHF grade 3 ☐ 2. DHF grade 1 ☐ 5. DHF grade 4 ☐ 3. DHF grade 2 ☐ 6. Non dengue	d5 []
Any remarkable aspects on this patient?	
PI	
Check by clerk	

Form 5:

Clinical Dx Vs Final Dx

Firstname Surname

Study number

□ □ □ □ - □ □ □ □ - □

Dengue Laboratory Result	
6 PCR ☐ 1. Negative ☐ 5. Dengue 4 ☐ 2. Dengue 1 ☐ 6. Un-interpretable ☐ 3. Dengue 2 ☐ 7. Not available ☐ 4. Dengue 3 ☐ 8. Other	d6 []
7 ELISA ☐ 1. 1° infection ☐ 4. Un-interpretable ☐ 2. 2° infection ☐ 5. Not available ☐ 3. Non dengue ☐ 6. Other	d7 []
8 Virus Isolation ☐ 1. Negative ☐ 5. Dengue 4 ☐ 2. Dengue 1 ☐ 6. Un-interpretable ☐ 3. Dengue 2 ☐ 7. Not available ☐ 4. Dengue 3 ☐ 8. Other	d8 []
9 Final Diagnosis ☐ 1. DF ☐ 4. DHF grade 3 ☐ 2. DHF grade 1 ☐ 5. DHF grade 4 ☐ 3. DHF grade 2 ☐ 6. Non dengue, (specify)	d9 []
10 Clinical data sheet consistent with patient course? ☐ 1. Yes ☐ 2. No If no, explain the reason	d10 []

Pl assistant	1 st Data entry by
Pl audit.....	2 nd Data entry by.....
	2 nd Verify by.....
	2 nd Audit by.....



ตารางผนวกที่ ข1 ตารางความหมายของแอททริบิวต์ไฟล์เอกสาร D01-01

ชื่อแอททริบิวต์	ความหมาย
sno	เลขประจำตัวผู้ป่วย
age	อายุผู้ป่วยอยู่ในช่วง 1-15 ปีหรือไม่
fever_inclusion	ผู้ป่วยมีอุณหภูมิสูงกว่า 38.5 หรือไม่
duration_fever	ช่วงเวลาการมีไข้ของผู้ป่วยมากกว่า 5 วันหรือไม่
tt1	พบจุดเลือดเกินกว่า 10 จุดต่อ ตร.นิ้วหรือไม่
tt2	พบจุดเลือดเกินกว่า 5 จุดต่อ ตร.นิ้วหรือไม่
infection	ผู้ป่วยไม่มีไข้จากสาเหตุอื่นๆ ไข้หรือไม่
consent	ผู้ป่วยยินยอมให้ข้อมูลหรือไม่
wbc	จำนวนเม็ดเลือดขาวของผู้ป่วยวันที่ 4,5 น้อยกว่าหรือเท่ากับ 7000 หรือไม่
pcr_screen	PCR screening positive
outbreak	มีการระบาดในพื้นที่อยู่อาศัยหรือไม่
nurse	รายละเอียดของพยาบาลที่บันทึกข้อมูล

ตารางผนวกที่ ข2 ตารางความหมายของแอททริบิวต์เอกสาร D01-02

ชื่อแอททริบิวต์	ความหมาย
bdate	วันเกิดของผู้ป่วย
province	จังหวัดที่ผู้ป่วยเกิด : จังหวัดขอนแก่น จังหวัดสงขลา
provincex	จังหวัดที่ผู้ป่วยเกิด : จังหวัดอื่นๆ
religion	ศาสนา: ไทย คริสต์อิสลาม
religionx	ศาสนา: อื่นๆ
address	ที่อยู่ของผู้ป่วย
zipcode	รหัสไปรษณีย์
map	แผนที่
gi_name	ชื่อผู้ปกครอง
gi_age	อายุของผู้ปกครอง

ตารางผนวกที่ ข2 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
gi_relation	ความสัมพันธ์ระหว่างผู้ป่วย : พ่อ แม่ ปู่/ตา ย่า/ยาย พี่/น้อง
gi_relationx	ความสัมพันธ์ระหว่างผู้ป่วย : อื่นๆ
gi_address	ที่อยู่ของผู้ปกครอง
je	ผู้ป่วยได้รับวัคซีนโรคไข้มองอักเสบ (JE vaccine) หรือไม่
je_source	แหล่งข้อมูล : สมุดบันทึกการฉีดวัคซีน การตรวจ จากญาติ
je1st	ผู้ป่วยได้รับครั้งแรก
je1st_era	ผู้ป่วยได้รับครั้งแรกเมื่อไหร่
je2nd	ผู้ป่วยได้รับครั้งที่2
je2nd_era	ผู้ป่วยได้รับครั้งที่2 เมื่อไหร่
je3rd	ผู้ป่วยได้รับครั้งที่3
je3rd_era	ผู้ป่วยได้รับครั้งที่3 เมื่อไหร่
height	ส่วนสูงของผู้ป่วย
dengue_fever	ผู้ป่วยเคยเป็นไข้เลือดออกมาก่อนหรือไม่
admission	วันที่ผู้ป่วยรับเข้าโรงพยาบาล
enrollment	วันที่ผู้ป่วยลงทะเบียนเข้าร่วมโครงการ
discharge	วันที่ผู้ป่วยกลับบ้าน
accident	ผู้ป่วยเคยประสบอุบัติเหตุมาก่อนหรือไม่
accidentx	ถ้ามี ระบุรายละเอียด
withdraw	วันที่ผู้ป่วยถอนตัวจากโครงการ
withdrawx	เหตุผลที่ผู้ป่วยถอนตัว
dx_admission	การวินิจฉัยเบื้องต้นจากแพทย์ : AFI, DF, DHF, R/ODF, R/ODHF, Acute dengue infection
dx_admissionx	การวินิจฉัยเบื้องต้นจากแพทย์จากสาเหตุอื่น
febrile	ผู้ป่วยมีอาการไข้มองอักเสบใน 2 เดือนที่ผ่านมาหรือไม่
fever	ผู้ป่วยมีไข้หรือไม่
fever_date	วันที่ผู้ป่วยมีไข้วันแรก
fever_time	เวลาที่ผู้ป่วยมีไข้

ตารางผนวกที่ ข2 (ต่อ)

ชื่อแอทริบิวต์	ความหมาย
headache	ผู้ป่วยมีอาการปวดหัวหรือไม่
headache_day	วันที่ผู้ป่วยมีอาการ
myalgia	ผู้ป่วยมีอาการปวดเมื่อยกล้ามเนื้อ
myalgia_day	วันที่ผู้ป่วยมีอาการ
bone	ผู้ป่วยมีอาการเจ็บกระดูก
bone_day	วันที่ผู้ป่วยมีอาการ
retro	ผู้ป่วยมีอาการปวดหลัง
retro_day	วันที่ผู้ป่วยมีอาการ
flushed	ผู้ป่วยมีอาการหน้าแดง
flushed_day	วันที่ผู้ป่วยมีอาการ
rashwithout	ผู้ป่วยมีผื่นหรือไม่
rashwithout_day	วันที่ผู้ป่วยมีอาการ
rashwith	ผู้ป่วยมีอาการคันหรือไม่
rashwith_day	วันที่ผู้ป่วยมีอาการ
anorexia	ผู้ป่วยไม่มีอาการอยากอาหาร
anorexia_day	วันที่ผู้ป่วยมีอาการ
nausea	ผู้ป่วยมีอาการคลื่นไส้
nausea_day	วันที่ผู้ป่วยมีอาการ
vomiting	ผู้ป่วยมีอาการอาเจียนหรือไม่
vomiting_day	วันที่ผู้ป่วยมีอาการ
abdominal	ผู้ป่วยมีปวดท้องหรือไม่
abdominal_day	วันที่ผู้ป่วยมีอาการ
drowsiness	ผู้ป่วยมีเซ้าซี้หลับ
drowsiness_day	วันที่ผู้ป่วยมีอาการ
bleeding	ผู้ป่วยมีเลือดออกหรือไม่ (พร้อมระบุรายละเอียด)
spont	ผู้ป่วยมีจุดเลือดที่สามารถเห็นได้ด้วยตาเปล่า
spont_day	วันที่ผู้ป่วยมีอาการ

ตารางผนวกที่ ข2 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
ecchy	ผู้ป่วยมีจ้ำเลือดที่ผิวหนัง
ecchy_day	วันที่ผู้ป่วยมีอาการ
gum	ผู้ป่วยมีเลือดออกที่เหงือก
gum_day	วันที่ผู้ป่วยมีอาการ
nose	ผู้ป่วยมีเลือดออกที่จมูก
nose_day	วันที่ผู้ป่วยมีอาการ
vomit_bleed	ผู้ป่วยอาเจียนเป็นเลือดหรือไม่
vomit_bleed_day	วันที่ผู้ป่วยมีอาการ
stool	ผู้ป่วยอุจจาระเป็นเลือดหรือไม่
stool_day	วันที่ผู้ป่วยมีอาการ
other	อาการอื่นๆ
otherx	คำอธิบายอื่นๆ
other_day	วันที่มีผู้ป่วยอาการอื่นๆ
tourniquet	ผู้ป่วยรัดแขนแล้วเห็นจุดเลือดหรือไม่
tourniquet_date	วันที่ตรวจด้วยการรัดแขน
quantity	จำนวนจุดเลือดที่สังเกตจากการรัดแขน
quality	ขนาดของจุดเลือดที่สังเกตจากการรัดแขน
uri	ผู้ป่วยมีการตรวจระบบทางเดินหายใจ
exud	ผู้ป่วยมีน้ำมูกข้นหรือไม่
exud_day	วันที่ผู้ป่วยมีอาการ
clear	ผู้ป่วยมีน้ำมูกใสหรือไม่
clear_day	วันที่ผู้ป่วยมีอาการ
non_prod	ผู้ป่วยไอแบบไม่มีเสมหะ
non_prod_day	วันที่ผู้ป่วยมีอาการ
prod	ผู้ป่วยไอแบบมีเสมหะ
prod_day	วันที่ผู้ป่วยมีอาการ
feeling	ผู้ป่วยมีอาการเจ็บคอหรือไม่

ตารางผนวกที่ ข2 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
feeling_day	วันที่ผู้ป่วยมีอาการ
diarrhea	ผู้ป่วยมีอาการท้องเสียหรือไม่
diarrhea_day	วันที่ผู้ป่วยมีอาการ
other_sym	ผู้ป่วยมีอาการอื่นๆหรือไม่
other_symx	คำอธิบายอาการอื่นๆ
other_sym_day	วันที่ผู้ป่วยมีอาการ
nurse_fl	ข้อมูลพยาบาลที่กรอกข้อมูล

ตารางผนวกที่ ข3 ตารางความหมายของแอททริบิวต์เอกสาร D02

ชื่อแอททริบิวต์	ความหมาย
sno	เลขประจำตัวผู้ป่วย
date	วันเวลาที่เก็บข้อมูลผู้ป่วย
antipyretics	ผู้ป่วยทานยาลดไข้มาบ้างหรือไม่
para	ชนิดพาราเซตามอล (Paracetamol)
para_total	ปริมาณยาพาราเซตามอลที่ทาน
para_source1	ที่มาของยา1
para_source2	ที่มาของยา2
para_source3	ที่มาของยา3 อื่นๆ
para_source3x	คำอธิบายที่มาของยา3
asa	ยาแอสไพริน (ASA)
asa_total	ปริมาณยาแอสไพรินที่ทาน
asa_source1	ที่มาของยาแอสไพริน 1
asa_source2	ที่มาของยาแอสไพริน 2
asa_source3	ที่มาของยาแอสไพริน 3 อื่นๆ
asa_source3x	คำอธิบายที่มาของยาแอสไพริน 3
brufen	ยาแก้ปวด (Brufen)

ตารางผนวกที่ ข3 (ต่อ)

ชื่อแอทริบิวต์	ความหมาย
brufen_total	ปริมาณยาที่ทาน
brufen_source1	ที่มาของยา 1
brufen_source2	ที่มาของยา 2
brufen_source3	ที่มาของยา 3 อื่นๆ
brufen_source3x	คำอธิบายที่มาของยา 3
other	ยาอื่นๆ
otherx	คำอธิบายยาอื่นๆ
other_total	ปริมาณยาอื่นๆที่ได้รับ
other_source1	ที่มาของยาอื่นๆ 1
other_source2	ที่มาของยาอื่นๆ 2
other_source3	ที่มาของยาอื่นๆ 3
other_source3x	คำอธิบายที่มาของยาอื่นๆ 3
antibiotics	ยาแก้อักเสบ
amoxy	ชนิดยา Amoxy
amoxy_total	ปริมาณยา Amoxy
amoxy_source1	ที่มาของยา Amoxy อื่นๆ 1
amoxy_source2	ที่มาของยา Amoxy อื่นๆ 2
amoxy_source3	ที่มาของยา Amoxy อื่นๆ 3
amoxy_source3x	คำอธิบายที่มาของยา Amoxy อื่นๆ 3
sulfa	ชนิดยา Sulfa
sulfa_total	ปริมาณยา Sulfa
sulfa_source1	ที่มาของยา Sulfa อื่นๆ 1
sulfa_source2	ที่มาของยา Sulfa อื่นๆ 2
sulfa_source3	ที่มาของยา Sulfa อื่นๆ 3
sulfa_source3x	คำอธิบายที่มาของยา Sulfa อื่นๆ 3
macrolide	ชนิดยา Macrolide
macro_total	ปริมาณยา Macrolide

ตารางผนวกที่ ข3 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
macro_source1	ที่มาของยา Macrolideอื่นๆ 1
macro_source2	ที่มาของยา Macrolideอื่นๆ 2
macro_source3	ที่มาของยา Macrolideอื่นๆ 3
macro_source3x	คำอธิบายที่มาของยา Macrolideอื่นๆ 3
calosporin	ชนิดยา Calosporin
calo_total	ปริมาณยา Calosporin
calo_source1	ที่มาของยา Calosporinอื่นๆ 1
calo_source2	ที่มาของยา Calosporinอื่นๆ 2
calo_source3	ที่มาของยา Calosporinอื่นๆ 3
calo_source3x	คำอธิบายที่มาของยา Calosporinอื่นๆ 3
penicillin	ชนิดยา Penicillin
pen_total	ปริมาณยา Penicillin
pen_source1	ที่มาของยา Penicillinอื่นๆ 1
pen_source2	ที่มาของยา Penicillinอื่นๆ 2
pen_source3	ที่มาของยา Penicillinอื่นๆ 3
pen_source3x	คำอธิบายที่มาของยา Penicillinอื่นๆ 3
other_bio	ยาอีกเสบอื่นๆ
other_biox	คำอธิบายยาอีกเสบอื่นๆ
other_bio_total	ปริมาณยาอีกเสบอื่นๆที่ได้รับ
other_bio_source1	ที่มาของยาอีกเสบอื่นๆ 1
other_bio_source2	ที่มาของยาอีกเสบอื่นๆ 2
other_bio_source3	ที่มาของยาอีกเสบอื่นๆ 3
other_bio_source3x	คำอธิบายที่มาของยาอีกเสบอื่นๆ 3
ors	ผู้ป่วยได้รับเกลือแร่หรือไม่
gi_drug	ผู้ป่วยได้รับยาระบบทางเดินอาหารหรือไม่
antimetic	ชนิดยา Antimetic
antimetic_source1	ที่มาของยา Antimeticอื่นๆ 1

ตารางผนวกที่ ข3 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
antimetic_source2	ที่มาของยา Antimeticอื่นๆ 2
antimetic_source3	ที่มาของยา Antimeticอื่นๆ 3
antimetic_source3x	คำอธิบายที่มาของยา Antimeticอื่นๆ 3
anta	ชนิดยา Anta-acid
anta_source1	ที่มาของยา Anta-acidอื่นๆ 1
anta_source2	ที่มาของยา Anta-acidอื่นๆ 2
anta_source3	ที่มาของยา Anta-acidอื่นๆ 3
anta_source3x	คำอธิบายที่มาของยา Anta-acidอื่นๆ 3
respiratory	ผู้ป่วยได้รับยาระบบทางเดินหายใจหรือไม่
anticough	ชนิดยา Anticough
anticough_source1	ที่มาของยา Anticoughอื่นๆ 1
anticough_source2	ที่มาของยา Anticoughอื่นๆ 2
anticough_source3	ที่มาของยา Anticoughอื่นๆ 3
anticough_source3x	คำอธิบายที่มาของยา Anticoughอื่นๆ 3
antihis	ชนิดยา Antihistamine
antihis_source1	ที่มาของยา Antihistamineอื่นๆ 1
antihis_source2	ที่มาของยา Antihistamineอื่นๆ 2
antihis_source3	ที่มาของยา Antihistamineอื่นๆ 3
antihis_source3x	คำอธิบายที่มาของยา Antihistamineอื่นๆ 3
injection	ผู้ป่วยมีการใช้ยาฉีดและสารน้ำหรือไม่
antipy	ชนิดยา Anapyrexia
antipy_source1	ที่มาของยา Anapyrexiaอื่นๆ 1
antipy_source2	ที่มาของยา Anapyrexiaอื่นๆ 2
antipy_source3	ที่มาของยา Anapyrexiaอื่นๆ 3
antipy_source3x	คำอธิบายที่มาของยา Anapyrexiaอื่นๆ 3
antibio	ชนิดยา Antibiotics
antibio_source1	ที่มาของยา Antibioticsอื่นๆ 1

ตารางผนวกที่ ข3 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
antibio_source2	ที่มาของยา Antibioticsอื่นๆ 2
antibio_source3	ที่มาของยา Antibioticsอื่นๆ 3
antibio_source3x	คำอธิบายที่มาของยา Antibioticsอื่นๆ 3
anticon	ชนิดยา Anticonvulsant
anticon_source1	ที่มาของยา Anticonvulsantอื่นๆ 1
anticon_source2	ที่มาของยา Anticonvulsantอื่นๆ 2
anticon_source3	ที่มาของยา Anticonvulsantอื่นๆ 3
anticon_source3x	คำอธิบายที่มาของยา Anticonvulsantอื่นๆ 3
antieme	ชนิดยา Antiemetic
antieme_source1	ที่มาของยา Antiemeticอื่นๆ 1
antieme_source2	ที่มาของยา Antiemeticอื่นๆ 2
antieme_source3	ที่มาของยา Antiemeticอื่นๆ 3
antieme_source3x	คำอธิบายที่มาของยา Antiemeticอื่นๆ 3
iv	ชนิดยา IV fluid
iv_source1	ที่มาของยา IV fluidอื่นๆ 1
iv_source2	ที่มาของยา IV fluidอื่นๆ 2
iv_source3	ที่มาของยา IV fluidอื่นๆ 3
iv_source3x	คำอธิบายที่มาของยา IV fluidอื่นๆ 3
unk_drug	ไม่ทราบชนิดยา
unk_drugx	รายละเอียดชื่อยาที่ไม่ทราบ
nurse	ข้อมูลพยาบาลที่กรอกข้อมูล

ตารางผนวกที่ ข4 ตารางความหมายแอททริบิวต์ตามเอกสาร D03

ชื่อแอททริบิวต์	ความหมาย
sno	เลขประจำตัวผู้ป่วย
date	วันที่เก็บข้อมูล

ตารางผนวกที่ ข4 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
day_zero	day of defervescence
temp_max	อุณหภูมิมากที่สุดที่วัดได้ในวันที่เก็บข้อมูล
temp_min	อุณหภูมิน้อยที่สุดที่วัดได้ในวันที่เก็บข้อมูล
temp_nd	ไม่ได้วัดอุณหภูมิของผู้ป่วย
sbp	ค่าblood pressure ตัวแรก
dbp	ค่าblood pressure ตัวหลัง
bp_nd	ไม่สามารถวัดค่าได้
pulse_pre_min	ค่าความต่างของ blood pressure ตัวแรกลบตัวหลัง
pulse_pre_nd	ไม่สามารถวัดค่าได้
pulse_rate_max	ค่าอัตราการเต้นมากที่สุดในวันนั้นๆ
pulse_rate_min	ค่าอัตราการเต้นน้อยสุดในวันนั้นๆ
pulse_rate_nd	ไม่สามารถวัดค่าได้
hct_max	ค่าความเข้มข้นของเลือดมากที่สุด
hct_min	ค่าความเข้มข้นของเลือดน้อยที่สุด
hct_nd	ไม่ได้วัดค่าความเข้มข้นของเลือด
intake1	ปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 1
intake1_nd	ไม่ได้วัดปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 1
intake2	ปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 2
intake2_nd	ไม่ได้วัดปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 2
intake3	ปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 3
intake3_nd	ไม่ได้วัดปริมาณน้ำเกลือที่ให้ไปในครั้งที่ 3
output1	ปริมาณปัสสาวะที่ออกมาครั้งที่ 1
output1_nd	ไม่ได้วัดปริมาณปัสสาวะที่ออกมาครั้งที่ 1
output2	ปริมาณปัสสาวะที่ออกมาครั้งที่ 2
output2_nd	ไม่ได้วัดปริมาณปัสสาวะที่ออกมาครั้งที่ 2
output3	ปริมาณปัสสาวะที่ออกมาครั้งที่ 3
output3_nd	ไม่ได้วัดปริมาณปัสสาวะที่ออกมาครั้งที่ 3

ตารางผนวกที่ ข4 (ต่อ)

ชื่อแอทริบิวต์	ความหมาย
note	บันทึกเพิ่มเติม
nurse	ข้อมูลพยาบาลที่กรอกข้อมูล

ตารางผนวกที่ ข5 ตารางความหมายแอทริบิวต์ตามเอกสาร D04

ชื่อแอทริบิวต์	ความหมาย
sno	เลขประจำตัวผู้ป่วย
date	วันที่ในการบันทึกข้อมูล
mental	สภาพจิตของผู้ป่วย
tourniquet	ผู้ป่วยมีจุดเลือดหรือไม่
quantity	จำนวนจุดเลือด
quality	ขนาดของจุดเลือด
rash	ผื่น
conflu	ลักษณะผื่นขาว (Confluent Petechiae)
maculo	ลักษณะผื่นดำ (Maculopapular)
conval	ลักษณะผื่นกำลังจะหาย (Convalescent)
rash_otherx	ลักษณะผื่นอื่นๆ
itching	ผู้ป่วยมีอาการคันที่ผื่นหรือไม่
bruising	ผู้ป่วยเจาะเลือดแล้วมีอาการฟกช้ำไหม
bleeding	ผู้ป่วยมีเลือดออกหรือไม่
spont	เลือดออกที่เกิดขึ้นเองตามธรรมชาติ (Spontaneous petechiae)
ecchy	ผู้ป่วยมีจ้ำเลือดที่ผิวหนังหรือไม่
gum	ผู้ป่วยมีเลือดออกเหงือกหรือไม่
nose	ผู้ป่วยมีเลือดออกจมูกหรือไม่
vomiting	ผู้ป่วยอาเจียนเป็นเลือดหรือไม่
stool	ผู้ป่วยอุจจาระเป็นเลือดหรือไม่
bleeding_otherx	ผู้ป่วยมีเลือดออกที่ส่วนอื่นๆหรือไม่

ตารางผนวกที่ ข5 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
diarrhea	ผู้ป่วยมีอาการท้องร่วงหรือไม่
uri	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจหรือไม่
exud	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจแบบ Exudative
clear	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจแบบ Clear rhinestones
non_prod	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจแบบ Non-productive cough
prod	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจแบบ Productive cough
feeling	ผู้ป่วยมีอาการติดเชื้อทางระบบทางเดินหายใจแบบ Feeling of sore throat
abdominal	ผู้ป่วยมีอาการปวดท้องหรือไม่
dyspnea	ผู้ป่วยมีอาการหายใจติดขัดหรือไม่
ascites	พบอาการมีน้ำรั่วในช่องท้องหรือไม่
juandice	ผู้ป่วยมีอาการดีซ่านหรือตัวเหลืองหรือไม่
liver	ผู้ป่วยมีอาการเจ็บเมื่อมีการกดที่ตับ
liver_size	ขนาดของตับที่โตจากปกติถ้ามีค่าแสดงว่ามีอาการตับโต
ab_circum	ขนาดรอบเอวของผู้ป่วย
spleen	ขนาดม้าม (Spleen size)
date2	วันที่บันทึกข้อมูล
lymph	ผู้ป่วยมีอาการต่อมน้ำเหลืองโตหรือไม่
cervical	ตำแหน่งที่คอ
epitro	ตำแหน่งที่ไหปลาร้า
inguinal	ตำแหน่งที่ขาหนีบ
injected	ผู้ป่วยมีอาการตาอักเสบหรือไม่
limbus	ตำแหน่ง limbus ในตา
bilbar	ตำแหน่ง bilbar ในตา
weight	น้ำหนักของผู้ป่วย
estimate	ปริมาณน้ำที่ผู้ป่วยดื่มเข้าไป

ตารางผนวกที่ ข5 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
transfusion	ปริมาณน้ำเกลือที่ผู้ป่วยได้รับ
crystalline	รายละเอียดทางกายภาพของน้ำเกลือ Crystalline fluid Tx amount
form_crystalline	รายละเอียดทางกายภาพของน้ำเกลือ Form crystalline fluid
form1	รายละเอียดทางกายภาพของน้ำเกลือ 5% D/NSS (5% DAR)
form2	รายละเอียดทางกายภาพของน้ำเกลือ 5% D/N/2
form3	รายละเอียดทางกายภาพของน้ำเกลือ 0.9% NSS
form4	รายละเอียดทางกายภาพของน้ำเกลือ Other
form4x	รายละเอียดทางกายภาพของน้ำเกลือรายละเอียด Other
dextran	รายละเอียดทางกายภาพของน้ำเกลือ Dextran
haesteryl	รายละเอียดทางกายภาพของน้ำเกลือ Haesteryl
blood	ผู้ป่วยได้รับการตรวจเลือด
whole_blood	ปริมาณเลือดทั้งหมด
red_cell	ปริมาณเม็ดเลือดแดง
plasma	ปริมาณน้ำเลือด
platelet	ปริมาณเกล็ดเลือด
cryopre	ปริมาณพลาสมาที่ตกตะกอนด้วยความเย็น (Cryoprecipitate)
note	บันทึกเพิ่มเติม
nurse	ข้อมูลพยาบาลที่กรอกข้อมูล

ตารางผนวกที่ ข6 ตารางความหมายแอททริบิวต์ตามเอกสาร D05

ชื่อแอททริบิวต์	ความหมาย
sno	เลขประจำตัวผู้ป่วย
date	วันที่บันทึกข้อมูล
time_blood	เวลาที่บันทึกการเจาะเลือด
tb_unk	ไม่ทราบเวลาที่เจาะเลือด
specimen	หมายเลขของข้อมูลเลือด

ตารางผนวกที่ ข6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
s_pro	ผู้ป่วยเข้าร่วมโครงการหรือไม่
s_nonpro	ผู้ป่วยไม่ได้เข้าร่วมโครงการ
cbc	ผู้ป่วยตรวจเม็ดเลือดหรือไม่
hct	ปริมาณเม็ดเลือดแดง (%)
wbc	ปริมาณเม็ดเลือดขาว (x1000 เซลล์)
diff_wbc	แยกชนิดเม็ดเลือดขาวหรือไม่
pmn	เม็ดเลือดขาวชนิด PMN
band	เม็ดเลือดขาวชนิด Band
lymp	เม็ดเลือดขาวชนิด Lymphocyte
atyp	เม็ดเลือดขาวชนิด Atypical Lymphocyte
mono	เม็ดเลือดขาวชนิด Monocyte
baso	เม็ดเลือดขาวชนิด Basophil
eosin	เม็ดเลือดขาวชนิด Eosinophil
wbc_ot	เม็ดเลือดขาวชนิดอื่นๆ
platelet	ปริมาณเม็ดเกล็ดเลือด (x1000) ที่วัดได้จากเครื่อง
plt	ปริมาณเม็ดเกล็ดเลือด (x1000) ที่วัดโดยไม่ใช่เครื่อง
lft	ผู้ป่วยมีการตรวจผลเคมีจากตับหรือไม่
lft_protein	ค่าโปรตีนในตับ (4.1)
lft_albumin	ค่าalbumin (4.2)
lft_ast	ค่าเอนไซม์ Aspartate Aminotransferase (AST)
lft_alt	ค่าเอนไซม์ Alanine Aminotransferase (ALT)
lft_x	ค่าอื่นๆที่วัดได้จากตับ
note_x	บันทึกเพิ่มเติม
ua	ผู้ป่วยมีการตรวจปัสสาวะหรือไม่
ua_color	สีของปัสสาวะ
appear	ลักษณะที่ปรากฏของปัสสาวะ
ph	ค่า ph ของปัสสาวะ

ตารางผนวกที่ ข6 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
spec_g	ค่าความถ่วงจำเพาะของปัสสาวะ
ua_protein	ค่าโปรตีนในปัสสาวะ
ua_sugar	ค่าน้ำตาลในปัสสาวะ
ua_blood	ค่าเลือดในปัสสาวะ
ua_wbc	ค่าเม็ดเลือดขาวในปัสสาวะ
ua_rbc	ค่าเม็ดเลือดแดงในปัสสาวะ
date2	วันที่บันทึกข้อมูล
coagul	อัตราเร็วการแข็งตัวของเลือดของผู้ป่วย
pt	มีการตรวจ Prothrombin Time (PT) หรือไม่
pt_test	ค่าผลตรวจ
pt_control	ค่า PT control
pt_inr	ค่า International Normalized Ratio (INR)
ptt	มีการตรวจ PTT
ptt_test	ค่าผลตรวจ
ptt_control	ค่า PTT control
tt	มีการตรวจ TT
tt_test	ค่าผลตรวจ
tt_control	ค่า TT control
decubitusfilm	ผู้ป่วย X-rays หรือไม่
eff_1	ผลการทำ X-rays มีค่าเป็น negative (ปกติ)
eff_2	ค่าผลการตรวจถ้าพบว่ามีอาการน้ำรั่วในช่องท้องเกิดขึ้น
decubitus_ultrasound	ผู้ป่วย Ultrasound หรือไม่
eff_amount	ถ้าได้ทำมีผลที่วัดได้จากการทำ ultrasound เท่าไหร่
note2	บันทึกเพิ่มเติม
nurse	ข้อมูลพยาบาลที่กรอกข้อมูล

ตารางผนวกที่ ข7 ตารางความหมายแอททริบิวต์ตามเอกสาร D06

ชื่อแอททริบิวต์	ความหมาย
barcode	หมายเลขที่อ่านได้จากเอกสาร
sno	เลขประจำตัวผู้ป่วย
outcome	ผลลัพธ์การรักษา:ฟื้นตัว ไม่ฟื้นตัว ตาย
timefever_date	วันที่ไข้ลดลง
tfdate_unk	ไม่ทราบวันที่ไข้ลดลง
timefever_time	เวลาที่ไข้ลดลง
tftime_unk	ไม่ทราบเวลาที่ไข้ลดลง
leakage	สรุปอาการรั่วจากการตรวจ 4 อย่าง
hemo	ดูจากค่า Hemoconcentration
ascites	ดูจากค่า Ascites อาการน้ำรั่วในช่องท้อง
pleural	ดูจากผลการทำ X-ray และ Ultrasound
plasma	ดูค่าการรั่วของ plasma
shock	มีอาการช็อกหรือไม่
from_date	ถ้าใช้ระบุวันแรกที่มีอาการ
from_unk	ไม่รู้วันที่มีอาการช็อก
from_time	ถ้าใช้ ระบุเวลา
from_tunk	ไม่รู้เวลาที่มีอาการช็อก
to_date	ถ้าใช้ระบุวันสุดท้ายวันที่มีอาการ
to_unk	ไม่รู้วันที่มีอาการช็อก
to_time	ถ้าใช้ ระบุเวลา
to_tunk	ไม่รู้เวลาที่มีอาการช็อก
lowsbp	ค่าความดันค่าสูง sbp มีค่าต่ำเมื่อมีอาการช็อก
pulse	มีแรงบีบตัวของหัวใจน้อยกว่า 20 หรือไม่เมื่อมีอาการช็อก
nopressure	ไม่สามารถวัดค่าความดันคือ น้อยจนไม่สามารถวัดได้เมื่อมีอาการช็อก
clinical	ลักษณะทางกายภาพเมื่อมีอาการช็อก
mental	ภาวะทางจิตเมื่อมีอาการช็อก

ตารางผนวกที่ ข7 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
cold	ตัวเย็นเมื่อมีอาการช็อก
delay	เลือดไม่กลับเมื่อมีอาการช็อกทดลองโดยบีบเล็บแล้วเลือดไม่กลับ
dx	ผลการวินิจฉัย
remark	บันทึกข้อความรายละเอียดอื่นๆเพิ่มเติม
clinical_data	ข้อมูลทางกายภาพ
clinicalx	รายละเอียดทางกายภาพอื่นๆ
pi	ชื่อแพทย์ที่วินิจฉัย

ตารางผนวกที่ ข8 ตารางความหมายแอททริบิวต์ตามเอกสาร D07

ชื่อแอททริบิวต์	ความหมาย
barcode	หมายเลขที่อ่านได้จากเอกสาร
sno	เลขประจำตัวผู้ป่วย
followup	ผู้ป่วยให้ข้อมูลหลังจากรักษา
followupx	ถ้าไม่มาเหตุผลเพราะอะไร
date	วันที่ผู้ป่วยเดินทางกลับมาให้ข้อมูล
time_blood	เวลาที่ตรวจเลือด
time_blood_unk	ไม่ทราบเวลาตรวจเลือด
specimen	ข้อมูลเข้าร่วมโครงการ
s_proj	หมายเลขโครงการ
s_nonproj	ผู้ป่วยที่ไม่ได้เข้าร่วมโครงการ
cbc	ผู้ป่วยตรวจเม็ดเลือดหรือไม่
hct	ปริมาณเม็ดเลือดแดง (%)
wbc	ปริมาณเม็ดเลือดขาว (x1000 เซลล์)
diff_wbc	แยกชนิดเม็ดเลือดขาวหรือไม่
pmn	เม็ดเลือดขาวชนิด PMN
band	เม็ดเลือดขาวชนิด Band

ตารางผนวกที่ ข8 (ต่อ)

ชื่อแอททริบิวต์	ความหมาย
lymp	เม็ดเลือดขาวชนิด Lymphocyte
atyp	เม็ดเลือดขาวชนิด Atypical Lymphocyte
mono	เม็ดเลือดขาวชนิด Monocyte
baso	เม็ดเลือดขาวชนิด Basophil
eosin	เม็ดเลือดขาวชนิด Eosinophil
wbc_ot	เม็ดเลือดขาวชนิดอื่นๆ
platelet	ปริมาณเม็ดเกล็ดเลือด (x1000) ที่วัดได้จากเครื่อง
plt	ปริมาณเม็ดเกล็ดเลือด (x1000) ที่วัดโดยไม่ใช่เครื่อง
lft	ผู้ป่วยมีการตรวจผลเคมีจากตับหรือไม่
protein	ค่าโปรตีนในตับ (4.1)
albumin	ค่าalbumin (4.2)
ast	ค่าเอนไซม์ Aspartate Aminotransferase (AST)
alt	ค่าเอนไซม์ Alanine Aminotransferase (ALT)
otherx	ค่าอื่นๆที่วัดได้จากตับ
note	บันทึกเพิ่มเติม
nurse	ข้อมูลพยาบาลที่กรอกข้อมูล



ตารางผนวกที่ ค1 ตารางตัวอย่างผลการหาความสัมพันธ์ जोทย์ที่ 5

คลาส DF

Best rules found:

1. dx=DHF 349 ==> AST_avg=H 349 conf:(1)
2. AST_avg=H 349 ==> dx=DHF 349 conf:(1)
3. leakage=Yes 346 ==> AST_avg=H 346 conf:(1)
4. leakage=Yes 346 ==> dx=DHF 346 conf:(1)
5. leakage=Yes dx=DHF 346 ==> AST_avg=H 346 conf:(1)
6. AST_avg=H leakage=Yes 346 ==> dx=DHF 346 conf:(1)
7. leakage=Yes 346 ==> AST_avg=H dx=DHF 346 conf:(1)
8. juandice_found=No 345 ==> AST_avg=H 345 conf:(1)
9. juandice_found=No 345 ==> dx=DHF 345 conf:(1)
10. juandice_found=No dx=DHF 345 ==> AST_avg=H 345 conf:(1)
11. juandice_found=No AST_avg=H 345 ==> dx=DHF 345 conf:(1)
12. juandice_found=No 345 ==> AST_avg=H dx=DHF 345 conf:(1)
13. atyp_found=Yes 343 ==> AST_avg=H 343 conf:(1)
14. atyp_found=Yes 343 ==> dx=DHF 343 conf:(1)
15. atyp_found=Yes dx=DHF 343 ==> AST_avg=H 343 conf:(1)
16. atyp_found=Yes AST_avg=H 343 ==> dx=DHF 343 conf:(1)
17. atyp_found=Yes 343 ==> AST_avg=H dx=DHF 343 conf:(1)
18. juandice_found=No leakage=Yes 342 ==> AST_avg=H 342 conf:(1)
19. juandice_found=No leakage=Yes 342 ==> dx=DHF 342 conf:(1)
20. juandice_found=No leakage=Yes dx=DHF 342 ==> AST_avg=H 342 conf:(1)

คลาส DHF

Best rules found:

1. dx=DHF 349 ==> AST_avg=H 349 conf:(1)
2. AST_avg=H 349 ==> dx=DHF 349 conf:(1)
3. leakage=Yes 346 ==> AST_avg=H 346 conf:(1)

4. leakage=Yes 346 ==> dx=DHF 346 conf:(1)
5. leakage=Yes dx=DHF 346 ==> AST_avg=H 346 conf:(1)
6. AST_avg=H leakage=Yes 346 ==> dx=DHF 346 conf:(1)
7. leakage=Yes 346 ==> AST_avg=H dx=DHF 346 conf:(1)
8. juandice_found=No 345 ==> AST_avg=H 345 conf:(1)
9. juandice_found=No 345 ==> dx=DHF 345 conf:(1)
10. juandice_found=No dx=DHF 345 ==> AST_avg=H 345 conf:(1)
11. juandice_found=No AST_avg=H 345 ==> dx=DHF 345 conf:(1)
12. juandice_found=No 345 ==> AST_avg=H dx=DHF 345 conf:(1)
13. atyp_found=Yes 343 ==> AST_avg=H 343 conf:(1)
14. atyp_found=Yes 343 ==> dx=DHF 343 conf:(1)
15. atyp_found=Yes dx=DHF 343 ==> AST_avg=H 343 conf:(1)
16. atyp_found=Yes AST_avg=H 343 ==> dx=DHF 343 conf:(1)
17. atyp_found=Yes 343 ==> AST_avg=H dx=DHF 343 conf:(1)
18. juandice_found=No leakage=Yes 342 ==> AST_avg=H 342 conf:(1)
19. juandice_found=No leakage=Yes 342 ==> dx=DHF 342 conf:(1)
20. juandice_found=No leakage=Yes dx=DHF 342 ==> AST_avg=H 342 conf:(1)

ตารางผนวกที่ ค2 ตารางตัวอย่างผลการหาความสัมพันธ์ जोทย์ที่ 7

ผลัด DF

Best rules found:

1. juandice_found=No 486 ==> dx=DF 486 conf:(1)
2. dyspnea_found=No 484 ==> dx=DF 484 conf:(1)
3. Effusion_Result=No 484 ==> dx=DF 484 conf:(1)
4. shock=No 484 ==> dx=DF 484 conf:(1)
5. dyspnea_found=No juandice_found=No 483 ==> dx=DF 483 conf:(1)
6. juandice_found=No Effusion_Result=No 483 ==> dx=DF 483 conf:(1)
7. juandice_found=No shock=No 483 ==> dx=DF 483 conf:(1)
8. ascites_found=No 482 ==> dx=DF 482 conf:(1)

9. dyspnea_found=No Effusion_Result=No 481 ==> dx=DF 481 conf:(1)
10. dyspnea_found=No shock=No 481 ==> dx=DF 481 conf:(1)
11. ascites_found=No juandice_found=No 481 ==> dx=DF 481 conf:(1)
12. Effusion_Result=No shock=No 481 ==> dx=DF 481 conf:(1)
13. dyspnea_found=No ascites_found=No 480 ==> dx=DF 480 conf:(1)
14. dyspnea_found=No juandice_found=No Effusion_Result=No 480 ==> dx=DF 480 conf:(1)
15. dyspnea_found=No juandice_found=No shock=No 480 ==> dx=DF 480 conf:(1)
16. juandice_found=No Effusion_Result=No shock=No 480 ==> dx=DF 480 conf:(1)
17. ascites_found=No Effusion_Result=No 479 ==> dx=DF 479 conf:(1)
18. ascites_found=No shock=No 479 ==> dx=DF 479 conf:(1)
19. dyspnea_found=No ascites_found=No juandice_found=No 479 ==> dx=DF 479 conf:(1)
20. dyspnea_found=No Effusion_Result=No shock=No 478 ==> dx=DF 478 conf:(1)

คลาส DHF

Best rules found:

1. leakage=Yes 510 ==> dx=DHF 510 conf:(1)
2. AST_avg=H 509 ==> dx=DHF 509 conf:(1)
3. juandice_found=No 507 ==> dx=DHF 507 conf:(1)
4. AST_avg=H leakage=Yes 506 ==> dx=DHF 506 conf:(1)
5. juandice_found=No leakage=Yes 504 ==> dx=DHF 504 conf:(1)
6. atyp_found=Yes 503 ==> dx=DHF 503 conf:(1)
7. juandice_found=No AST_avg=H 503 ==> dx=DHF 503 conf:(1)
8. AST_ALT=Yes 501 ==> dx=DHF 501 conf:(1)
9. atyp_found=Yes leakage=Yes 500 ==> dx=DHF 500 conf:(1)
10. AST_avg=H AST_ALT=Yes 500 ==> dx=DHF 500 conf:(1)
11. juandice_found=No AST_avg=H leakage=Yes 500 ==> dx=DHF 500 conf:(1)
12. atyp_found=Yes AST_avg=H 499 ==> dx=DHF 499 conf:(1)
13. juandice_found=No atyp_found=Yes 498 ==> dx=DHF 498 conf:(1)
14. AST_ALT=Yes leakage=Yes 498 ==> dx=DHF 498 conf:(1)
15. AST_avg=H AST_ALT=Yes leakage=Yes 497 ==> dx=DHF 497 conf:(1)

16. atyp_found=Yes AST_avg=H leakage=Yes 496 ==> dx=DHF 496 conf:(1)
17. juandice_found=No AST_ALT=Yes 495 ==> dx=DHF 495 conf:(1)
18. juandice_found=No atyp_found=Yes leakage=Yes 495 ==> dx=DHF 495 conf:(1)
19. juandice_found=No atyp_found=Yes AST_avg=H 494 ==> dx=DHF 494 conf:(1)
20. juandice_found=No AST_avg=H AST_ALT=Yes 494 ==> dx=DHF 494 conf:(1)



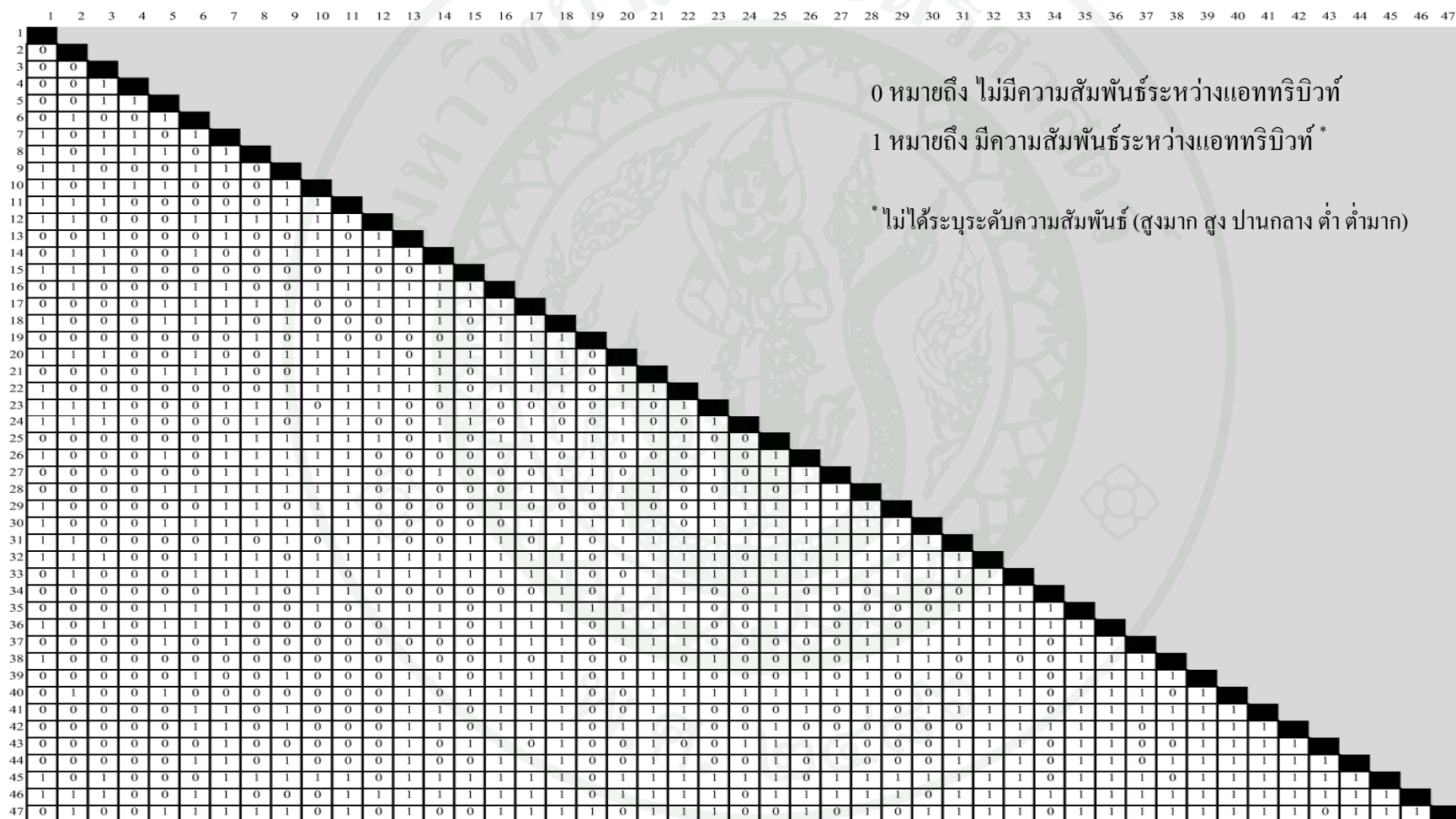


ตารางผนวกที่ ๑1 ตารางความสัมพันธ์ของอาการวิบัติที่เทียบกับคลาสข้อมูลที่ไม่ปรากฏบนต้นไม้

ชื่ออาการวิบัติ	ความหมาย	ค่า r	ความสัมพันธ์	ระดับ
leakage	อาการรั่วของพลาสมาในเลือด	0.8456	มี	สูง
shock	อาการช็อก	0.5377	มี	ปานกลาง
Bleeding	อาการเลือดออก	0.3719	มี	ต่ำ
platelet_min	ปริมาณเกล็ดเลือดน้อยสุด	-0.5853	มี	ปานกลาง
ALT_avg	ค่า ALT โดยเฉลี่ย	0.1941	มี	ต่ำมาก
white_blood_cell_avg	ปริมาณเม็ดเลือดขาวโดยเฉลี่ย	0.0842	มี	ต่ำมาก
lymph_found	อาการต่อมน้ำเหลืองโต	-0.0905	มี	ต่ำมาก
white_blood_cell_min	ปริมาณเม็ดเลือดขาวน้อยสุด	-0.0785	มี	ต่ำมาก
je	วัคซีนไข้มองอักเสบ	-0.0677	มี	ต่ำมาก
hematocrit_max	ปริมาณเม็ดเลือดแดงมากที่สุด	0.3267	มี	ต่ำ
hematocrit_max_d05	ปริมาณเม็ดเลือดแดงมากที่สุด	0.2642	มี	ต่ำมาก
AST_max	ค่า AST มากสุด	0.2393	มี	ต่ำมาก
liver_size_avg	ขนาดตับที่โตขึ้น	0.4177	มี	ต่ำ
temperature_max	อุณหภูมิมากที่สุด	0.0649	มี	ต่ำมาก

ตารางผนวกที่ ๑2 ตารางความสัมพันธ์ของอาการวิบัติที่เทียบกับคลาสข้อมูลที่ไม่ปรากฏบนต้นไม้

ชื่ออาการวิบัติ	ความหมาย	ค่า r	ความสัมพันธ์	ระดับ
ascites_found	อาการรั่วในช่องท้อง	0.5024	มี	ปานกลาง
abdominal_pain	อาการปวดท้อง	0.3592	มี	ต่ำ
dyspnea_found	อาการหายใจติดขัด	0.3765	มี	ต่ำ
protein_avg	ค่าโปรตีนในตับ	-0.3847	มี	ต่ำ
lft_albumin_avg	ค่า Albumin	-0.4701	มี	ต่ำ



ตารางผนวกที่ 3 ตารางอธิบายหมายเลขและความหมายของแอททริบิวต์ในภาพผนวกที่ 3

ลำดับ	ชื่อแอททริบิวต์	ความหมาย
1	je	รับวัคซีนโรคไข้มองอักเสบ
2	uri	อาการติดเชื้อทางเดินหายใจครั้งแรกที่เข้ามารับการรักษา
3	temperature_max	อุณหภูมิสูงสุด
4	temperature_min	อุณหภูมิน้อยสุด
5	sbp_minus_dbp_avg	ค่าผลต่างระหว่างค่าสูง-ต่ำ
6	pulse_pre_min_found	พบค่าความน้อยสุดมีผลต่างน้อยกว่า 20 มิลลิเมตรปรอท
7	hematocrit_max	ค่าความเข้มข้นของเลือดมากสุดในแต่ละวัน
8	hematocrit_min	ค่าความเข้มข้นของเลือดน้อยสุดในแต่ละวัน
9	quantity_max_found	ปริมาณจุดเลือดมากที่สุด (ค่าข้อมูล 3 ช่วงคือ 0-10, 11-20 และ มากกว่า 20)
10	rash_found	อาการผื่นขึ้น ไม่มีอาการคัน
11	itching_found	อาการคัน
12	bruising_found	อาการฟกช้ำหลังจากการเจาะเลือด
13	Bleeding	อาการเลือดออก
14	diarrhea_found	อาการท้องเสีย
15	uri_found	อาการติดเชื้อที่ระบบทางเดินหายใจช่วงที่รักษา
16	abdominal_pain	อาการปวดท้อง
17	dyspnea_found	อาการหายใจติดขัด
18	ascites_found	อาการมีน้ำรั่วในช่องท้องของผู้ป่วย
19	juandice_found	อาการดีซ่าน ตัวเหลือง
20	liver_tenderness	อาการเจ็บเมื่อกดที่ตับ
21	liver_size_avg	ขนาดของตับโต (ซม.)
22	liver_found	อาการตับโต
23	lymph_found	อาการต่อมน้ำเหลืองโต
24	injected_found	อาการตาอักเสบ
25	hematocrit_max_d05	ค่าความเข้มข้นของเลือดมากสุดในช่วงที่รับการรักษา
26	hematocrit_min_d05	ค่าความเข้มข้นของเลือดน้อยสุดในช่วงที่รับการรักษา

ตารางผนวกที่ 3 (ต่อ)

ลำดับ	ลำดับ	ลำดับ
27	hematocrit_avg_d05	ค่าความเข้มข้นของเลือดโดยเฉลี่ยในช่วงที่รับการรักษา
28	white_blood_cell_max	ค่าเม็ดเลือดขาว (x1000 cells) มากสุดในช่วงที่รับการรักษา
29	white_blood_cell_min	ค่าเม็ดเลือดขาว (x1000 cells) น้อยสุดในช่วงที่รับการรักษา
30	white_blood_cell_avg	ค่าเม็ดเลือดขาว (x1000 cells) โดยเฉลี่ยในช่วงที่รับการรักษา
31	platelet_max	ค่าปริมาณเกล็ดเลือด (x1000 cells) มากสุดในช่วงที่รับการรักษา
32	platelet_min	ค่าปริมาณเกล็ดเลือด (x1000 cells) น้อยสุดในช่วงที่รับการรักษา
33	platelet_avg	ค่าปริมาณเกล็ดเลือด (x1000 cells) โดยเฉลี่ยในช่วงที่รับการรักษา
34	atyp_found	พบความผิดปกติในเม็ดเลือดของผู้ป่วย
35	protein_avg	ค่าโปรตีนในเลือดเฉลี่ยในช่วงที่รับการรักษา
36	lft_albumin_avg	ค่า Albumin
37	Globurin_avg	ค่า Globurin
38	ratio_albumin_avg	ค่าสัดส่วนเฉลี่ยของ Albumin หารด้วย globurin
39	AST_max	ค่า AST มากสุด
40	AST_min	ค่า AST น้อยสุด
41	AST_avg	ค่า AST โดยเฉลี่ย
42	ALT_max	ค่า ALT มากสุด
43	ALT_min	ค่า ALT น้อยสุด
44	ALT_avg	ค่า ALT โดยเฉลี่ย
45	Effusion_Result	อาการรั่วของน้ำในปอด
46	leakage	อาการรั่วของพลาสมาในเลือด
47	shock	อาการช็อก



1. DATA MINING ON DENGUE VIRUS DISEASE, Proceeding of the 13th International Conference on Enterprise Information Systems (ICEIS2011), June 8-11, 2011. pp. 32-41 Beijing, China
2. Data Mining of Dengue Infection Using Decision Tree, The 12th WSEAS International Conference on APPLIED COMPUTER SCIENCE (ACS'12), May 11-13, 2012. pp. 154-159 Singapore city, Singapore
3. A Data Mining Framework for building Dengue infection disease Model, The 26th Annual Conference of Japanese Society for Artificial Intelligence (JSAI 2012), June 12-15, 2012 Reference Number 1K2-IOS-1b-7. pp. 1-8 Yamaguchi, Japan

DATA MINING ON DENGUE VIRUS DISEASE

Darane Thitiprayoonwongse^a, Prapat Suriyaphol^b, and Nuanwan Soonthomphisaj^{a,*}

^a*Department of Computer Science, Faculty of Science Kasetsart University, Bangkok, Thailand.
g521440162@ku.ac.th, fscinws@ku.ac.th*

^b*Bioinformatics and Data Management for Research Unit Office for Research and Development, Siriraj Hospital
Mahidol University, Bangkok, Thailand. sipur@mahidol.ac.th*

Abstract. Dengue infection is an epidemic disease typically found in tropical region. Symptoms of the disease show rapid and violent for patients in a short time. The World Health Organization (WHO) classifies the dengue infection as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF). Symptoms of DHF are divided into 4 types. The problem might be happen when an expert misdiagnoses dengue infection. For Example, an expert diagnosed a patient as non dengue or DF even if a patient was a DHF patient. That might be the cause of dead if patient did not receive treatment. Therefore, we selected data mining approach to solve this problem. We employed decision tree algorithm to learn from data set in order to create new knowledge. The first experimental result shows useful knowledge to classify dengue infection levels into 4 groups (DF, DHF I, DHF II, and DHF III). An average accuracy is 96.50 %. The second experimental result shows the tree and a set of rules to classify dengue infection levels into 2 groups followed by our assumption. An accuracy is 96.00 %. Furthermore, we compared our performance in term of false negative values to WHO and Tarig's criteria [7] and found that our research outperforms those criteria, as well.

Keywords: Data Mining, Decision Tree, Dengue virus Disease.

1 INTRODUCTION

Dengue Fever is an acute viral infection characterized by fever. It is caused by a bite from mosquitoes carrying dengue virus. The World Health Organization (WHO) classifies the dengue infection as DF and DHF. Symptoms of DF are rapidly fever, headache, myalgia, loss of appetite food, vomiting, abdominal pain and thrombocytopenia. The severity of DHF is divided into 4 types. First, DHF I is a DF patient who has fever and hemorrhagic appearance. Second, DHF II is a DHF I patient who has spontaneous bleeding. Next, DHF III is a DHF II patient who has sign of physiological failure such as rapid/weak pulse, narrow pulse pressure and cold/clammy skin. Lastly, DHF IV is a DHF III patient who shock and can't detect blood pressure or pulse [7]. The objectives of our research are as following (1). We would like to know a set of significant attributes that classify the type of dengue infections (2) Physician would like to know the criteria or patterns found in each class.

We selected decision tree learning as an approach to find knowledge in order to classify type of dengue infection. The total number of patients is 258 patients from Siriraj Hospital, Bangkok, Thailand. The data set consists of 128 DF, 65 DHF I, 52 DHF II and 13 DHF III (There is no patient who was diagnosed as DHF IV). We focus on patients whose ages are lower than 15 years old because the infection in children is more severe than adults. Forty-eight attributes are selected as a feature set for decision tree learning. These attributes are divided into 2 groups, which are categorical attributes and numerical attributes. The value of categorical attributes represented the evidence of symptom. Whereas the numerical attributes (integer and real number) are obtained from hematological evidence such as percentage of hematocrit increase (HCT), white blood cell (WBC), etc. The original data set is high dimension and has some missing values. Therefore, we need to preprocess data to clean up and clarify some error. We set up 2 experiments. The objective of the first

experiment is to find knowledge for each type of dengue infection. The second experiment explores the hypothesis to find the pattern of severe and non severe dengue patients.

2 CRITERIA OF DENGUE VIRUS DISEASE

Dengue is the most common virus transmitted by mosquitoes which causes up to 100 million infections and 25,000 deaths worldwide each year. WHO has launched criteria for diagnosing dengue patients.

2.1 WHO criteria

Table 1: WHO criteria [9]

DF/DHF	Grade	Symptoms	Laboratory
DF		Fever with two or more of the following signs: headache, retro-orbital pain, myalgia, arthralgia.	Leukopenia occasionally. Thrombocytopenia, may be present, no evidence of plasma loss.
DHF	I	Above signs plus positive tourniquet test	Thrombocytopenia < 100,000, HCT rise $\geq$ 20 %
DHF	II	Above signs plus spontaneous bleeding	Thrombocytopenia < 100,000, HCT rise $\geq$ 20 %
DHF	III	Above signs plus circulatory failure (weak pulse, hypotension, restlessness)	Thrombocytopenia < 100,000, HCT rise $\geq$ 20 %
DHF	IV	Profound shock with undetectable blood pressure and pulse.	Thrombocytopenia < 100,000, HCT rise $\geq$ 20 %

2.2 Tanner's criteria

Tanner, et al [3] employed decision tree to classify data into 4 levels which are Probable dengue, Likely dengue, Likely non-dengue and Probable non-dengue. Their data set contains 1,200 patients (1,012 patients from the EDEN study and 188 patients from Vietnam). They found 6 significant features which were platelet count (PLT), white blood cell count (WBC), body temperature (T), hematocrit (HCT), absolute number of lymphocytes (Lymphocyte) and absolute number of neutrophils (Neutrophil) . They got 84.7% correctness.

2.3 Tarig's Criteria

There were 3 research papers written by Tarig Faisal [6, 7, 8]. In March 2010, they predicted the risk of dengue patients using Self Organizing Map (SOM) and Multilayer Feed-forward Neural

WHO announced a set of criteria for classifying dengue patients according to DF and DHF (see Table 1 for details)

However, WHO criteria are not sufficient to classify the dengue patients. Since they are just a common criteria for dengue virus disease. We believe that there are some different clinical evidence and laboratory results that fit to our regional disease. There are some researchers work in this area such as Tanner, et al [3] and Tarig, et al. [6]. They tried to find new criteria in order to classify dengue patients.

Networks (MFNN). Nevertheless, their accuracy rate was only 70 %. Their next research was to do data clustering on patients into 2 groups as low risk and high risk patients. They classified 195 patients using three criteria obtained from SOM. There are 3 risk criteria which were platelet counts (PLT) (less or equal than 40,000 cell per mm³, HCT (greater than or equal to 25%) and aspartate aminotransferase (AST) (rose by fivefold the normal upper limit for AST or alanine aminotransferase) (ALT) (rose by fivefold, the normal upper limit for ALT)

A high risk patient was a patient who had at least 2 criteria. A low risk patient was a patient who had less than 2 criteria. Their finding supported the criteria of WHO. Lastly, in June 2010, they classified the risk of dengue patients using MLP. The accuracy only 75 %. Fatimah Ibrahim et al. [2] predicted the day of defervescence of fever (day0). Their data set consists of 252 dengue patients (4 DF

and 248 DHF). They applied Multi-Layer Perceptron (MLP) and got 90 % correctness.

3 DATA PROCESSING

Data integration is a step that integrated the data from several sources. In this study, Siriraj Hospital integrated patient's data from Srinagarindra Hospital and Songklanagarind Hospital. Next step is data cleaning. Sometimes the data sets contained noise data that results from human error or machine error. In case of missing value found in the data set, we will replace them with mean value. Feature selection is a step to exclude attributes that are not

important to improve the efficiency of experimental result. Data transformation is a step that transformed some attribute values in order to qualify the requirement of the algorithm. Feature extraction is an important step to pick up suitable attributes or create new feature set to represent some data pattern.

In this paper, we created new feature set as shown in Table 2 and transformed some numerical attributes to categorical attributes. During the treatment period, we observed the clinical information and hematological information. These attributes were extracted, as well.

Table 2: Feature extraction obtained from the treatment period

Attribute	Meaning
Bleeding	Evidence of Bleeding (Yes/No)
uri	Evidence of upper respiratory infection (Yes/No)
hematocrit_max	Maximum value of hematocrit concentration (%)
hematocrit_min	Minimum value of hematocrit concentration (%)
AST_max	Maximum value of AST (U/L)
AST_min	Minimum value of AST (U/L)
AST_avg	Average value of AST (U/L)
ALT_max	Maximum value of ALT (U/L)
ALT_min	Minimum value of ALT (U/L)
ALT_avg	Average value of ALT (U/L)
temp_max	Maximum temperature of patient (celsius)
temp_min	Minimum temperature of patient (celsius)
sbp_minus_dbp_avg	Average value of the difference between systolic blood pressure (sbp) and diastolic blood pressure (dbp) (mm.Hg)
liver_size_average	Average size of liver (cm)
hematocrit_max_dx	Maximum value of hematocrit concentration (%)
hematocrit_min_dx	Minimum value of hematocrit concentration (%)
hematocrit_avg_dx	Average value of hematocrit concentration (%)
white_blood_cell_max	Maximum number of white blood cells (x1000 cells/ μ l)
white_blood_cell_min	Minimum number of white blood cells (x1000 cells/ μ l)
white_blood_cell_avg	Average number of white blood cells (x1000 cells/ μ l)
platelet_max	Maximum of platelet count (x1000 cells/ μ l)
platelet_min	Minimum of platelet count (x1000 cells/ μ l)
platelet_avg	Average of platelet count (x1000 cells/ μ l)
protein_avg	Average value of protein in liver (g/dl)
albumin_avg	Average value of albumin (g/dl)
globulin_avg	Average value of globulin (g/dl)
ratio_albumin_avg	Average value of ratio between albumin and globulin
quantity_max_found	Maximize quantity obtained from tourniquet test.
pulse_pre_min_found	The pulse pre min values of a patient.
rash_found	Evidence of rash (Yes/No)
itching_found	Evidence of itching (Yes/No)
bruising_found	Evidence of bruising (Yes/No)
diarrhea_found	Evidence of diarrhea (Yes/No)
uri_found	Evidence of upper respiratory infection (Yes/No)
abdominal_found	Evidence of abdominal (Yes/No)
dyspnea_found	Evidence of dyspnea (Yes/No)

ascites_found	Evidence of ascites (Yes/No)
jaundice_found	Evidence of jaundice (Yes/No)
liver_tenderness	Evidence of liver tenderness (Yes/No)
liver_found	Evidence of Grown liver (Yes/No)
lymph_found	Evidence of lymph node enlargement (Yes/No)
injected_found	Evidence of injected conjunctive.
atypical_lymp_found	Evidence of atypical lymphocyte.
Effusion_Result	Evidence of effusion obtained from X-ray or Ultrasound test (Yes/No)
leakage	Evidence of plasma leakage (Yes/No)

4 DECISION TREE APPROACH

Decision tree learning [4, 5] is a supervised learning method. The algorithm constructs a tree which consists of a set of selected attributes. These attributes are qualified by the gain ratio since they

can reduce the entropy of the classes. Consider the entropy equation (see equation 1). For the multiclass problem, entropy equation is defined as shown in equation 2. Finally the gain value is calculated in equation 3.

$$Entropy(S) = \frac{-P}{P+N} \log_2 \frac{P}{P+N} - \frac{N}{P+N} \log_2 \frac{N}{P+N} \quad (1)$$

Where S is the training data set, P is the number of positive class and N is the number of negative class.

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

Note that S is the training data set, p_i is a ratio of class i compare with all data, and c is the number of class.

$$Gain(S, A) = Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (3)$$

Note that S is the prior data set before classified by attribute A , $|S_v|$ is the number of examples those value of attribute A are v , $|S|$ is the total number of records in the data set.

5 PERFORMANCE MEASUREMENT

We use sensitivity, specificity and accuracy as performance measures. Three equations are defined as following. Sensitivity (see equation 4) measures the proportion of the positive class which are correctly identified (e.g. the percentage of dengue patients who are correctly identified as having the

condition). Specificity (see equation 5) measures the proportion of the negative class which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition). Moreover, we apply accuracy measurement (see equation 6) in order to evaluate the proportion of the true results.

$$Sensitivity = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}} \quad (4)$$

$$Specificity = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}} \quad (5)$$

$$Accuracy = \frac{\text{number of True Positives} + \text{number of True Negatives}}{\text{number of True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \quad (6)$$

6 EXPERIMENTAL RESULTS

6.1 Data Set

The total number of patients was 258 patients that obtained from Siriraj Hospital, Bangkok, Thailand. The data set consists of 128 DF, 65 DHF I, 52 DHF II and 13 DHF III. These attributes value are clinical attributes and hematological attributes. There are 48 attributes (26 numerical attributes, 21 categorical attributes.

Attributes in Table 3 were recorded during the first visit of each patient. Some attributes were preprocessed such as Bleeding. The Bleeding value was determined from any evidences found from spontaneous petechiae, ecchymosis, gum, nose, vomiting, stool and others.

During the treatment period, nurses and physicians followed the symptoms as shown in Table 4 and 5. Temporal attributes are summarized in terms of maximum, minimum and average values.

Table 3: Attributes obtained in the early phrase of treatment

Attribute	Type	Meaning
JE vaccine	Categorical	Received JE vaccine
URI	Categorical	Upper respiratory tract infection
Bleeding	Categorical	Bleeding

Table 4: Attributes obtained during the treatment period (numerical values)

Attribute	Meaning
hematocrit_max	Maximum value of hematocrit concentration
hematocrit_min	Minimum value of hematocrit concentration
AST_max	Maximum value of AST
AST_min	Minimum value of AST
AST_avg	Average value of AST
ALT_max	Maximum value of ALT
ALT_min	Minimum value of ALT
ALT_avg	Average value of ALT
temperature_max	Maximum of temperature
temperature_min	Minimum of temperature
sbp_dbp_avg	The difference between sbp and dbp
liver_size_avg	Average size of grown liver
hematocrit_max_dx	Maximum value of hematocrit concentration
hematocrit_min_dx	Minimum value of hematocrit concentration
hematocrit_avg_dx	Average value of hematocrit concentration
white_blood_cell_max	Maximum of WBC (x1000)
white_blood_cell_min	Minimum of WBC (x1000)
white_blood_cell_avg	Average of WBC (x1000)
platelet_max	Maximum of platelet count (x1000) by machine
platelet_min	Minimum of platelet count (x1000) by machine
platelet_avg	Average of platelet count (x1000) by machine
protein_avg	Average value of protein in liver
albumin_avg	Average value of albumin
globulin_avg	Average value of globulin
ratio_albumin_avg	Average value of ratio between albumin and globulin
quantity_max_found	Maximize quantity value of tourniquet test

Table 5: Attributes obtained during the treatment period (categorical values)

Attributes	Meaning
pulse_pre_min_found	Minimum of different pressure value evidence
rash_found	Rash on skin evidence
itching_found	Itching related to rash evidence
bruising_found	Bruising evidence
diarrhea_found	Diarrhea evidence
uri_found	Upper respiratory infection evidence
abdominal_found	Abdominal pain
dyspnea_found	Evidence of dyspnea

ascites_found	Evidence of ascites
jaundice_found	Evidence of jaundice
liver_tenderness	Evidence of liver tenderness
liver_found	Evidence of grown liver
lymph_found	Evidence of lymph node enlargement
injected_found	Injected conjunctive evidence
atypical_lymph_found	Atyp lymphocyte evidence
Effusion_Result	Effusion evidence
leakage	Evidence of plasma leakage
shock	Evidence of shock
dx	Class

6.2 The First Experiment

In the first experiment, we used decision tree learning algorithm in order to find the knowledge in dengue patient's data set. The data set consists of 4 classes which were DF, DHF I, DHF II and DHF III. We obtained the decision tree as shown in Figure 1. We found 7 significant attributes needed to classify patients. These attributes were **leakage** - leakage of plasma in blood, **shock** - shock evidence found during treatment period, **Bleeding** - bleeding evidence found, **lymph_found** - lymph node enlargement found, **quantity_max_found** - bleeding spot found under skin, **platelet_avg** - the average of platelet count and **AST_max** - the level of aspartate aminotransferase. There was only one

rule found for the DF patients. If there was no leakage evidence found in the patient, he/she would be diagnose as DF. There were three rules found for the DHF I patients. If leakage evidence was found and no shock and no bleeding evidence were found, those patients would be diagnose as DHF I. The second rule, if bleeding evidence was found, no lymph node enlargement and the tourniquet result were 14-17 bleeding spots obtained from tourniquet test, the patients would be diagnose as DHF I. The third rule, if the bleeding spots were less than 14 and the average number of platelet count was more than 88.8 cells/ μ L and the maximum level of AST was more than 131 U/L, then the patients would be diagnose as DHF I.

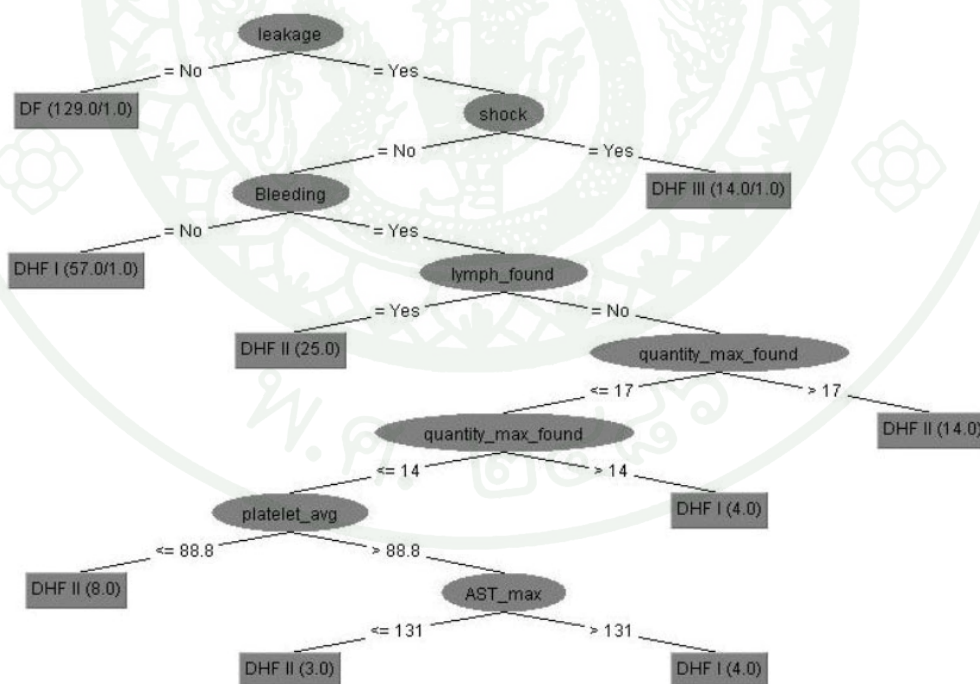


Figure 1: Decision tree with 4 classes.

Consider DHF II class; there were four rules. The patients would be diagnose as DHF II if there were leakages evidence, no shock, bleeding evidence and lymph node enlargement evidence. However, if lymph node enlargement evidence was not found and if the bleeding spots obtained from tourniquet test were more than 17, they would be diagnose as DHF II. The third rule of DHF II was that if the maximum quantity of bleeding spots

obtained from tourniquet test was less than 14 and the average of platelet count was less than 88.8 cells/ μ l. The forth rule was that if the average of platelet count was more than 88.8 cells/ μ l and the maximum of AST was less than 131 U/L, they would be diagnose as DHF II. For DHF III class, there was only one rule found. The patient would be diagnose as DHF III if they found leakage evidence and shock evidence.

Table 7: Performance of the first experiment.

Class	Sensitivity(%)	Specificity(%)	Accuracy(%)	Overall Accuracy (%)
DF	100.00	99.13	99.59	96.50
DHF I	86.15	96.88	94.16	
DHF II	86.54	96.10	94.16	
DHF III	100.00	99.57	99.59	

Table 8: Confusion matrix of the first experiment

	a	b	c	d	classified
128	0	0	0	0	a = DF
0	45	6	1	1	b = DHF II
1	8	56	0	0	c = DHF I
0	0	0	0	13	d = DHF III

We found that the decision tree completely classified patients in DF and DHF III with 100 % on sensitivity value. For DF class, the specificity performances of DHF II class were 86.54 %, 96.10 % and 94.16 % measured on sensitivity, specificity and the average accuracy, respectively. The specificity and the average accuracy of DHF III were 99.57 % and 99.59 %, respectively. The last column shows the overall accuracy of this model which was 96.5 % (see Table 6 for details) value was 99.13 %. The accuracy of DF class was 99.59 %. Consider DHF I class, we found that the sensitivity, specificity, and average accuracy were 86.15 %, 96.88 % and 94.16 %, respectively.

6.3 The Second Experiment

WHO has launched new criteria to classify patients into 2 classes. Therefore, we set up the second experiment in order to classify patients into 2 groups. We reassign DF and DHF I as Non Severe (Group1) and reassign DHF II and DHF III as Severe (Group2). Consider the decision tree shown in Fig. 2, we found 8 significant attributes useful for classifying data. Same attributes as the first experiment were found in the tree which were leakage evidence, shock evidence, bleeding evidence and lymph node enlargement evidence. Different attributes were abdominal pain, an average of white blood cell, an upper respiratory infection and a minimum of patient's temperature. There were 5 rules for non-severe group and 5 rules for severe group.

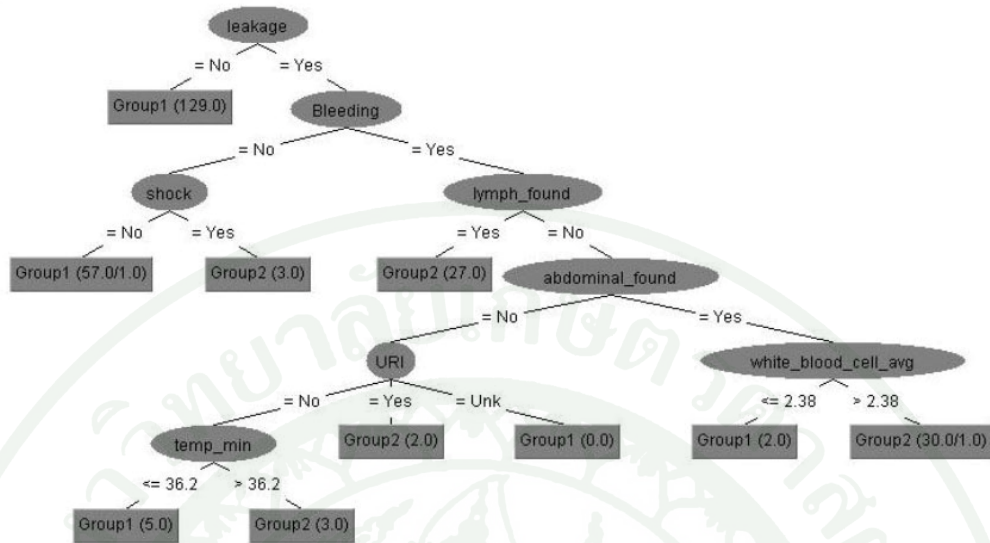


Figure 2: Decision tree with 2 classes.

Table 8: Confusion matrix of the second experiment

a	b	Classified as
185	8	a = Group1
5	60	b = Group2

Table 9: Performance of second experiment

Class	Sensitivity (%)	Specificity (%)	Accuracy (%)	Overall Accuracy (%)
Group1	95.85	92.31	94.96	96.00
Group2	92.31	95.85	94.96	

Table 8 represented the number of correctly classified patients performed by decision tree. We found that 185 patients from 193 patients can be correctly classified as a non-severe group. Sixty patients from 65 patients were correctly classified as a severe group.

For non-severe group, we found that the sensitivity, specificity and the average accuracy were 95.85 %, 92.31 % and 94.96 %, respectively. For severe group, we found that the sensitivity, specificity and average accuracy were 92.31 %, 95.85 % and 94.96 %, respectively. We obtained 96.00 % of the overall accuracy from ROC Area in the second experimental result (see Table 9 for details)

7 DISCUSSION

7.1 Experimental Result

Tanner, et al. [3], applied decision tree algorithm to classify the patients into 4 levels of dengue patients and non-dengue patients (Probable dengue, Likely dengue, Likely non-dengue and Probable non-dengue). The accuracy was only 84.70%. Thus in this paper, we used decision tree algorithm in order to learn and classify type of the patients (called "grading"). We would like to classify the patients into 4 classes. The first experimental result showed new knowledge in Fig. 1. Our accuracy rate was 96.50 % in the first experiment. For the second experiment, we would like to classify dengue infection into 2 groups (severe and non-severe). Our second experiment was similar to Tarig's experiment [7]. They applied Self Organization Map to characterize the low risk and high risk dengue patients. They used three features to cluster

dengue patients which were HCT, PLT and AST/ALT. Tarig Faisal [8] classified the dengue patients using two algorithms of neural networks. They used three features from the research [7]. Their best algorithm was MLPSCG [8]. Their accuracy was 75.00 % whereas our accuracy was better than their result. Our experimental result found different feature set compared to those of Tarig Faisal [8]. We found 8 significant features which were leakage, bleeding evidence, shock, lymph node enlargement, abdominal pain, upper respiratory infection, white blood cell and body temperature. Our accuracy of the second experiment was 96.00 %.

7.2 Result validation with other criteria

As stated before that WHO has launched a set of criteria for physician to classify dengue infection patients. The decision tree classified the patients into 4 classes which were shown in Table 10. The false negative values were examined for comparison between WHO criteria and the decision tree. The false negative of decision tree were 10.77% and 17.31% for DHF I and DHF II. We applied the WHO criteria to compared with the first experimental result, the patients (258 patients) were classified by WHO criteria (see Table 11)

Table 10: False Negative value obtained from the first experiment

Class	Label data	Number of patients classified by Decision tree	False Negative (%)
DF	128	129	0
DHF I	65	53	10.77
DHF II	52	62	17.31
DHF III	13	14	0

Table 11: Result of WHO criteria

Class	Label data	Number of patients classified by WHO criteria	False Negative (%)
DF	128	165	5.47
DHF I	65	39	44.62
DHF II	52	35	40.38
DHF III	13	0	100
DHF IV	0	17	-
Non Dengue	0	2	-

Consider DF class, the false negative value obtained from WHO was 5.47 % higher than the decision tree. Moreover the false negative value of DHF I, DHF II, DHF III were 33.85 %, 23.07 % and 100 % higher than the decision tree. It means that the criteria from WHO were not sufficient to classify type of dengue patients. However, our decision tree provides better performance in classifying patients. We hope that the knowledge obtained from the decision tree algorithm may help physicians in diagnosis process.

From Table 12, we found the value of false negative were 4.15 % and 7.69 %. After that, we considered the data using Tarig's criteria (see Table 13). We found that a false negative value of non-severe group was 20 and a false negative value of severe group was 50. We calculated them in term of percentage in Table 14. We found that the false negatives were increased when we used Tarig's criteria to classify the data.

Table 12: Our second experimental result

Class	Label Data	Number of patients classified by decision tree	False Negative (%)
Group1 (Low risk)	193	190	4.15
Group2 (High risk)	65	68	7.69

Table 13: Confusion Matrix using Tarig's criteria

a	b	Classified as
173	20	a = Group1
50	15	b = Group2

Table 14: The result of Tarig's criteria

Class	Label data	Number of patients Classified by Tarig's criteria	False Negative (%)
Group1 (Low risk)	193	223	10.36
Group2 (High risk)	65	35	76.92

Table 15: False Negative values

Class	Number of patients		
	Decision Tree classifier	WHO criteria	Tarig's criteria
DF	0	7	-
DHF I	7	29	-
DHF II	9	21	-
DHF III	0	13	-
Group1	8	-	20
Group2	5	-	50

Using Tarig's criteria, their result also gave more false negative value than that of our experimental result. That means their criteria were not sufficient in classifying the data because they had much of the false negative value.

Table 15 represented details about the false negative value when decision tree, WHO criteria and Tarig's criteria were applied for classifying data. The number of false negative using WHO criteria were greater than that of decision tree classifier and the number of false negative patients using Tarig's criteria were greater than that of decision tree classifier. Our experiment gave better classifying result than WHO criteria and Tarig's criteria.

8 Conclusion

Our research work is in the framework of data mining. We try to find new knowledge that contributes to the more accurate classifying results. We got an accuracy as 96.50% for classify levels into 4 groups and 96.00 % for classify levels into 2 groups.

We create new feature set that make the learning algorithm succeeded in classifying task. Finally, we found some significant features such as lymph node enlargement and upper respiratory infection that are useful to differentiate the degree of dengue patients.

Acknowledgement

This research is supported by Faculty of Science and Kasetsart University and Research Development Institute, Bangkok Thailand.

REFERENCE

- [1] D. Delen, G. Walker and A. Kadam: Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*; 34 (2005) 113–27
- [2] F. Ibrahim, M. N Taib, W. A. B. Wan Abas, C. G. Chan and S. Sulaiman: A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), *Comput. Methods Programs Biomed.* 79 (2005) 273–281
- [3] L. Tanner, M. Schreiber, J.G. Low, A. Ong, T. Tolfvenstam, Y.L. Lai, L.C. Ng, Y.S. Leo, L. Thi Puong, S.G. Vasudevan, C.P. Simmons, M.L. Hibberd and E.E. Ooi: Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness. *PLoS Negl. Trop. Dis.* 2 (2008), 196.
- [4] C. NadaLavra: Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 16 (2) (June, 1999)
- [5] N. Soonthornphisaj: *Artificial Intelligence*. Chulalongkorn University Printing House. 2010.
- [6] T. Faisal, F. Ibrahim and M.N. Taib: A noninvasive intelligent approach for predicting the risk in dengue patients. *Expert Systems with Applications*, Volume 37, Issue 3 (15 March 2010) 2175-2181, ISSN 0957-4174
- [7] T. Faisal, M.N. Taib and F. Ibrahim: Reexamination of risk criteria in dengue patients using the self-organizing map. *Med. Biol.Eng.Comput.*48, (2010) 293-301
- [8] T. Faisal, M.N. Taib and F. Ibrahim: Neural network diagnostic system for dengue patients risk classification.(June, 2010)
- [9] World Health Organization: *Guideline for Treatment of Dengue Fever/Dengue Haemorrhagic Fever* (1999), 5

Data Mining of Dengue Infection Using Decision Tree

DARANEE THITIPRAYOONWONGSE¹, PRAPAT SURİYAPHOL² AND NUANWAN SOONTHORNPHISAJ^{1*}

¹Department of Computer Science, Faculty of Science, Kasetsart University
BANGKOK, THAILAND
g521440162@ku.ac.th, fscinws@ku.ac.th

²Bioinformatics and Data Management for Research Unit Office for Research and Development
Siriraj Hospital, Mahidol University
BANGKOK, THAILAND
sipur@mahidol.ac.th

Abstract: - Dengue infection is an epidemic disease typically found in tropical region. Nowadays, the experts need to know the set of features on dengue infection in order to correctly classify the patients since these classes require different treatment. Our temporal dataset consists of clinical data and laboratory data. The data was collected from the first visit of patient until the date of discharge. We obtained 2 sources of datasets from different regions of Thailand which are Srinagarindra Hospital and Songklanagarind Hospital. Each dataset consists of more than 400 attributes. To accomplish the knowledge discovery task, we consider to employ decision tree as a data mining tool. We propose a set of meaningful attributes from the temporal data. Our experiments are divided into 4 parts. The first two experimental results show the useful knowledge to classify dengue infection from Srinagarindra Hospital's dataset and Songklanagarind Hospital's dataset, respectively. Each set of knowledge is tested by different dataset to make sure that the test data was a real unseen data. The third experimental results show the useful knowledge when we integrated 2 datasets. Another objective of this research is to detect the day of defervescence of fever which is called day0. The day0 date is the critical date of dengue patients that some patients face the fatal condition. Therefore the physicians need to predict day0 in order to treat the patients. They expect to have an intelligent system that can trigger the day0 date of each patient.

Key- Words: - Data Mining, Dengue infection, the day of defervescence.

1 Introduction

Dengue Fever causes by a bite from mosquitoes carrying dengue virus. Symptoms of dengue infection show rapid and violent to patients in a short time. Dengue infection is divided into 4 types which are DHFI, DHF II, DHF III and DHF IV, respectively [6]. In this paper, we obtained 2 sources of datasets. The total number of patients is 524 patients from Srinagarindra Hospital and 477 patients from Songklanagarind Hospital. We selected a decision tree learning as an approach to find knowledge in order to classify the type of dengue infection. We propose forty-eight attributes as a set of meaningful attributes [1].

We selected decision tree for 4 experiments. The first two experimental results show knowledge for each type of dengue infection. The third experimental results show knowledge from all sources. In the forth experiment, we predicted the day of defervescence of fever or day0.

* corresponding author

2 Related Work

There are some researchers who work on dengue classification such as Tanner, et al and Tarig, et al. The team of Tanner classified 1,200 patients using decision tree approach. They found 6 significant features and they got 84.7 % correctness [3]. A combination of the self-organizing map (SOM) and multilayer feed-forward neural networks (MFNN) was employed for the risk prediction of dengue patients in the Tarig's research. They clustered patients into 2 groups which are low risk and high risk using three criteria [5]. They used only examples from Day0 until Day2 (Day2 refers to 2 days after the day of defervescence of fever). On the other hand, they got only 70% correctness.

Fatimah Ibrahim et al. [2] predicted the day of defervescence of fever (day0) from 252 dengue patients (4 DF and 248 DHF). They used Multi-Layer Perceptrons (MLP) and got 90% correctness.

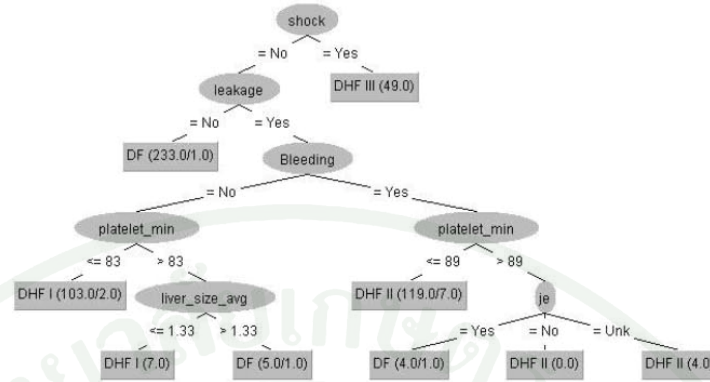


Fig.1 Decision tree of the first experiment.

3 Decision Tree Approach

Decision tree learning is a supervised learning method. The algorithm constructs a tree which consists of a set of selected attributes. These attributes are qualified by the gain ratio since they can reduce the entropy of the classes. Consider the entropy equation (see equation 1). For the multiclass problem, entropy equation is defined as shown in equation 2. Finally the gain value is calculated in equation 3.

$$\text{Entropy}(S) = \frac{-P}{P+N} \log_2 \frac{P}{P+N} - \frac{N}{P+N} \log_2 \frac{N}{P+N} \quad (1)$$

Where S is the training data set, P is the number of positive class and N is the number of negative class.

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

Note that S is the training data set, p_i is a ratio of class i compare with all data, and c is the number of class.

$$\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{S} \text{Entropy}(S_v) \quad (3)$$

Note that S is the prior data set before classified by attribute A, $|S_v|$ is the number of examples those value of attribute A are v, $|S|$ is the total number of records in the data set.

In this study, we used WEKA software in order to learn from the training set. We changed the value of some parameters in order to know more knowledge and get more accuracy. That parameter was "confidenceFactor" which was used for pruning set (smaller values incur more pruning).

4 Experimental Results

We need to preprocess data followed by the mining step [4]. We excluded noisy data such as outliers in some attributes. Then, we replaced missing value with the mean of some numeric attributes. In addition, we picked up the set of suitable attributes and created new features to represent some data pattern. Then we transformed some attribute values in order to qualify the requirement of the algorithms. In this paper, we used 48 attributes [1].

We set up 4 experiments. In the first three experiments, we would like to find knowledge in order to classify type of dengue infection. For forth experiment, we would like to predict the day of defervescence with the data before day0 date. We applied decision tree approach to all experiments.

Note that we use sensitivity, specificity and accuracy as performance measures.

4.1 Datasets

We divided dataset into 2 datasets. The total number of patients is 524 patients from Srinagarindra Hospital, Khon Kaen, Thailand (KK). There are 240 DF, 116 DHF I, 118 DHF II and 50 DHF III. In the dataset of Songklanagarind Hospital, Songkla, Thailand (SK), we obtained 477 patients that consists of 248 DF, 106 DHF I, 111 DHF II and 12 DHF III. There is no patient who was diagnosed as DHF IV in both datasets. These attributes consists of 26 numerical attributes, 21 categorical attributes and one class attribute.

4.2 The First Experiment

In the first experiment, we classified on KK's dataset. We set confidenceFactor value to 0.4. We obtained the decision tree as shown in Fig.1

We found 6 significant attributes which are needed to classify patients. These attributes were **shock** – shock evidence found during treatment period, **leakage** - leakage of plasma in blood, **Bleeding** – bleeding evidence found, **platelet_min** – the minimum of platelet count, **liver_size_avg** – the average number of liver size and **je** –JE vaccine.

There are three rules found for the DF patients. The first rule states that *if there is no shock and no leakage evidence then the patients would be diagnose as DF*. The second rule states that *if there is leakage evidence and there is no bleeding evidence, and the minimum number of platelet count is more than 83 and the average number of liver size more than 1.33 then those patients would be diagnose as DF*. The third rule states that *if bleeding evidence is found and the minimum of platelet count is more than 89 and the patient got JE vaccine then those patients would be diagnose as DF*.

There are two rules found for the DHF I patients. The first rule states that *if there is no shock and leakage evidence and no bleeding evidence and the minimum of platelet count is less than 83 then those patients would be diagnose as DHF I*. The second rule states that *if the minimum of platelet count more than 83 and the average number of liver size less than or equal 1.33 then those patients would be diagnose as DHF I*.

There are two rules found for DHF II. The patients would be diagnose as DHF II if there were no shock evidence and leakages evidence, bleeding evidence and the minimum of platelet count is less than 89. However, if the minimum of platelet count is more than 89 and there is no evidence of JE vaccine injection then the patient would be diagnose as DHF II.

For DHF III class, the rule states that the patient would be diagnose as DHF III if there was shock evidence.

We got three significant attributes which were shock evidence, leakage evidence and bleeding evidence. In the first experiment, the minimum of platelet count and the average number of liver size were correlated with the classes. If the minimum of platelet count was increased, the level of dengue fever was non-severe type. On the other hand, if the average number of liver size was increased, the level of dengue fever was severe type. Note that, JE vaccination attribute didn't relate with classes. The results conform to the decision tree.

Table 1 Performance of the first experiment using 10-fold cross validation.

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.08	97.47	97.29	97.6%
DHF I	93.10	98.50	97.29	
DHF II	95.76	97.99	97.48	
DHF III	98.00	100.00	99.80	

Table 1 shows the performance measured in terms of sensitivity (Sens.), specificity (Spec.) and the average accuracy (Acc.) We found that the decision tree classified patients in DHF III with 98.00 % on sensitivity value. The first experimental results show interesting attributes which were platelet_min, liver_size_avg and JE vaccination. Our accuracy rate was 97.6%.

To evaluate the potential of knowledge obtained from the decision tree, we also test the knowledge with the unseen test set using SK's dataset. We obtained the accuracy as shown in Table 2.

Table 2 Performance of 1st experiment measure on unseen testset

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.98	94.91	96.55	97.2%
DHF I	85.85	99.44	96.34	
DHF II	91.89	96.92	95.73	
DHF III	100.00	98.87	98.90	

We found that the decision tree can completely classified patients in DHF III with 100 % on sensitivity value. We obtained high accuracy when we cross check with different dataset. That means our knowledge can be applied on real data.

From the experimental results, the decision tree algorithm was learned from KK dataset and was tested using SK dataset. We got 97.2% for accuracy.

4.3 The Second Experiment

In the second experiment, we used SK's dataset. The dataset also consists of 4 classes. We set confidenceFactor value to 0.3. We obtained the decision tree as shown in Fig 2.

We found 9 significant attributes needed to classify patients. These attributes were **leakage** - leakage of plasma in blood, **shock** – shock evidence found during treatment period, **Bleeding** – bleeding evidence found, **platelet_min** – the minimum of platelet count, **abdominal_pain** – abdominal pain evidence found during treatment period, **rash_found** – rash evidence found during treatment period, **uri** - the appearance of upper respiratory infection, **hematocrit_min_d05** – the minimum

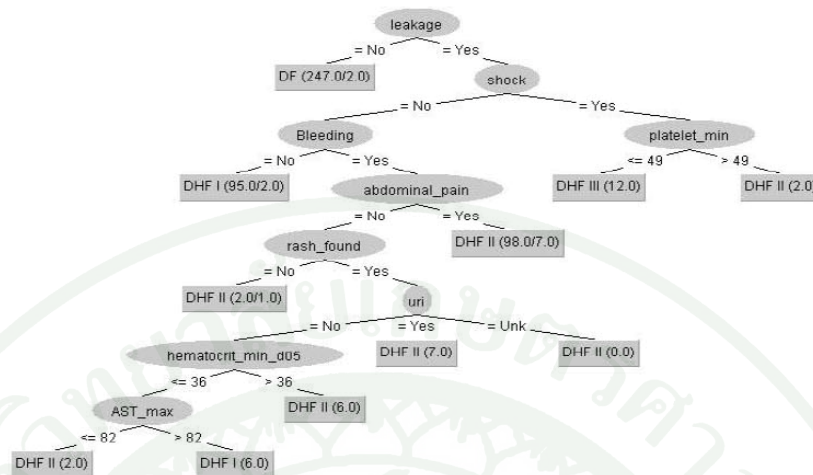


Fig.2 Decision tree of the second experiment

number of hematocrit test during treatment period and **AST_max** – the maximum value of aspartate aminotransferase.

There was only one rule found for the DF patients. The rule states that *If there is no leakage evidence, then the patients would be diagnosed as DF.*

There were two rules found for the DHF I patients. The first rule states that *if there is leakage evidence and no shock evidence and no bleeding evidence then those patients would be diagnosed as DHF I.* The second rule states that *if there were bleeding evidence, no abdominal pain, rash evidence found, no upper respiratory infection, the minimum number of hematocrit is less than 36 and the maximum value of AST is more than 82 then those patients would be diagnosed as DHF I.*

There were six rules found for DHF II class. The patients would be diagnose as DHF II if there were leakage evidence, shock evidence, and the minimum of platelet count more than 49. The second rule, if there were no shock evidence, bleeding evidence and they had abdominal pain, then those patients would be diagnosed as DHF II. The third rule, if there was no abdominal pain and no rash evidence, the patients would be diagnosed as DHF II. But if rash evidence and URI were found, then those patients would be diagnosed as DHF II. They also diagnose as DHF II if there were no URI and the minimum number of hematocrit more than 36. However, if that number less than 36 and the maximum value of AST less than 82, they would be diagnose as DHF II.

For DHF III class, there was only one rule found. The patient would be diagnose as DHF III if they found leakage evidence and shock evidence and the minimum of platelet count is less than or equal to 49

Table 3 Performance of the second experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	98.79	99.05	98.91	96.6%
DHF I	90.57	97.81	96.18	
DHF II	90.99	96.70	95.37	
DHF III	91.67	99.55	99.34	

We found that the decision tree classified patients in DF with 98.79 % on sensitivity value. In addition, we selected KK's dataset to be the test set. We obtained the accuracy as shown in Table 4. For the second experiment, there were 6 interested attribute. These attributes were platelet_min, abdominal_pain, rash_found, uri, hematocrit_min_d05 and AST_max. We obtained 96.6% of accuracy.

Table 4 Performance of 2nd experiment (on KK dataset)

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	96.67	99.62	98.21	95.7%
DHF I	93.97	96.97	96.29	
DHF II	94.07	95.50	95.17	
DHF III	82.00	100.00	98.21	

Consider the unseen data test, if we used SK as a training set and used KK as a test set, we got 95.7%.

In the second experiment, the minimum of platelet count was still related with classes. The maximum of AST related with low correlation

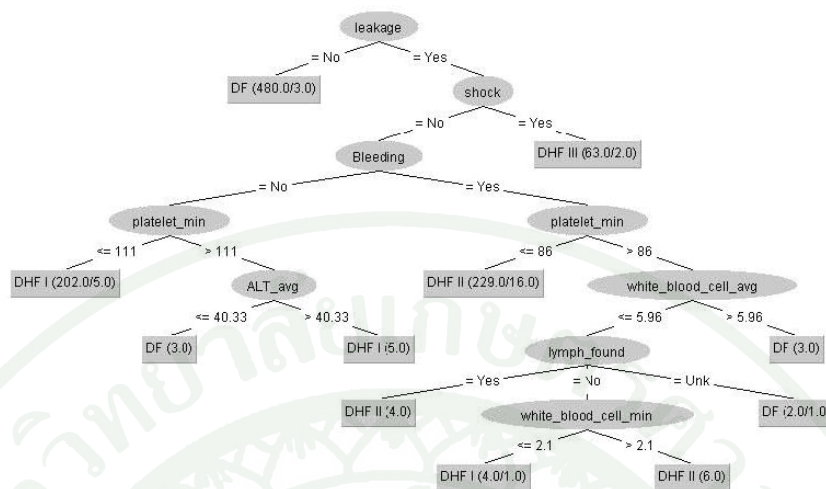


Fig.3 Decision tree of the third experiment.

coefficient value. When we compared with decision tree in the second experiment, `AST_max` node was a criterion which after `hematocrit_min_05` node. So that refer `AST_max` also didn't relate with classes. There were 4 attributes which didn't relate with classes. There were evidence of abdominal pain, rash found evidence, upper respiratory infection appearance and the minimum number of hematocrit test in the laboratory test.

4.4 The Third Experiment

This experiment, we merged 2 datasets and used decision tree algorithm in order to extract the total knowledge of dengue infection in Thailand. We obtained the decision tree as shown in Fig.3.

We found 8 significant attributes needed to classify patients. These attributes were **leakage** - leakage of plasma in blood, **shock** - shock evidence found during treatment period, **Bleeding** - bleeding evidence found, **platelet_min** - the minimum of platelet count, **ALT_avg** - the average of alanine aminotransferase value, **white_blood_cell_avg** - the average number of white blood cell in laboratory test, **lymp_found** - evidence of lymphocyte node enlargement, and **white_blood_cell_min** - the minimum number of white blood cell in laboratory test.

There were three rules found for the DF patients. If there was no leakage evidence, he/she would be diagnose as DF. The second rule, if there were leakage evidence, no bleeding, the minimum of platelet count more than 111 and the average number of ALT less than or equal 40.33, those patients would be diagnose as DF. However, if there

were bleeding evidence, the minimum of platelet count more than 86 and the average number of white blood cell more than 5.96, those patients would be diagnose as DF.

For DHF I, there were three rules. If there were leakage evidence, no shock evidence, no bleeding evidence and the minimum of platelet count less than or equal 111, those patients would be diagnose as DHF I. If the minimum of platelet count more than 111 and the average number of ALT more than 40.33, those patients would be diagnose as DHF I. The third rule, there were bleeding evidence, the minimum of platelet count more than 86, the average number of white blood cell less than or equal 5.96, no lymphocyte node enlargement and the minimum number of white blood cell less than or equal 2.1, those patients would be diagnose as DHF I.

Consider DHF II class; there were three rules. The patients would be diagnosed as DHF II if there was leakage evidence and no shock evidence, and bleeding evidence was found and the minimum of platelet count was less than or equal 86 then those patients would be diagnosed as DHF II. The second rule, if the minimum of platelet count was more than 86, the average value of white blood cell less than or equal 5.96 and lymphocyte node enlargement evidence was found, then those patients would be diagnosed as DHF II. The third rule, if there were no lymphocyte node enlargement evidence and the minimum number of white blood cell more than 2.1, the patients would be diagnosed as DHF II.

For DHF III class, there was only one rule found. The patient would be diagnosed as DHF III if they found leakage evidence and shock evidence.

In the third experiment, the minimum of platelet count, the average number of ALT and the minimum number of white blood cell were correlated with classes. We obtained two attributes which were not correlated with the classes. These attributes were the evidence of lymphocyte node enlargement and the average number of white blood cell. This results also conform to the decision tree.

Table 5 Performance of third experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
DF	97.34	98.56	97.95	96.7%
DHF I	89.64	98.44	96.46	
DHF II	95.63	96.84	96.56	
DHF III	98.39	99.55	99.48	

We found that the decision tree classified patients in each class. As shown in Table 5. The overall accuracy of this model was 96.7 %.

4.5 The Forth Experiment

This experiment, we focus on day0 prediction. We used daily data obtained from dengue patients without missing value records. We selected the data before day0 and day0 date. The training set was labeled into 5 classes which were day0, day-1, day-2, day-3 and day-4. Note that “day0” referred the day of defervescence, where as “day-1” referred to 1 day before day0 and so on. We obtained the accuracy as shown in Table 6.

Table 6 Performance of third experiment

Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
day-4	0.00	97.79	96.09	67.8
day-3	5.00	94.02	87.01	
day-2	40.91	82.20	73.18	
day-1	49.32	64.50	58.62	
day0	68.36	67.11	67.79	

The algorithm can correctly classification the day-1 patients with the accuracy rate 58.62%. For the day-2 patients, we obtained 73.18% correctness. We obtained 67.79% of accuracy. The overall accuracy was 67.8%.

5 Conclusion

This study proposed to use a decision tree approach to classify dengue patients from two datasets. We got an accuracy as 97.6%, 96.6 % from the first experiment and the second experiment, respectively. After we explored the correlation of attributes found in the decision trees, we found some significant

features which were platelet_min, liver_size_avg, ALT_avg and white_blood_cell_min.

We applied the unseen data in order to test knowledge in the first two experiments. We found that the first decision tree completely classified patients in DHF III with 100% (see Table 2 for details). Because the training set of KK contained more examples compared to those of SK dataset. The accuracy of both experiments in unseen test set were more than 95%.

In the experiment of day0, we obtained very low accuracy in day-4 and day-3 (see Table 6 for details). We found that the tree is overfit. The experimental results shown that the decision tree approach didn't suit this task thus we think we should to select a new classification approach in the future works.

Acknowledgement

This research is supported by Kasetsart University Research and Development Institute (KURDI) under Grant no. 130.55.

References:

- [1] D. Thitiprayoonwongse, P. Suriyaphol, and N. Soonthornphisaj, Data Mining on Dengue Virus Disease, *13th International Conference on Enterprise Information Systems (ICEIS 2011)*, No.1, 2011, pp. 32-41.
- [2] F. Ibrahim, M. N Taib, W. A. B. Wan Abas, C. G. Chan and S. Sulaiman, A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), *Computer Methods and Programs in Biomedicine*, No.79, 2005, pp. 273-281.
- [3] L. Tanner, M. Schreiber, J.G. Low, A. Ong, T. Tolfvenstam, Y.L. Lai, L.C. Ng, Y.S. Leo, L. Thi Puong, S.G. Vasudevan, C.P. Simmons, M.L. Hibberd and E.E. Ooi, Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness, *PLoS Neglected Tropical Disease*, Vol.2, 2008.
- [4] N. Soonthornphisaj, *Artificial Intelligence*, Chulalongkorn University Printing House, 2010.
- [5] T. Faisal, F. Ibrahim and M.N. Taib, A noninvasive intelligent approach for predicting the risk in dengue patients, *Expert Systems with Application*, Vol.37, No.3, 2010, pp. 2175-2181.
- [6] World Health Organization, *Guideline for Treatment of Dengue Fever/Dengue Haemorrhagic Fever*, 1999

A Data Mining Framework for Building Dengue Infection Disease Model

Darane Thitiprayoonwongse¹ Prapat Suriyaphol² Nuanwan Soonthornphisaj^{*1}

^{*1} Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok Thailand

^{*2} Bioinformatics and Data Management for Research Unit, Office for Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok Thailand

Dengue infection is a virus-caused disease that is spread by mosquitoes. Currently, there is no specific treatment, moreover there is no vaccine to prevent the people from the disease. Symptoms of this infection show rapid and violent to patients in a short time. Two classification problems are explored in this study, which are the Dengue Classification Problem and the Day of defervescence of fever (day0) detection problem. Two data mining models, fuzzy logic and decision tree, are studied in this work. We propose to use the knowledge obtained from the decision tree with the fuzzy logic approach in order to obtain better performance. The experimental result shows that using fuzzy logic approach for Dengue Classification is a suitable method. We get 97.39% of the average accuracy from fuzzy logic approach. Consider the day0 detection problem, using the attributes obtained from the dengue patient can provide the early warning on one day before the day0 with the sensitivity as 71.11% by using fuzzy logic.

Key-word: Data Mining, Dengue infection, the day of defervescence of fever, fuzzy logic.

1. Introduction

Dengue infection is one of the major health problems in Thailand. According to World Health Organization, Dengue is divided into 4 types which are DHF I, DHF II, DHF III and DHF IV, respectively [WHO 1999]. Symptoms of dengue infection show rapid and violent to patients in a short time. In this paper, we obtain 2 sources of datasets which contain DF, DHF I, DHF II and DHF III. The data was collected from the first visit of patient until the date of discharge. The number of patients is 524 patients from Srinagarindra Hospital and 477 patients from Songklanagarind Hospital. We propose forty-eight attributes as a set of meaningful attributes. We selected a decision tree learning as an approach to find the set of informative attributes in order to classify the type of dengue infection. Moreover we selected a fuzzy reasoning approach to classify the type of dengue infection to compare the performance.

Another objective of this research is to provide the early warning for the day of defervescence of fever which is called day0. The day0 date is the critical date of Dengue patients that some patients face the fatal condition. Therefore the target class is day-1 (a day before the day0 date). Therefore we use a set of attributes found in the decision tree as an attribute in fuzzy reasoning and we generate the fuzzy rules from the set of rules obtained from the decision tree.

2. Related Work

The problem of dengue classification was explored by Tanner, et al and Faisal, et al. The team of Tanner applied decision tree approach to classified patients into 4 levels which were Probable dengue, Likely dengue, Likely non-dengue and Probable non-dengue. Their dataset contained 1,012 patients from the EDEN study and 188 patients from Vietnam. They found 6 significant

features which were platelet count, white blood cell count, body temperature, hematocrit, absolute number of lymphocytes and absolute of neutrophils. They obtained only 84.7 % correctness. [Tanner 2008].

The team of Faisal's research [Faisal 2010] used a combination of the self-organizing map (SOM) and multilayer feed-forward neural networks (MFNN) to predict the risk of dengue patients. They classified patients into 2 groups which were low risk and high risk using three criteria. These criteria were platelet counts less or equal than 40,000 cell per mm³, hematocrit greater than or equal to 25% and aspartate aminotransferase or alanine aminotransferase rose by fivefold the normal upper limit. They used only examples from Day0 until Day2 (Day2 refers to 2 days after the day of defervescence of fever). Finally, they got only 70% correctness.

Fatimah Ibrahim et al. [Ibrahim 2005] predicted the day of defervescence of fever (day0) from 252 dengue patients (4 DF and 248 DHF). They used Multi-Layer Perceptrons (MLP) and got 90% correctness. However they just used eight signs and symptoms in their research. In this paper, seven significant attributes from decision tree and the fuzzy logic approach were employed to detect the day of defervescence of fever.

3. Method

The raw data obtained from two hospitals in north eastern and southern of Thailand was prepared before applying to the algorithm.

3.1 Datasets

Our datasets were obtained from Srinagarindra Hospital, Khon Kaen, Thailand (KK) and Songklanagarind Hospital, Songkla, Thailand (SK). KK data consists of 524 patients. There are 240 DF, 116 DHF I, 118 DHF II and 50 DHF III. In SK dataset, we get 477 patients that consists of 248 DF, 106 DHF I, 111 DHF II and 12 DHF III. No DHF IV is found in both data sets therefore

¹Contact: Nuanwan Soonthornphisaj, Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand. Email: fscinws@ku.ac.th

we will discard this dengue class. These attributes consists of 26 numerical attributes, 21 categorical attributes and one class attribute.

We excluded noisy data such as outliers in some attributes. Then, we replaced missing value with the mean of the same attributes. In addition, we picked up the set of suitable attributes and created new features to represent some data pattern. Then we transformed some attribute values in order to qualify the requirement of the algorithms. In this paper, we got 48 attributes from the clinical and laboratory data [Thitiprayoonwongse 2011].

Table 1. Feature set obtained after preprocessing step.

Attribute name	Meaning
JE vaccine	Received JE vaccine
URI	Upper respiratory tract infection
Bleeding	Bleeding
hematocrit_max	Maximum value of hematocrit concentration
hematocrit_min	Minimum value of hematocrit concentration
AST_max	Maximum value of AST
AST_min	Minimum value of AST
AST_avg	Average value of AST
ALT_max	Maximum value of ALT
ALT_min	Minimum value of ALT
ALT_avg	Average value of ALT
temperature_max	Maximum of temperature
temperature_min	Minimum of temperature
sbp_dbp_avg	The difference between sbp and dbp
liver_size_avg	Average size of grown liver
hematocrit_max_dx	Maximum value of hematocrit concentration
hematocrit_min_dx	Minimum value of hematocrit concentration
hematocrit_avg_dx	Average value of hematocrit concentration
white_blood_cell_max	Maximum of WBC (x1000)
white_blood_cell_min	Minimum of WBC (x1000)
white_blood_cell_avg	Average of WBC (x1000)
platelet_max	Maximum of platelet count (x1000) by machine
platelet_min	Minimum of platelet count (x1000) by machine
platelet_avg	Average of platelet count (x1000) by machine
protein_avg	Average value of protein in liver
albumin_avg	Average value of albumin
globulin_avg	Average value of globulin
ratio_albumin_avg	Average value of ratio between albumin and globulin
quantity_max_found	Maximize quantity value of tourniquet test
pulse_pre_min_found	Minimum of different pressure value evidence

Attribute name	Meaning
rash_found	Rash on skin evidence
itching_found	Itching related to rash evidence
bruising_found	Bruising evidence
diarrhea_found	Diarrhea evidence
uri_found	Upper respiratory infection evidence
abdominal_found	Abdominal pain
dyspnea_found	Evidence of dyspnea
ascites_found	Evidence of ascites
jaundice_found	Evidence of jaundice
liver_tenderness	Evidence of liver tenderness
liver_found	Evidence of grown liver
lymph_found	Evidence of lymph node enlargement
injected_found	Injected conjunctive evidence
atypical_lymph_found	Atypical lymphocyte evidence
Effusion_Result	Effusion evidence
leakage	Evidence of plasma leakage
shock	Evidence of shock
dx	Class

3.2 Decision Tree Approach

Decision tree is an inductive learning approach. It is a powerful and popular tool for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them.

A decision tree can be used to classify an example by starting from the root node and moving through it until a leaf node is found which provides the classification of the instance.

The requirements to do data with decision tree are 1) a set of predefined classes, 2) Attribute-value description: object or case must be expressible in terms of a fixed collection of properties or attributes, and 3) Sufficient training data.

The algorithm constructs a tree which consists of a set of informative attributes. These attributes are qualified by the gain ratio since they can reduce the entropy of the classes. Consider the entropy equation (see equation 1). For the multiclass problem, entropy equation is defined as shown in equation 2. Finally the gain value is calculated in equation 3.

$$\text{Entropy} = \frac{-P}{P+N} \log_2 \frac{P}{P+N} - \frac{N}{P+N} \log_2 \frac{N}{P+N} \quad (1)$$

Where S is the training data set, P is the number of positive class and N is the number of negative class.

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

Note that S is the training data set, p_i is a ratio of class i compare with all data, and c is the number of class.

$$\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{S} \text{Entropy}(S_v) \quad (3)$$

Note that S is the prior data set before classified by attribute A, $|S_v|$ is the number of examples those value of attribute A are v, |S| is the total number of records in the data set.

For the decision tree approach, we used a software namely WEKA to learn from the training set.

3.3 Fuzzy Logic

Fuzzy logic is a form of probabilistic logic. It evaluates with reason that is approximate rather than fixed value. In traditional logic theory, the binary sets have true or false but fuzzy logic variables may have a degree value of truth value that ranges between 0 and 1. The reasoning of fuzzy logic is similar to human reasoning.

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns discerned. The process of fuzzy inference involves all of the pieces that are described in the previous sections: Membership Functions, Logical Operations, and If-Then Rules. Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, and computer vision.

There are 5 steps for fuzzy inference system which are 1) fuzzify inputs 2) apply fuzzy operators 3) apply implication method 4) aggregate all output and 5) defuzzify.

In the first step, given the input attributes, the system determines the degree to each of the appropriate fuzzy sets via membership functions.

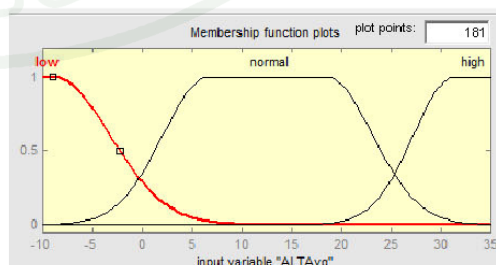
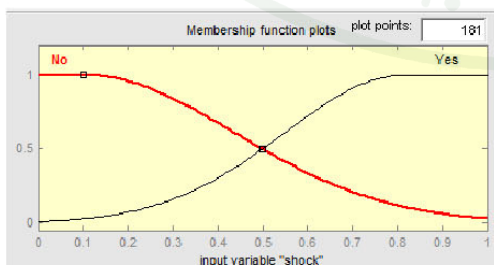
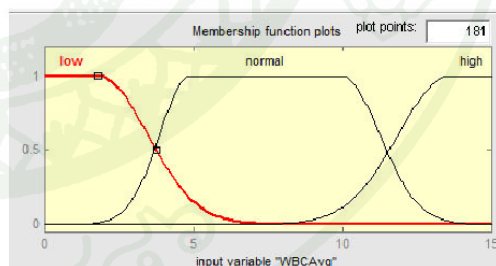
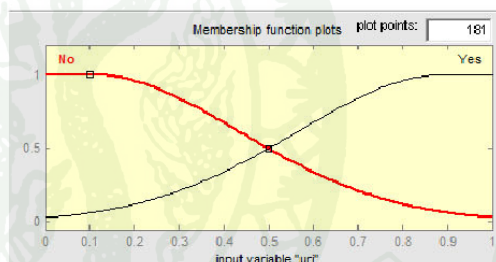
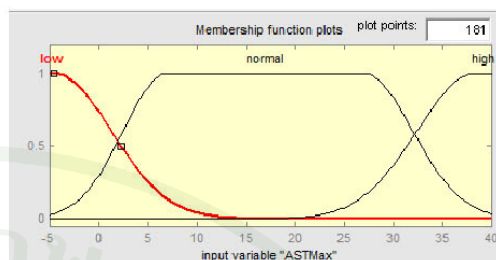
In the second step, we obtained the degree of inputs which were each part of the antecedent. We applied fuzzy operator when there is more than one condition in the rule. Note that In this work, we apply the set of rules obtained from the decision tree.

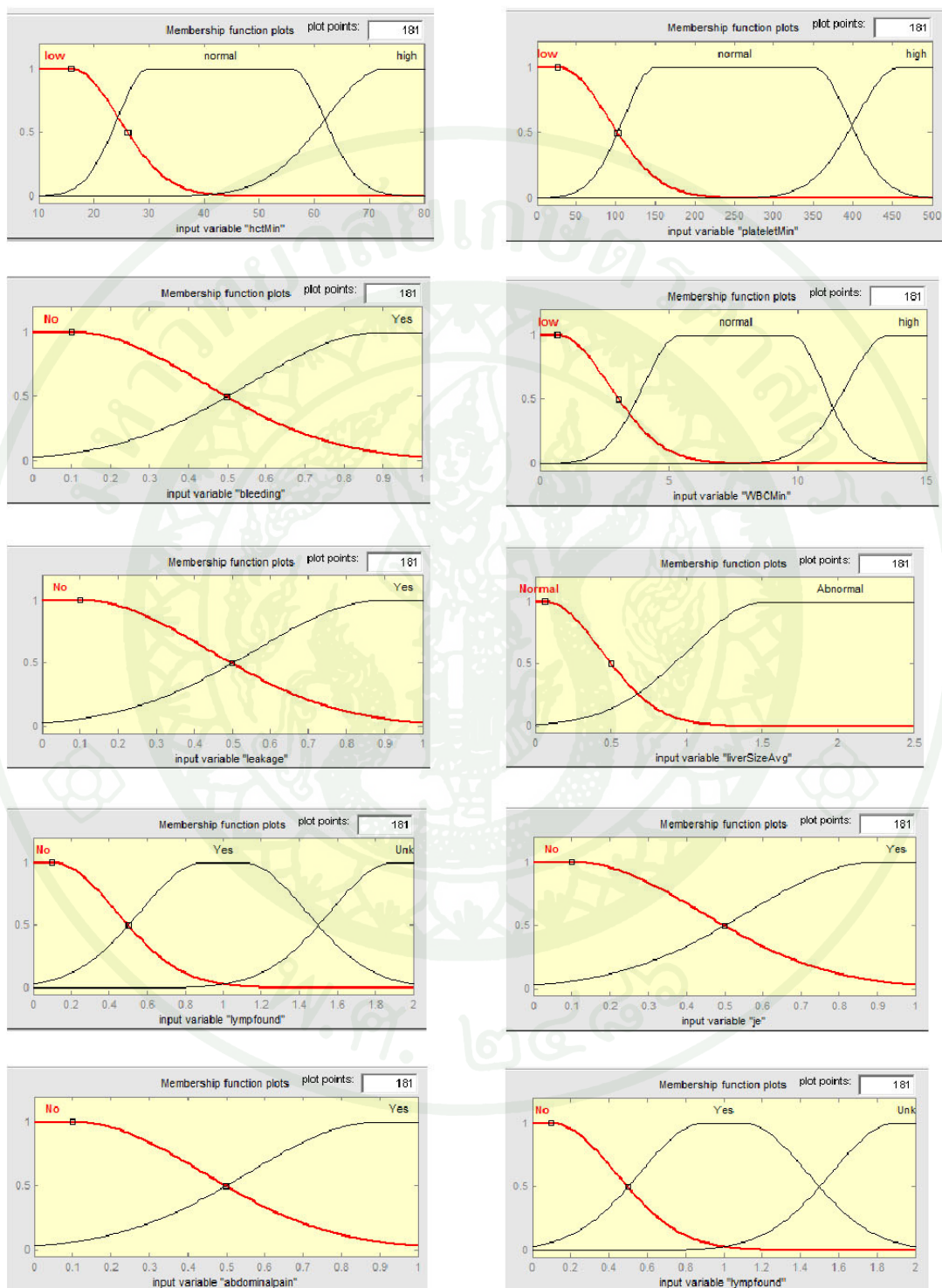
In the Apply Implication Method step, we can assign weight to each rule (a number between 0 and 1). This step is defined as the shaping of the fuzzy sets.

In the forth step, we aggregated the outputs of each rule and combine them into a single fuzzy set in preparation for final step.

For the defuzzification step, the centroid is calculated which correspond to the center of area under the curve. In this paper, we test our approach with the data which contains 968 instances and seven attributes. We create membership functions are built on Gaussian distribution curve with the normal range value in each input.

We provide the detail of each membership function used in Dengue Classification Problem and Day0 Detection Problem. Note that the range of each feature was obtained from expert. For the Dengue Fever Classification Problem, we propose to use 16 attributes as shown in Fig 1.





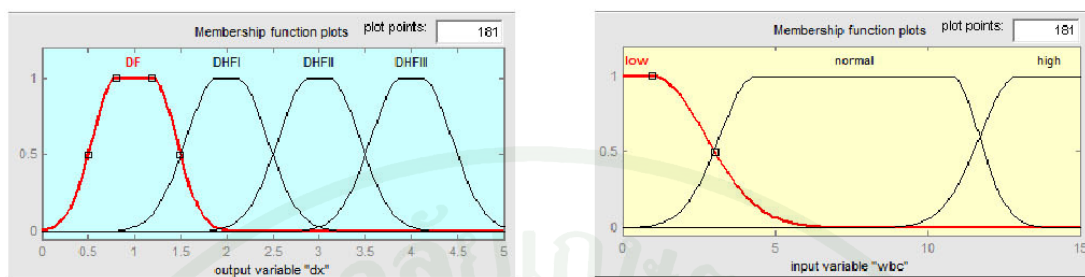


Fig. 1 Membership Functions of each input
in Dengue classification

For the Day0 Detection Problem, we propose to use 7
attributes as shown in Fig 2.

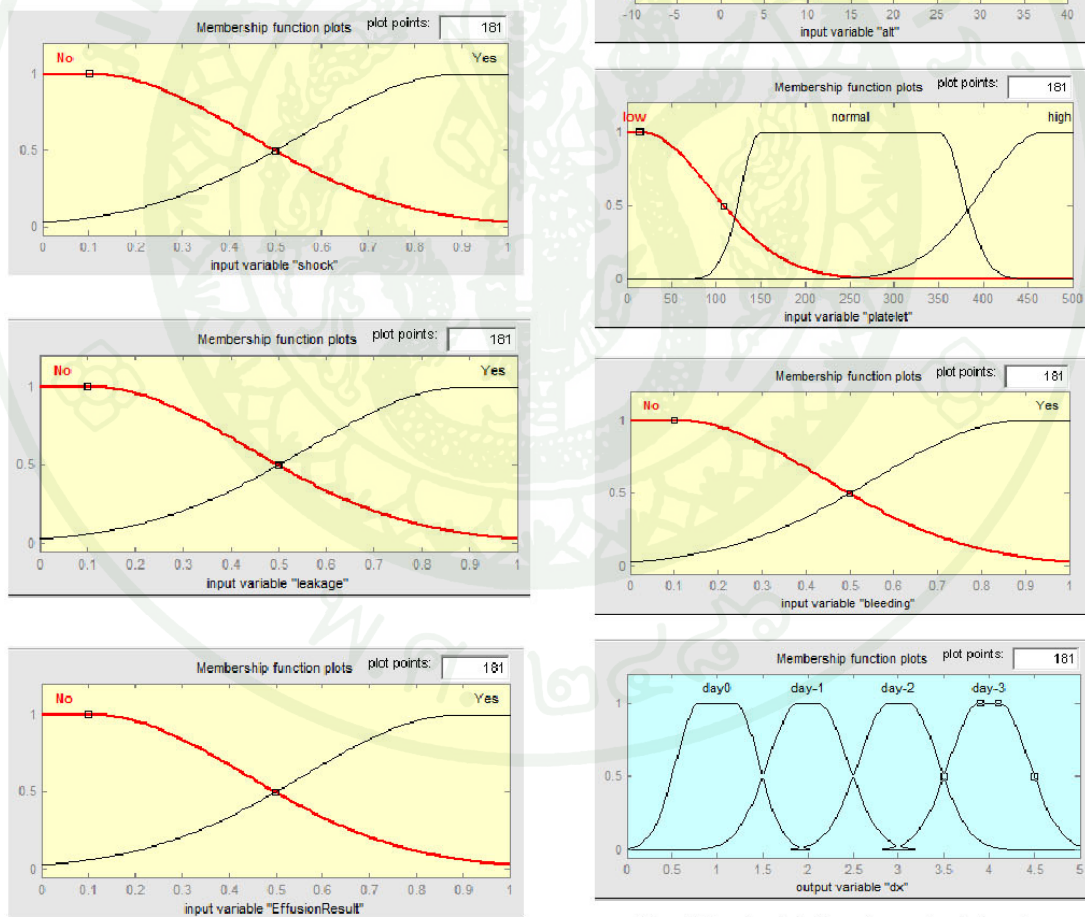


Fig. 2 Membership Functions of each input
in Day0 detection problem

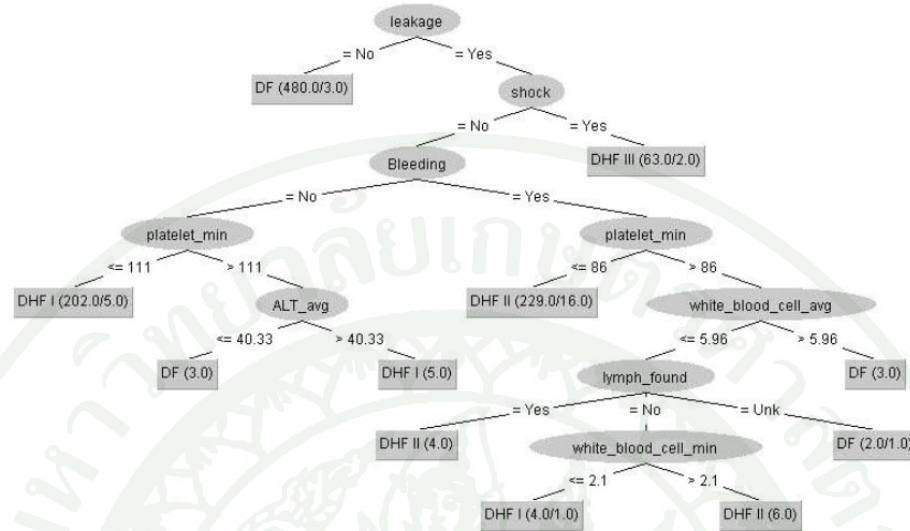


Fig. 3 Decision Tree for Dengue Classification Problem

4. Performance Measurement

We use sensitivity, specificity and accuracy as performance measures. Three equations are defined as following.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Where TP is the number of True Positives, TN is the number of True Negatives instance, FP is the number of False Positives and FN is the number of False Negative.

4.1 Result on Dengue Classification Problem

In this experiment, we applied fuzzy logic and decision tree algorithm in order to compare the performance of both algorithms. We obtained the decision tree as shown in Fig. 3.

We found 8 significant attributes needed to classify patients. These attributes were **leakage** :: leakage of plasma in blood, **shock** :: shock evidence found during treatment period, **Bleeding** :: bleeding evidence found, **platelet_min** :: the minimum of platelet count, **ALT_avg** :: the average of alanine aminotransferase value, **white_blood_cell_avg** :: the average number of white blood cell in laboratory test, **lymph_found** :: evidence of lymphocyte node enlargement, and **white_blood_cell_min** :: the minimum number of white blood cell in laboratory test.

There were three rules found for the DF patients. If there was no leakage evidence, he/she would be diagnose as DF. The

second rule, if there were leakage evidence, no bleeding, the minimum of platelet count more than 111 and the average number of ALT less than or equal 40.33, those patients would be diagnose as DF. However, if there were bleeding evidence, the minimum of platelet count more than 86 and the average number of white blood cell more than 5.96, those patients would be diagnose as DF.

For DHF I, there were three rules. If there were leakage evidence, no shock evidence, no bleeding evidence and the minimum of platelet count less than or equal 111, those patients would be diagnose as DHF I. If the minimum of platelet count more than 111 and the average number of ALT more than 40.33, those patients would be diagnose as DHF I. The third rule, there were bleeding evidence, the minimum of platelet count more than 86, the average number of white blood cell less than or equal 5.96, no lymphocyte node enlargement and the minimum number of white blood cell less than or equal 2.1, those patients would be diagnose as DHF I.

Consider DHF II class; there were three rules. The patients would be diagnosed as DHF II if there was leakage evidence and no shock evidence, and bleeding evidence was found and the minimum of platelet count was less than or equal 86 then those patients would be diagnosed as DHF II. The second rule, if the minimum of platelet count was more than 86, the average value of white blood cell less than or equal 5.96 and lymphocyte node enlargement evidence was found, then those patients would be diagnosed as DHF II. The third rule, if there were no lymphocyte node enlargement evidence and the minimum number of white blood cell more than 2.1, then the patients would be diagnosed as DHF II.

For DHF III class, there was only one rule found. The patient would be diagnosed as DHF III if they found leakage evidence and shock evidence.

We found that the minimum of platelet count, the average number of ALT and the minimum number of white blood cell were correlated with classes. We obtained two attributes which were not correlated with the classes. These attributes were the evidence of lymphocyte node enlargement and the average number of white blood cell. These results also conform to the decision tree.

We found that the decision tree classified patients in each class. As shown in Table 2, the overall accuracy of this model was 96.7 %.

Next, we propose 8 significant attributes to be inputs in fuzzy logic approach. These attributes are shock, leakage, bleeding, the minimum of platelet count, the average of ALT, the minimum of white blood cells, the average of white blood cells and evidence found of lymphocyte node enlargement. The data contained 1001 dengue patients.

As shown in Table 2, Fuzzy logic approach gets 97.39% of overall accuracy.

Table 2 Performance of Fuzzy Logic and Decision Tree on Dengue Classification Problem

Method	Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
Fuzzy Logic	DF	97.75	99.37	98.55	97.39
	DHF I	92.79	96.63	95.77	
	DHF II	90.39	97.51	95.87	
	DHF III	98.39	99.78	99.69	
Decision Tree	DF	97.34	98.56	97.95	96.70
	DHF I	89.64	98.44	96.46	
	DHF II	95.63	96.84	96.56	
	DHF III	98.39	99.55	99.48	

4.2 Day0 Detection Problem

We used daily data obtained from dengue patients without missing value records. We selected the data obtained on the date before day0 and day0 date. The training set was labeled into 4 classes which were day0, day-1, day-2, and day-3. Note that "day0" referred the day of defervescence, whereas "day-1" referred to 1 day before day0 and so on.

First, we selected only significant attributes which had correlation value more than 0.5 from the third experiment. There are bleeding appearance, the number of white blood cell, platelet count, the number of ALT, shock evidence and leakage evidence.

Table 3 The normal range of membership functions

Attribute Name	Normal Range
White Blood Cell (wbc)	4-11 x 10 ³ /ul
Platelet count	150-440 x 10 ³ /ul
Systolic BP – Diastolic BP	25-35
Albumin	3.5-5.5 g/dl
Hematocrit (HCT)	37-52 %

After that, we applied decision tree approach in order to extract the knowledge and assigned the rules in fuzzy logic

approach. We got 64.8% of accuracy as shown in Table 4. The decision tree contained 62 rules to classify the day of defervescence of fever.

The Decision Tree algorithm can correctly classification the day-1 class with the accuracy at 51.56%. For the day-2 patients, we obtained 23.08% correctness. We obtained 64.8 % of overall accuracy.

Next step, we create membership function and assigned the rules. We used Fuzzy Logic Toolbox in MATLAB in order to detect the day0. First, we took the inputs and determine the degree to each of the fuzzy sets via membership functions. We used Gaussian distribution curve of membership function for input and output variables. Then, we created a set of rules by using the rules from decision tree. We transformed the rules of decision tree to the rules of fuzzy. Finally, we obtained the accuracy as shown in Table 4. Note that the sensitivity of day -1 was 71.11% which was better than accuracy from the decision tree. Overall accuracy was also increased to 65.45%.

Table 4 Performance of Fuzzy Logic and Decision Tree on Day0 Detection Problem

Method	Class	Sens(%)	Spec(%)	Acc (%)	Overall Acc(%)
Fuzzy logic	day 0	50.86	75.13	64.88	65.45
	day -1	71.11	45.37	54.75	
	day -2	3.21	96.92	81.82	
	day -3	2.00	99.67	94.63	
Decision Tree	day 0	61.61	57.92	59.82	64.80
	day -1	51.56	54.66	53.43	
	day -2	23.08	85.74	71.11	
	day -3	10.00	96.91	88.79	

5. Discussion and Conclusion

This study proposes a hybrid method which combines a decision tree and fuzzy logic approach in order solve 2 well-known problems of Dengue infection. We found that better performance could be achieved by using the combination of these two approaches.

Since the mechanism of decision tree explores the set of informative attributes. We found some significant features that conform with the knowledge of physician. These attributes are shock evidence, leakage evidence, bleeding evidence and the minimum of platelet count.

For the Day0 detection problem, we use the significant features obtained from the decision tree. Then, we create decision tree with J48 and got the knowledge of day0 detection. The knowledge was created the rules for fuzzy logic approach. We obtained better accuracy in day-1 which was 71.11% (see Table 4. for details). Note that the class day-1 is the target class because the physician needs to know in advance before the fatal condition occurs on the patient. However, the performance of day-2 and day-3 class is very low for fuzzy logic because missing value is found in the original data. We know that fuzzy logic takes advantage from the continuous value through the

membership function. From the medicine literature [Green, 1999] the main feature that determines the Dengue severity is the plasma leakage which can be measured in term of the Pleural effusion index. However most patients from the dataset were not test with this measure.

Acknowledgements

This research was supported by the grant from Kasetsart University Research and Development Institute (KURDI). We would like to thank Department of Computer Science, Faculty of Science and Kasetsart University for their support.

References

- [Ibrahim 2005] F. Ibrahim, M. N. Taib, W. A. B. Wan Abas, C. G. Chan and S. Sulaiman, A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), Computer Methods and Programs in Biomedicine, No.79, 2005, pp. 273-281.
- [Faisal 2010] T. Faisal, F. Ibrahim and M.N. Taib, A noninvasive intelligent approach for predicting the risk in dengue patients, Expert Systems with Application, Vol.37, No.3, 2010, pp. 2175-2181.
- [Green 1999] S. Green, D. W. Vaughn, S. Kalayanaroj, S. Nimmannitya, S. Suntayakorn, A. Nisalak, A. L. Rothman, and F. A. Ennis, Elevated Plasma Interleukin-10 Levels in Acute Dengue Correlate With Disease Severity, Journal of Medical Virology, 1999.
- [Roger Jang 1997] JSR Jang and N. Gulley, MATLAB: Fuzzy Logic Toolbox User's Guide, The MathWorks Inc., 24 Prime Park Way, Natick, Mass. 1997.
- [Tanner 2008] L. Tanner, M. Schreiber, J.G. Low, A. Ong, T. Tolfvenstam, Y.L. Lai, L.C. Ng, Y.S. Leo, L. Thi Puong, S.G. Vasudevan, C.P. Simmons, M.L. Hibberd and E.E. Ooi, Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness, PLoS Neglected Tropical Disease, Vol.2, 2008.
- [Thitiprayoonwongse 2011] D. Thitiprayoonwongse, P. Suriyaphol, and N. Soonthornphisaj, Data Mining on Dengue Virus Disease, 13th International Conference on Enterprise Information Systems (ICEIS 2011), No.1, 2011, pp. 32-41.
- [WHO 1999] World Health Organization, Guideline for Treatment of Dengue Fever/Dengue Haemorrhagic Fever, 1999.

ประวัติการศึกษาและการทำงาน

ชื่อ-นามสกุล	นางสาวดารณี ฐิติประยูรวงศ์
วัน เดือน ปีที่เกิด	28 พฤศจิกายน 2528
สถานที่เกิด	อำเภอเมือง จังหวัดนครราชสีมา
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) (เกียรตินิยมอันดับสอง) มหาวิทยาลัยขอนแก่น
ทุนการศึกษาที่ได้รับ	ทุนผู้ช่วยสอนจากบัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์ (พ.ศ. 2552 – 2554)