



ใบรับรองวิทยานิพนธ์
บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตร์มหาบัณฑิต (สถิติ)

ปริญญญา

สถิติ

สถิติ

สาขา

ภาควิชา

เรื่อง การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบทางเดินหายใจ

Applications of Decision Tree Algorithms with Diagnosis of the Respiratory System

นามผู้วิจัย นายดิษฐ์พล มั่นธรรม

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ลีลี อิงศรีสว่าง, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(รองศาสตราจารย์ปรีชา นุช อภิณ โยภาส, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(อาจารย์สุภาวดี อิงศรีสว่าง, Ph.D.)

หัวหน้าภาควิชา

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญญา วีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่

เดือน

พ.ศ.

วิทยานิพนธ์

เรื่อง

การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบทางเดินหายใจ

Applications of Decision Tree Algorithms with Diagnosis of the Respiratory System

โดย

นายดิษฐ์พล มั่นธรรม

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)

พ.ศ. 2553

ลิขสิทธิ์ มหาวิทยาลัยเกษตรศาสตร์

ดิษฐพล มั่นธรรม 2553: การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรกระบบทางเดิน
หายใจ ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาวิชาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษา
วิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ลีลี อิงศรีสว่าง, Ph.D. 126 หน้า

การศึกษานี้มีวัตถุประสงค์ 1. เพื่อประยุกต์วิธีสืบค้นความรู้จากฐานข้อมูล โดยใช้อัลกอริทึมต้นไม้
ตัดสินใจกับการวินิจฉัยโรกระบบทางเดินหายใจ ในการจำแนกผู้ป่วย 3 โรค คือ โรคการติดเชื้อทางเดินหายใจ
ส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ของโรงพยาบาลพระนครศรีอยุธยา
และ 2. เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ 3 อัลกอริทึม ได้แก่ ID3 C4.5 และ CART
ในการจำแนกหรือคัดกรองผู้ป่วยทั้ง 3 โรค ข้อมูลที่ใช้ศึกษาเป็นข้อมูลเวชระเบียนผู้ป่วยนอกโรกระบบทางเดิน
หายใจ ของโรงพยาบาลพระนครศรีอยุธยา ระหว่างปี พ.ศ.2547 - 2549 จำนวน 7,327 คน โดยตัวแปรที่นำมา
พิจารณาประกอบด้วย อายุ อุณหภูมิร่างกาย เขตที่อยู่อาศัย อาชีพ อาการต่างๆ เช่น มีน้ำมูก มีไข้ คัดจมูก
แน่นหน้าอก ปวดกระบอกตา ปวดศีรษะ หายใจเหนื่อยหอบ และไอ เป็นต้น วิธีการศึกษา ทำการสืบค้น
ความรู้จากฐานข้อมูลผู้ป่วยโรกระบบทางเดินหายใจด้วยอัลกอริทึม ID3 C4.5 และ CART และเปรียบเทียบ
ประสิทธิภาพของอัลกอริทึมทั้งสาม โดยมีการคัดเลือกตัวแปรที่มีความสำคัญต่อการจำแนกผู้ป่วยแต่ละโรค
ด้วยอัลกอริทึม Best First Greedy Stepwise และ Genetic Search แล้วจึงนำตัวแปรที่ได้มาผ่านอัลกอริทึม
ต้นไม้ตัดสินใจข้างต้น เพื่อจำแนกผู้ป่วยทั้ง 3 โรค ด้วยระบบการเรียนรู้และตรวจสอบความถูกต้องของการ
จำแนกผู้ป่วยแต่ละโรคด้วยการแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบตามหลัก Cross-
validation และ Percentage split

ผลการศึกษา วิธีการสืบค้นความรู้สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบ
เฉียบพลัน พบว่าการใช้ตัวแปรที่คัดเลือกได้เพียง 7 ตัวแปร กับอัตราส่วนข้อมูลฝึกสอนต่อข้อมูลทดสอบ
70 : 30 อัลกอริทึม C4.5 ให้ผลค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้อง 92.32% ค่าความไว 92.40%
ค่าพยากรณ์บวก 95.63% ค่าพยากรณ์ลบ 90.18% และค่า AUC 96.20% ส่วนการสืบค้นความรู้สำหรับ
จำแนกผู้ป่วยโรคปอดอักเสบ พบว่าการใช้ตัวแปรที่คัดเลือกได้เพียง 8 ตัว และใช้อัตราส่วนข้อมูลฝึกสอนต่อ
ข้อมูลทดสอบ 70 : 30 อัลกอริทึม C4.5 ให้ค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้อง 94.70% ค่าความ
ไว 93.42% ค่าพยากรณ์บวก 83.24% ค่าพยากรณ์ลบ 98.16% และค่า AUC 92.20% และการสืบค้นความรู้
สำหรับจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน พบว่าตัวแปรที่คัดเลือกได้เพียง 7 ตัว และใช้อัตราส่วนข้อมูล
ฝึกสอนต่อข้อมูลทดสอบ 50:50 อัลกอริทึม CART ให้ค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้อง 94.69%
ค่าความไว 100% ค่าพยากรณ์บวก 81.57% ค่าพยากรณ์ลบ 100% และค่า AUC 94.70% ซึ่งผลลัพธ์ที่ได้จาก
การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจ สามารถนำไปเป็นแนวทางสนับสนุนการคัดกรองผู้ป่วยโรกระบบ
ทางเดินหายใจเบื้องต้นเพื่อการวินิจฉัยต่อไป

ลายมือชื่อนิติ

ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Ditapol Muntham 2010: Applications of Decision Tree Algorithms with Diagnosis of the Respiratory System. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Lily Ingsrisawang, Ph.D. 126 pages.

The objectives of this study involved: 1) the application of methods of knowledge discovery from data base using decision tree algorithms for Respiratory System Diagnosis to classify patient at Pranakornsri Ayudthaya Hospital into three groups: acute URI, acute sinusitis, and pneumonia, and 2) the comparison of the performance of three decision tree algorithms : ID3, C4.5, and CART, for the classification or screening the patients with the three diseases. The data used in this study came from the medical records of 7,327 out-patients with respiratory diseases at Pranakornsri Ayudthaya Hospital in the period of 2003 to 2006. The variables considered were age, body temperature, residence, occupation, and some symptoms such as rhinorrhea, fever, nasal congestion, periorbital pain, headache, wheezing and coughing. The study methods were the knowledge discovery with the employment of ID3, C4.5, and CART decision tree algorithms from hospital's patient registration database and the effective comparison of the three algorithms. The algorithms used for the important variable selection were Best First, Greedy Stepwise, and Genetic Search. The decision tree algorithms were employed to classify the patients with those three diseases into three groups. The validity of the decision tree algorithms was studied by the division of the data into two sets, a training and a testing data sets, which were based on the cross-validation and the percentage split methods.

The study of the methods of knowledge discovery for the classification found that, for the patients with acute URI with only 7 selected variables and the ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective. The classification accuracy was 92.31%; the sensitivity was 90.60%; the PPV was 95.63%; the NPV was 90.18%; and the AUC was 96.20%. For the classification of the patients with acute sinusitis with only 8 selected variables and the ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective. The classification accuracy was 94.70%; the sensitivity was 93.42%; the PPV was 83.24; the NPV was 98.16%; and the AUC was 92.20%. For the classification of the patients with pneumonia with only 7 selected variables and the ratio 50:50 of the training data set and the testing data set, the CART algorithm was the most effective. The classification accuracy was 94.69%; the sensitivity was 100%; the PPV was 81.57; the NPV was 100%; and the AUC was 94.70%. The obtained results could be used for the support of the diagnosis of patients with Respiratory System disease.

Student's signature

Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ โดยคำแนะนำ การสนับสนุนให้ข้อคิดเห็นและแก้ไขข้อบกพร่องต่างๆ จาก ผู้ช่วยศาสตราจารย์ ดร.ลีลี อิงศรีสว่าง อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รองศาสตราจารย์ ดร.ปริยานุช อภิบุณโยภาส ดร.สุภาวดี อิงศรีสว่าง อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม รองศาสตราจารย์สายสุดา สมจิต ประธานกรรมการสอบวิทยานิพนธ์ และ ผู้ช่วยศาสตราจารย์ ดร.นิภา โรจน์รุ่งวศินกุล ผู้ทรงคุณวุฒิภายนอก ที่กรุณาให้คำปรึกษาแนะนำ และให้การช่วยเหลือทำให้วิทยานิพนธ์สำเร็จลุล่วงไปด้วยดี

นอกจากนี้ผู้วิจัยขอขอบคุณ นายแพทย์วีระพล ชีระพันธ์เจริญ ผู้อำนวยการโรงพยาบาลพระนครศรีอยุธยา คุณพรทิพย์ รัตนวิชัย รองผู้อำนวยการฝ่ายการพยาบาล นายแพทย์สุรชัย โชคครรชิตไชย หัวหน้ากลุ่มพัฒนาระบบสุขภาพ และคุณภิรมย์ คำภิธัมโม หัวหน้างานเวชระเบียนและสถิติ รวมถึงเจ้าหน้าที่ห้องเวชระเบียนและสถิติทุกท่านที่ช่วยจัดหาข้อมูลและให้การสนับสนุนในการศึกษานี้เป็นอย่างยิ่ง

ขอขอบคุณ นายแพทย์สรรรพล วุฒิภักดากร วิทยาลัยแพทยศาสตร์กรุงเทพมหานครและวชิรพยาบาล ที่ให้คำปรึกษา คำแนะนำ ตลอดจนแก้ไขในส่วนแนวทางการซักประวัติผู้ป่วยโรคทางระบบหายใจ และความรู้เบื้องต้นเกี่ยวกับโรคระบบทางเดินหายใจ

ขอขอบคุณ คุณพ่อ คุณแม่ ที่คอยสนับสนุน คอยให้กำลังใจและความเข้าใจทุกอย่าง และขอขอบคุณ นางสาวชลพรช พันธ์พานิชย์ เพื่อนที่เป็นกำลังและเข้าใจเสมอมา ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์เล่มนี้จนสำเร็จลุล่วงด้วยดี

ประโยชน์อันใดที่เกิดเนื่องมาจากวิทยานิพนธ์เล่มนี้ ขอมอบแต่คุณพ่อ คุณแม่ และคณาจารย์ทุกท่าน ที่ได้เมตตา อบรมสั่งสอน ประสิทธิ์ประสาทวิชาความรู้ให้แก่ข้าพเจ้า

ดิษฐ์พล มั่นธรรม

ตุลาคม 2552

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(6)
คำนำ	1
วัตถุประสงค์	5
การตรวจเอกสาร	8
อุปกรณ์และวิธีการ	38
อุปกรณ์	38
วิธีการ	38
ผลและวิจารณ์	47
ผล	47
วิจารณ์	76
สรุปผลและข้อเสนอแนะ	78
สรุป	78
ข้อเสนอแนะ	84
เอกสารและสิ่งอ้างอิง	85
ภาคผนวก	90
ประวัติการศึกษาและการทำงาน	126

สารบัญตาราง

ตารางที่		หน้า
1	ตัวอย่างการประยุกต์ใช้ต้นไม้ตัดสินใจในการวิเคราะห์สภาพอากาศ สำหรับการออกไปเล่นกีฬานอกบ้าน	11
2	ตัวอย่างข้อมูลของปัจจัยที่ทำให้ผิวหนังไหม้	16
3	การตรวจสอบความสามารถของตัวจำแนกที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจ ในการวินิจฉัยการเป็นโรคของผู้ป่วย	27
4	การแปลความหมายค่า Kappa statistics	28
5	การแปลความหมายค่า AUC	31
6	รายละเอียดของตัวแปรและการกำหนดรหัสที่ศึกษา	39
7	จำนวนและร้อยละของผู้ป่วยกลุ่มโรคระบบทางเดินหายใจที่เข้ารับการรักษา ครั้งแรกด้วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา ปี พ.ศ. 2547 – 2549	48
8	จำนวนและร้อยละของลักษณะอาการของผู้ป่วยที่เข้ามารับการรักษาครั้งแรก ด้วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา ระหว่างปี พ.ศ. 2547 – 2549	50
9	เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับ จำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตาม วิธีการแบ่งชุดข้อมูล	53
10	เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับ จำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูล	55
11	เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับ จำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูล	57

สารบัญตาราง (ต่อ)

ตารางที่		หน้า
12	การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี	59
13	การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคไซนัสอักเสบเฉียบพลัน จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี	60
14	การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคปอดอักเสบไม่ระบุเชื้อ จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี	61
15	สรุปจำนวนตัวแปรที่คัดเลือกได้จากอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search จำแนกตามโรค และการแบ่งชุดข้อมูล	62
16	เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่ใช้การคัดเลือกตัวแปร สำหรับการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ	63
17	เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่ใช้การเลือกตัวแปร สำหรับการจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ	65
18	เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่มีการคัดเลือกตัวแปร สำหรับการจำแนกผู้ป่วยโรคปอดอักเสบ ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ	67
19	เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีไม่มีการคัดเลือกตัวแปรและกรณีมีการคัดเลือกตัวแปร ในการจำแนกผู้ป่วยกลุ่มโรคระบบทางเดินหายใจใน 3 โรค	82
20	ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีคัดเลือกตัวแปร กับจำนวนตัวแปรในการจำแนกผู้ป่วยกลุ่มโรคระบบทางเดินหายใจใน 3 โรค	83

สารบัญตาราง (ต่อ)

ตารางผนวกที่	หน้า
1 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการคัดเลือกตัวแปร ด้วยวิธี Genetic Search กับข้อมูลแบบ Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	91
2 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรแบบ Best First และ Greedy Stepwise เมื่อใช้การแบ่งข้อมูลแบบ Use full training set หรือ Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	92
3 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรด้วยวิธี Genetic Search จาก Use full training set สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	94
4 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการคัดเลือกตัวแปรด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูลที่แบ่งแบบ Cross – validation สำหรับจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	95
5 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูล Cross – validation สำหรับจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	97
6 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่มีการเลือกตัวแปรด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูล Cross – validation สำหรับจำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ	98

สารบัญตาราง (ต่อ)

ตารางผนวกที่

หน้า

7	ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ ด้วย Best First, Greedy Stepwise และ Genetic Search จากการใช้ Use full training set โดย สำหรับ จำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูล	100
---	---	-----

สารบัญภาพ

ภาพที่	หน้า
1 ขั้นตอนของกระบวนการ KDD	9
2 ต้นไม้ตัดสินใจสำหรับจำแนกการไปออกกำลังกายตามเงื่อนไขสภาพอากาศต่างๆ	12
3 รูปแบบของต้นไม้ตัดสินใจ	13
4 ต้นไม้ตัดสินใจที่ได้จากการใช้โปรแกรม Weka กับชุดข้อมูลปจจัยที่ทำให้ผิวหนังไหม้	18
5 ตัวอย่างการเปรียบเทียบค่า AUC	30
6 แสดงการแบ่งข้อมูลเพื่อการเรียนรู้การสร้างตัวจำแนกแบบ 5 – Folds Cross-validation	42
7 ขั้นตอนการพัฒนาตัวจำแนกโรกระบบทางเดินหายใจ ด้วยอัลกอริทึมต้นไม้ตัดสินใจ ID3, C4.5 และ CART กรณีไม่มีการคัดเลือกตัวแปร	44
8 ขั้นตอนการพัฒนาตัวจำแนกโรกระบบทางเดินหายใจ ด้วยอัลกอริทึมต้นไม้ตัดสินใจ ID3, C4.5 และ CART กรณีมีการคัดเลือกตัวแปร	45
9 ต้นไม้ตัดสินใจสำหรับการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลันด้วยอัลกอริทึม C4.5	69
10 ต้นไม้การจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลันด้วยอัลกอริทึม CART	72
11 ต้นไม้ตัดสินใจสำหรับการจำแนกผู้ป่วยโรคปอดอักเสบด้วยอัลกอริทึม C4.5	74

ภาพผนวกที่

1 ต้นไม้การจำแนก สำหรับอัลกอริทึม CART กรณีที่มีการเลือกตัวแปรด้วยวิธี Genetic Search จาก Use full training set และสร้างต้นไม้การจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI) จากข้อมูลที่แบ่งแบบ 5 Folds Cross-validation	125
--	-----

คำอธิบายสัญลักษณ์และคำย่อ

AUC	=	Area under the Receiver Operator Characteristic curve
Acute Sinusitis	=	Acute Sinusitis
Acute URI	=	Acute Upper Respiratory Infection
CART	=	Classification and Regression Trees
FP Rate	=	False Positive Rate
HA	=	Hospital Accreditation
ID3	=	Iterative Dichotomizer 3
KDD	=	Knowledge Discovery in Database
NPV	=	Negative Predictive Value
PPV	=	Positive Predictive Value
Pneumonia	=	Other Pneumonia
TP Rate	=	True Positive Rate

การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบทางเดินหายใจ

Applications of Decision Tree Algorithms with Diagnosis of the Respiratory System

คำนำ

ปัจจุบันการสืบค้นความรู้ที่เป็นประโยชน์จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Database : KDD) กำลังได้รับความนิยมโดยอาศัยวิธีการเหมืองข้อมูล (Data mining) (Fayyad *et al.*, 1996) ซึ่งเป็นศาสตร์ทางด้านสถิติและคอมพิวเตอร์ สำหรับจัดการความรู้ที่สำคัญให้อยู่ในฐานความรู้ (Knowledge Base) ของการสืบค้นต่อไป เช่น การวิเคราะห์ความสัมพันธ์ของข้อมูลหรือการทำนายเหตุการณ์ที่จะเกิดขึ้นในอนาคต การวิเคราะห์ความเสี่ยงของลูกค้านิติบุคคล การพยากรณ์ปริมาณยอดขายสินค้า การแก้ปัญหาทางวิศวกรรม การรักษาทางการแพทย์ เป็นต้น

KDD ประกอบด้วยเทคนิคต่างๆ เช่น การจำแนกข้อมูล (Data Classification) การจัดกลุ่มข้อมูล (Data Clustering) การค้นหากฎความสัมพันธ์ (Association Rule Discovery) และการวิเคราะห์อนุกรมเวลา (Time-Series Analysis) เป็นต้น ซึ่งแต่ละเทคนิคมีความเหมาะสมกับงานต่างประเภทกัน (Larose, 2004)

Data Classification สามารถใช้กลไกการเรียนรู้ต่อไปนี้ เช่น Decision Tree, Bayesian Classification, Neural Network, Genetic Algorithm, Cased-based Reasoning, Rough Set Theory, Fuzzy Sets, Association-based, Classification, Naive-Bayes และ K-Nearest Neighbor เป็นต้น ไปประยุกต์กับงานด้านต่างๆ เช่น การประกันภัย การจัดกลุ่มลูกค้า การจัดวางสินค้า เป็นต้น แต่กลไกการเรียนรู้ที่ได้รับความนิยมสูงสุด คือ Decision Tree (Larose, 2004)

Decision Tree หรือ ต้นไม้ตัดสินใจ เป็นกลไกการเรียนรู้ที่ใช้กระบวนการสร้างตัวแบบ (Model) และจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด ซึ่งต้นไม้ตัดสินใจมีความสามารถในการเรียนรู้พฤติกรรมของข้อมูลและสร้างตัวจำแนก (Classifies) ได้เร็วกว่ากลไกวิธีอื่นๆ โดยต้นไม้ตัดสินใจมีหลายอัลกอริทึมที่ใช้ในการจำแนกข้อมูล เช่น ID3 (Iterative Dichotomizer 3), Chi-Squared

Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 และ C5.0 algorithm เป็นต้น และแต่ละอัลกอริทึมก็จะให้โครงสร้างต้นไม้ตัดสินใจที่แตกต่างกัน (Larose, 2004) อัลกอริทึมที่มีการนำไปใช้มาก ได้แก่ ID3, C4.5 และ CART (Teknomo, 2009)

ID3 : เป็นอัลกอริทึมต้นไม้ตัดสินใจที่ง่าย โดยใช้ค่าเกน (Gain) ของข้อมูลเป็นเงื่อนไขในการแบ่งกลุ่ม ลักษณะข้อมูลที่ให้กับ ID3 ต้องเป็นข้อมูลเชิงกลุ่มและไม่มีค่าสูญหาย

C4.5 : เป็นอัลกอริทึมที่ได้รับการพัฒนาจาก ID3 โดยใช้ค่า Gain Ratio ของข้อมูลเป็นเงื่อนไขในการแบ่งกลุ่ม สามารถใช้กับข้อมูลทั้งที่มีลักษณะเชิงกลุ่มและเชิงปริมาณ และสามารถจัดการกับกรณีที่มีค่าสูญหายได้ เป็นการแก้ไขจุดด้อยของ ID3

CART : เป็นสถิติอนพารามิเตอร์ ที่ใช้ในการวิเคราะห์ในการจำแนกกลุ่ม สามารถแสดงภาพต้นไม้การจำแนกได้ ใช้ได้ทั้งข้อมูลที่มีลักษณะเชิงกลุ่มและเชิงปริมาณ แต่ CART มีข้อจำกัดในการสร้างต้นไม้การจำแนก ซึ่งสามารถสร้างต้นไม้การจำแนกแบบทวิภาค (Binary Tree) ได้เท่านั้น (Lior and Oded, 2008)

ต้นไม้ตัดสินใจถูกนำมาใช้ในงานด้านการแพทย์ เช่น Kamila (2007) ได้ศึกษาการเลือกใช้อัลกอริทึมเหมืองข้อมูลที่เหมาะสมสำหรับระบบสนับสนุนการตัดสินใจทางการแพทย์ โดยใช้ อัลกอริทึมต้นไม้ตัดสินใจ (ID3, C4.5) อัลกอริทึมโครงข่ายประสาทเทียม และอัลกอริทึม Naïve Bayes กับตัวอย่างศึกษา ได้แก่ ผู้ป่วยโรคเบาหวาน จำนวน 768 คน ผู้ป่วยโรคมะเร็งเต้านม จำนวน 699 คน ผู้ป่วยโรคตับอักเสบ จำนวน 155 คน ผู้ป่วยโรคหัวใจ จำนวน 303 คน และ ผู้ป่วยโรคผิวหนัง จำนวน 366 คน ปรากฏว่า อัลกอริทึม Naïve Bayes ให้ค่าความถูกต้องของการจำแนกผู้ป่วยในกลุ่มโรคเบาหวาน โรคผิวหนัง และโรคหัวใจ ได้สูงสุดถึงร้อยละ 76.30 97.40 และ 55.50 ตามลำดับ สำหรับอัลกอริทึมต้นไม้ตัดสินใจ C4.5 ให้ค่าความถูกต้องของการจำแนกผู้ป่วยโรคมะเร็งเต้านม และโรคตับอักเสบได้ถึงร้อยละ 96.00 และ 79.10 หรือในการศึกษาของ Porcel *et al.* (2008) ที่ประยุกต์อัลกอริทึมต้นไม้ตัดสินใจ C4.5 เพื่อจำแนกผู้ป่วยโรคผิวหนัง และผู้ป่วยโรคเชื้อหุ่มปอดอักเสบ ใช้กลุ่มตัวอย่างทั้งหมด 238 คน ที่เป็นผู้ป่วยโรคผิวหนัง จำนวน 64 คน และผู้ป่วยโรคเชื้อหุ่มปอดอักเสบ จำนวน 174 คน ตรวจสอบความถูกต้องของตัวแบบ ด้วยวิธีการ 10-Folds Cross-validation ผลการศึกษาพบว่า อัลกอริทึม C4.5 ให้ค่าความไว (Sensitivity) ร้อยละ 92.20 ค่าความจำเพาะ (Specificity) ร้อยละ 98.30 และค่า ROC 0.976 แสดงว่า

อัลกอริทึมต้นไม้ตัดสินใจสามารถช่วยในการจำแนกกลุ่มผู้ป่วยโรคหัวใจ และผู้ป่วยโรคเชื้อหุ้มปอดอักเสบได้ดี

สำหรับในประเทศไทยก็ได้มีการนำเทคนิคต้นไม้ตัดสินใจมาใช้ในการแพทย์เช่นกัน ดังเช่น พุทธิ (2551) ที่ทำการศึกษาระบบจำแนกประเภทแบบทดสอบสำหรับผู้ทดสอบสุขภาพจิตด้วยเทคนิคต้นไม้ตัดสินใจ โดยนำเทคนิคต้นไม้ตัดสินใจมาจำแนกประเภทแบบทดสอบที่เหมาะสมสำหรับผู้ทดสอบ โดยใช้กลุ่มตัวอย่าง 550 คน ผลจากการใช้ต้นไม้ตัดสินใจ พบว่าสามารถจำแนกได้ทั้งสิ้นจำนวน 15 ตัวแบบ และตัวแบบที่ได้จากการศึกษาครั้งนี้ได้ถูกนำมาเป็นองค์ความรู้เพื่อใช้ในการพัฒนาระบบต่อไป หรือในการศึกษาของ วราภรณ์ (2551) ที่ทำการวิเคราะห์ปัจจัยเสี่ยงและการจัดจำแนกกลุ่มของการดื้อยาในผู้ป่วยวัณโรคปอดที่เกิดโรคกลับด้วยการวิเคราะห์การถดถอยโลจิสติก และการวิเคราะห์ Classification Tree โดยศึกษาในกลุ่มผู้ป่วยจำนวน 196 ราย ผลการศึกษาพบว่า การวิเคราะห์การถดถอยโลจิสติก มีค่าความไวร้อยละ 24.10 ค่าความจำเพาะร้อยละ 95.50 ค่าพยากรณ์บวกร้อยละ 70.00 และค่าพยากรณ์ลบร้อยละ 74.10 สำหรับการวิเคราะห์ Classification Tree ให้ค่าความไวร้อยละ 16.70 ค่าความจำเพาะร้อยละ 98.50 ค่าพยากรณ์บวกร้อยละ 83.30 และค่าพยากรณ์ลบร้อยละ 72.80 เมื่อพิจารณาค่าความไว ค่าความจำเพาะ และค่าพยากรณ์ลบ ของการวิเคราะห์การถดถอยโลจิสติก และการวิเคราะห์ Classification Tree ให้ค่าใกล้เคียงกัน แต่สำหรับค่าพยากรณ์บวก พบว่า การวิเคราะห์ Classification Tree ให้ค่าที่สูงกว่าการวิเคราะห์การถดถอยโลจิสติก

จากงานวิจัยข้างต้นแสดงให้เห็นว่ามีการประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับผู้ป่วยโรคต่างๆ แต่ในประเทศไทยยังไม่มีการใช้ที่แพร่หลายนัก ดังนั้นผู้วิจัยจึงเห็นว่าควรจะประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับข้อมูลผู้ป่วยในรายโรคที่พบบ่อยในประเทศ จึงได้ศึกษาจากรายงานการเฝ้าระวังโรคประจำปี 2546-2548 (สำนักโรคบาวิทยา, 2550) พบว่ากลุ่มโรคระบบทางเดินหายใจมีอัตราการป่วยสูงขึ้น โดยมีอัตราการป่วยต่อแสนประชากรเท่ากับ 308.21, 311.80 และ 315.33 ในปี 2546, 2547 และ 2548 ตามลำดับ และในการประชุมโรคระบบการหายใจแห่งภาคพื้นเอเชีย-แปซิฟิก ครั้งที่ 13 พ.ศ.2551 นายแพทย์อรรถ นานา นายกสมาคมอุรเวชช์แห่งประเทศไทยได้แสดงความคิดเห็นว่า “ในปัจจุบันจะพบว่าโรคต่างๆ ที่เกี่ยวกับระบบการหายใจเพิ่มขึ้น โดยเฉพาะโรคที่เกี่ยวกับการทำงาน เนื่องจากการพัฒนาประเทศไปสู่ประเทศอุตสาหกรรมและกึ่งอุตสาหกรรมทำให้เกิดผลกระทบต่อสุขภาพของผู้ทำงานและผู้ประกอบการอย่างยากที่จะหลีกเลี่ยง เช่น โรคปอดอุดกั้นเรื้อรัง โรคภูมิแพ้ โรคหอบหืด และวัณโรค ซึ่งเป็นโรคที่อยู่ในกลุ่มโรคระบบทางเดินหายใจทั้งสิ้น” (สรพวงศ์, 2551) และจากรายงานจังหวัดในประเทศไทยที่มี

นิคมอุตสาหกรรม ได้แก่ กรุงเทพมหานคร สมุทรสาคร สมุทรปราการ ราชบุรี สระบุรี ลำพูน พิจิตร ขอนแก่น ฉะเชิงเทรา ชลบุรี ระยอง ปัตตานี และพระนครศรีอยุธยา พบว่าจังหวัดพระนครศรีอยุธยาถือได้ว่าเป็นจังหวัดที่มีนิคมอุตสาหกรรมเป็นอันดับ 1 ของภาคกลาง มีทั้งสิ้น 4 นิคมอุตสาหกรรม ได้แก่ สวนนิคมอุตสาหกรรมโรจนะ นิคมอุตสาหกรรมไฮเทค นิคมอุตสาหกรรมบางปะอิน และนิคมอุตสาหกรรมสหรัษฎนคร (การนิคมอุตสาหกรรมแห่งประเทศไทย, 2551) และมีรายงานสรุปจำนวนผู้ป่วยนอกฝ่ายเวชระเบียนและสถิติ โรงพยาบาลพระนครศรีอยุธยา ประจำปี 2547 - 2549 พบว่า กลุ่มโรคระบบทางเดินหายใจ (Diseases of the Respiratory System) เป็นกลุ่มที่มีจำนวนผู้ป่วยเข้ามารับการรักษามากที่สุดเป็นอันดับ 1 โดยมีอัตราการเข้ารับการรักษาคิดเป็นร้อยละ 14.47 15.13 และ 15.73 ตามลำดับ โดยโรคที่เข้ารับการรักษามากที่สุด 3 อันดับ ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute Upper Respiratory Infection) โรคไซนัสอักเสบเฉียบพลัน (Acute Sinusitis) และโรคปอดอักเสบ (Pneumonia) ตามลำดับ และมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง ทำให้อัตราค่าเฉลี่ยของแพทย์ที่มีอยู่อย่างจำกัดไม่เพียงพอต่อจำนวนผู้ป่วยที่มีมากขึ้น ส่งผลถึงการพัฒนาและรับรองคุณภาพโรงพยาบาล (HA)

ด้วยเหตุนี้ การศึกษานี้จึงสนใจที่จะประยุกต์อัลกอริทึมต้นไม้ตัดสินใจ 3 อัลกอริทึม ได้แก่ ID3 C4.5 และ CART กับการวินิจฉัยโรคระบบทางเดินหายใจ เพื่อจำแนกผู้ป่วย 3 โรค คือ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ เพื่อนำผลที่ได้ไปใช้สนับสนุนหรือเป็นแนวทางการคัดกรองผู้ป่วยในเบื้องต้นเพื่อการวินิจฉัยต่อไป

วัตถุประสงค์

1. เพื่อประยุกต์วิธีสืบค้นความรู้จากฐานข้อมูล โดยใช้อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบทางเดินหายใจ ในการจำแนกผู้ป่วย 3 โรค ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ

2. เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ 3 อัลกอริทึม ได้แก่ ID3 C4.5 และ CART ในการจำแนกผู้ป่วย 3 โรค ข้างต้น

ขอบเขตการวิจัย

1. ข้อมูลที่นำมาศึกษาเป็นข้อมูลเวชระเบียนผู้ป่วยนอกโรคระบบทางเดินหายใจ ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา โดยศึกษาเฉพาะผู้ที่เข้ามารับการรักษาในครั้งแรก ตั้งแต่เดือนมกราคม พ.ศ. 2547 ถึงเดือนธันวาคม พ.ศ. 2549
ตัวแปรที่ศึกษามีทั้งหมด 22 ตัวแปร

- | | |
|--------------------|-------------------------------------|
| 1) เพศ | 12) อาการคัดจมูก |
| 2) อายุ | 13) อาการอาเจียน |
| 3) ที่อยู่อาศัย | 14) อาการหายใจเหนื่อยหอบ/หายใจลำบาก |
| 4) อาชีพ | 15) อาการหายใจมีเสียงหวีดๆ |
| 5) อุณหภูมิร่างกาย | 16) อาการหายใจมีกลิ่นเหม็น |
| 6) อาการไอ | 17) อาการแน่นหน้าอก |
| 7) อาการเวียนศีรษะ | 18) อาการเจ็บหน้าอก |
| 8) อาการเจ็บคอ | 19) อาการปวดท้อง |
| 9) มีเสมหะ | 20) อาการปวดกระบอกตา |
| 10) อาการไอ | 21) อาการปวดจมูก |
| 11) มีน้ำมูก | 22) อาการปวดหู |

2. อัลกอริทึมต้นไม้ตัดสินใจที่ศึกษาเฉพาะ 3 อัลกอริทึม คือ ID3 C4.5 และ CART

3. โรคระบบทางเดินหายใจที่ทำการศึกษาทั้ง 3 โรค ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ทั้ง 3 โรคเป็นแบบไม่ระบุเชื้อ

นิยามศัพท์เฉพาะ

1. อุณหภูมิปกติ หมายถึง อุณหภูมิที่ทำการวัดในช่วงเวลา 6.00 น. จะอยู่ในช่วง 36.80 ± 0.4 องศาเซลเซียส ซึ่งความแตกต่างของอุณหภูมิของร่างกายนั้นขึ้นอยู่กับเวลาที่วัด โดยในการวัดอุณหภูมิช่วงเวลา 6.00 น. จะวัดอุณหภูมิปกติของร่างกายได้ต่ำที่สุด และในช่วงเวลา 16.00 น. จะวัดอุณหภูมิปกติของร่างกายได้สูงสุด ดังนั้น อุณหภูมิของร่างกายที่ผิดปกติในเวลา 6.00 น. จะมากกว่า 37.2 องศาเซลเซียส และในช่วงเวลา 16.00 น. จะมากกว่า 37.7 องศาเซลเซียส (Fauci *et al.*, 2008)

2. ลมหายใจมีกลิ่นเหม็น หมายถึง กลิ่นเหม็นที่ออกมาจากลมหายใจซึ่งสามารถได้กลิ่นไม่รวมถึงก๊าซในความเข้มข้นมากกว่าปกติที่ไม่สามารถได้กลิ่น (Rosenberg, 2002)

3. อาการหายใจลำบาก หมายถึง การรับรู้ถึงความยากลำบากในการหายใจ และความรู้สึกที่ต้องออกแรงในการหายใจมากขึ้นกว่าปกติ ร่วมกับการใช้กล้ามเนื้ออื่นช่วยในการหายใจ เป็นอาการบ่งบอกจากการหายใจที่ไม่เพียงพอต่อความต้องการ (จันทร์จิรา, 2544)

4. โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute Upper Respiratory Infection) หมายถึง โรคติดต่อกจากการติดเชื้อไวรัสในระบบทางเดินหายใจส่วนบน ได้แก่ จมูก และคอ โดยทั่วไปเรียกว่า โรคหวัด หรือไข้หวัด (Common cold) (สุรเกียรติ, 2544)

5. โรคไซนัสอักเสบเฉียบพลัน (Acute Sinusitis) หมายถึง การอักเสบเฉียบพลันที่เกิดจากการติดเชื้อจากเยื่อบุโพรงอากาศไซนัสที่มีอาการน้อยกว่า 4 สัปดาห์ และอาการหายไปอย่างสมบูรณ์ (สุรเกียรติ, 2544)

6. โรคปอดอักเสบ (Pneumonia) หมายถึง การอักเสบของเนื้อปอด มีหนองขัง บวม จึงทำหน้าที่ไม่ได้เต็มที่ ทำให้การหายใจสะดุด เกิดอาการหายใจเหนื่อยหอบ โดยทั่วไปเรียกว่า โรคปอดบวม (สุรเกียรติ์, 2544)



การตรวจเอกสาร

การตรวจเอกสารจะแยกออกเป็น 3 ส่วน คือ ส่วนแรกกล่าวถึง เทคนิคเหมืองข้อมูล ส่วนที่สองอธิบาย แนวทางการชักประวัติผู้ป่วยโรคทางระบบหายใจ และความรู้เบื้องต้นเกี่ยวกับ โรคระบบทางเดินหายใจ ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ สำหรับส่วนสุดท้ายสรุป ผลงานวิจัยที่เกี่ยวข้อง

ส่วนที่ 1 เทคนิคเหมืองข้อมูล

เหมืองข้อมูล (Data mining)

เหมืองข้อมูล (Fayyad *et al.*, 1996) เป็นขั้นตอนที่สำคัญในกระบวนการค้นหาลักษณะแฝงของข้อมูลที่เป็นประโยชน์และน่าสนใจจากฐานข้อมูลขนาดใหญ่ (KDD) เช่น รูปแบบความสัมพันธ์ การเปลี่ยนแปลง โครงสร้างที่เด่นชัด หรือ ลักษณะที่ผิดปกติของข้อมูลจากข้อมูลขนาดใหญ่ที่เก็บอยู่ในฐานข้อมูลหรือแหล่งข้อมูลอื่นๆ ซึ่งวิธีการต่างๆ ที่นำมาใช้ในการทำเหมืองข้อมูล ก็มีวัตถุประสงค์ต่างกันไปขึ้นอยู่กับผลลัพธ์ของกระบวนการโดยรวมที่ต้องการ

โดยทั่วไปกระบวนการของ KDD นั้นประกอบด้วยขั้นตอนต่างๆ ดังนี้

1) ขั้นตอนการคัดเลือกข้อมูล (Data Selection) คือ การเลือกเฉพาะข้อมูลที่ต้องการและเป็นประโยชน์ไปใช้

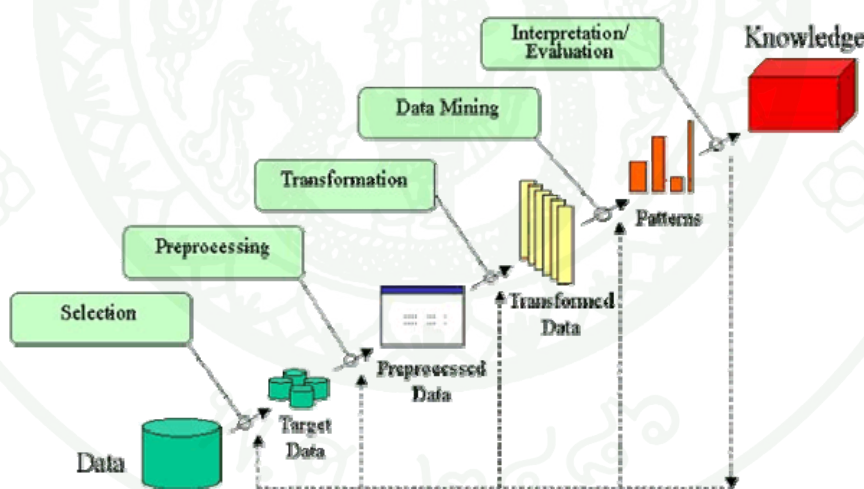
2) ขั้นตอนการเตรียมข้อมูล (Pre-processing) คือ การเตรียมข้อมูลให้เหมาะสม และให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้

3) ขั้นตอนการปรับแต่งข้อมูล (Data Transformation) คือ การเปลี่ยนแปลงรูปแบบหรือปรับค่าของข้อมูลเพื่อให้เหมาะสมสำหรับการนำไปใช้ เพื่อวิเคราะห์ตามอัลกอริทึมที่ใช้ในการทำเหมืองข้อมูลต่อไป

ขั้นตอนต่างๆ เหล่านี้จะเป็นตัวกำหนดความถูกต้อง และขอบเขตวัตถุประสงค์ที่ต้องการของฐานความรู้ที่ได้ เนื่องจากถ้าเลือกข้อมูลไม่ดี ข้อมูลมีขยะมาปะปนหรือมีการปรับแต่งข้อมูลไม่ถูกต้อง ก็จะทำให้ฐานความรู้ที่ได้ไม่ถูกต้องตามไปด้วย

4) ขั้นตอนเหมืองข้อมูล (Data Mining) คือ การเลือกเทคนิคที่เหมาะสมสำหรับงานที่ต้องการ สามารถใช้เทคนิคได้มากกว่าหนึ่งเทคนิคมาประมวลผลเพื่อดึงความรู้หรือสิ่งที่สนใจจากข้อมูลโดยผลลัพธ์ที่ได้จากขั้นตอนนี้ คือ ฐานความรู้

5) ขั้นตอนการแปลความ (Interpretation) คือ การนำฐานความรู้ที่ได้จากขั้นตอนเหมืองข้อมูลมาทดสอบ พิจารณาว่าถูกต้องตามความต้องการและมีประโยชน์หรือไม่ ซึ่งบางครั้งอาจต้องปรับแก้ค่าตัวแปรเพื่อกลับเข้าสู่ขั้นตอนเหมืองข้อมูลใหม่ ในบางครั้งก็ต้องเลือกข้อมูลชุดใหม่ หรือทำความสะอาดข้อมูลใหม่อีกครั้ง หลังจากการตรวจสอบความรู้จนถูกต้องแล้ว จึงนำความรู้นี้ไปใช้ซึ่งสามารถนำไปใช้ในการสืบค้นความรู้ การวิเคราะห์ข้อมูล หรือการทำนายปรากฏการณ์ที่จะเกิดขึ้นในอนาคตต่อไป



ภาพที่ 1 ขั้นตอนของกระบวนการ KDD

ที่มา: Fayyad *et al.* (1996)

การจำแนกข้อมูล (Data Classification)

การจำแนกข้อมูล (Chan and Stolfo, 1995; Gehrke *et al.*, 1998) เป็นเทคนิคหนึ่งใน KDD ที่หาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำมาใช้สร้างตัวจำแนกกลุ่มหรือคลาส (class) ของข้อมูล ซึ่งตัวจำแนกที่ได้สามารถนำไปทำนายกลุ่มของข้อมูลต่อไป โดยหลักการทำงานของเทคนิคนี้ จะนำข้อมูลส่วนหนึ่งมาสอนระบบให้สามารถเรียนรู้ (Training) เพื่อให้ระบบจำแนกข้อมูลนั้นออกมาเป็นกลุ่มต่างๆ ตามที่กำหนดไว้ ผลที่ได้จากการเรียนรู้ คือ ตัวจำแนกกลุ่มข้อมูล (Classifier) เมื่อมีข้อมูลใหม่เข้ามาสู่ระบบ ตัวจำแนกกลุ่มข้อมูลจะสามารถทำนายกลุ่มของข้อมูลที่ควรจะเป็นให้โดยอัตโนมัติ

สำหรับระบบเรียนรู้ (Learning) แบ่งออกเป็น 2 แบบ ได้แก่

1) Unsupervised Learning เป็นการเรียนรู้เพื่อค้นหาลักษณะบางอย่างที่เหมือนกันของข้อมูลแต่ละรายการ โดยที่ไม่ทราบชัดเจนว่าข้อมูลรายการนั้นถูกจัดอยู่ในกลุ่มใดหรือมีทั้งหมดกี่กลุ่ม โดยกระบวนการเรียนรู้จะพยายามที่จะหารูปแบบ (pattern) ความสัมพันธ์ภายในชุดข้อมูล ตัวอย่างแล้วทำการวิเคราะห์ว่าควรจัดข้อมูลออกเป็นกี่กลุ่ม และข้อมูลรายการใดควรอยู่ในกลุ่มใด แล้วทำการจัดกลุ่มข้อมูล (Clustering) (Witten and Frank, 2005)

2) Supervised Learning เป็นการเรียนรู้ที่มีเป้าหมายชัดเจนว่าต้องการศึกษาหรือจำแนกกลุ่มข้อมูลออกเป็นอย่างไร โดยนำเทคนิคและอัลกอริทึมต่างๆ มาใช้เพื่อจำแนกข้อมูล ประมาณค่าของข้อมูล หรือทำนายลักษณะข้อมูล จากความสัมพันธ์ของสิ่งต่างๆ ที่เกี่ยวข้อง ซึ่งกระบวนการเรียนรู้จะพยายามหาแบบจำลองของตัวจำแนก (Classifier) ในการแบ่งข้อมูลจากชุดของตัวอย่างออกเป็นกลุ่มๆ เพื่อให้เกิดความผิดพลาดน้อยที่สุดเมื่อนำไปใช้กับชุดข้อมูลจริงที่ไม่ได้อยู่ในชุดข้อมูลตัวอย่าง โดยชุดข้อมูลตัวอย่างจำเป็นที่จะต้องทราบและได้รับการกำหนดว่าอยู่กลุ่มใดก่อนเข้ากระบวนการเรียนรู้ สำหรับการจำแนกข้อมูล (Data Classification) (Witten and Frank, 2005)

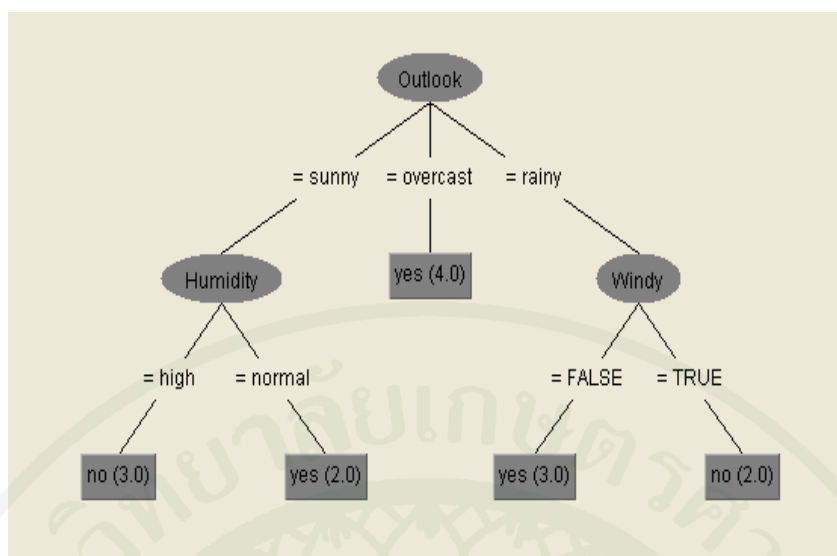
ตัวอย่างของการประยุกต์ใช้เทคนิคการจำแนกข้อมูล (Data Classification) คือ การวิเคราะห์สภาพอากาศเพื่อตัดสินใจว่าในขณะนั้น สมควรที่จะออกไปเล่นกีฬานอกบ้านหรือไม่ โดยเริ่มต้นจากการนำข้อมูลที่มีอยู่ (ตารางที่ 1) ซึ่งประกอบด้วยสภาพอากาศ (Outlook) อุณหภูมิ (Temperature) ความชื้น (Humidity) และแรงลม (Windy) พร้อมกับผลของการตัดสินใจว่าจะออกไปเล่นกีฬานอกบ้านหรือไม่ (Play) มาสอนให้ระบบเรียนรู้เพื่อการจำแนก ดังภาพที่ 2 คือ

ผลลัพธ์ที่ได้ของกระบวนการดังกล่าว แสดงเป็นแผนผังต้นไม้ โดยที่เส้นต่อระหว่าง โหนด
รากกับโหนดลูก คือ เงื่อนไขต่างๆ ของการตัดสินใจ เช่น no คือ ไม่ออกไปเล่นกีฬา หรือ
yes คือ ออกไปเล่นกีฬา จากแผนผังนี้จะช่วยให้ตัดสินใจได้ว่าเมื่อสภาพอากาศ อุณหภูมิ
ความชื้น และแรงลมในขณะนั้นมีลักษณะตามเงื่อนไขใดเงื่อนไขหนึ่ง จะตัดสินใจออกไปเล่นกีฬา
นอกบ้านหรือไม่ (Witten and Frank, 2005)

ตารางที่ 1 ตัวอย่างการประยุกต์ใช้ต้นไม้ตัดสินใจในการวิเคราะห์สภาพอากาศ สำหรับการออกไป
เล่นกีฬานอกบ้าน

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

ที่มา: Witten and Frank (2005)



ภาพที่ 2 ต้นไม้ตัดสินใจสำหรับจำแนกการไปออกกำลังกายตามเงื่อนไขสภาพอากาศต่างๆ

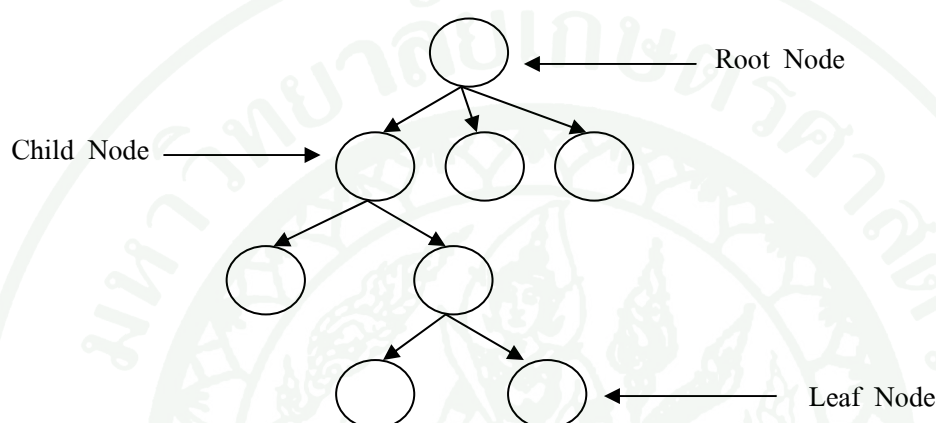
ที่มา: Witten and Frank (2005)

ที่ปลายกิ่งของต้นไม้ตัดสินใจ หรือโหนดใบ จะมีผลลัพธ์ของการตัดสินใจแสดงในรูปแบบ a/b โดย a หมายถึง จำนวนของรายการข้อมูลทั้งหมดที่มีคุณสมบัติตามเงื่อนไขในกิ่งก้าน และ b หมายถึง จำนวนของรายการข้อมูลที่แสดงผลลัพธ์การตัดสินใจที่โหนดใบตามเงื่อนไขของกิ่งก้าน สำหรับกรณีที่ a = b แผนภาพต้นไม้ตัดสินใจจะแสดงเฉพาะค่า a เพียงตัวเดียว เช่น จากภาพที่ 2 โหนดใบที่แสดงผลลัพธ์เป็น “no (3.0)” หมายถึง Outlook มีลักษณะ sunny แล้ว Humidity อยู่ในกลุ่ม high มีรายการข้อมูลทั้งหมด 3 รายการข้อมูล และทั้ง 3 รายการนี้มีผลลัพธ์การตัดสินใจเหมือนกัน คือ ไม่ออกนอกบ้าน (no) และในทำนองเดียวกันโหนดใบที่แสดงผลลัพธ์เป็น “yes (2.0)” หมายถึง ถ้า Outlook มีลักษณะ sunny แล้ว Humidity อยู่ในกลุ่ม normal จะมีรายการข้อมูลทั้งหมด 2 รายการ และทั้ง 2 รายการมีผลลัพธ์การตัดสินใจเหมือนกัน คือ ออกนอกบ้าน (yes) เป็นต้น

ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Cabema et al., 1998) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของต้นไม้ตัดสินใจ ซึ่งต้นไม้ตัดสินใจนั้นมีการทำงานแบบ Supervised learning คือ สามารถสร้างแบบจำลองการจัดกลุ่มข้อมูลจากชุดข้อมูลตัวอย่างที่ได้มีการกำหนดกลุ่มข้อมูล

ไว้แล้ว และเรียกชุดข้อมูลตัวอย่างนี้ว่า ชุดข้อมูลฝึกสอน และนำแบบจำลองที่ได้นี้ ไปใช้กับชุดข้อมูลอีกชุดหนึ่ง เรียกว่า ชุดข้อมูลทดสอบ เพื่อพยากรณ์กลุ่มของรายการข้อมูลที่ยังไม่เคยนำมาจัดกลุ่ม รูปแบบของต้นไม้ตัดสินใจจะประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (Root Node) จากโหนดรากก็จะแตกออกเป็นโหนดลูก (Child Node) และที่โหนดลูกก็จะมีลูกของตัวเองซึ่งโหนดที่ระดับสุดท้ายจะเรียกว่า โหนดใบ (Leaf Node) ดังภาพที่ 3



ภาพที่ 3 รูปแบบของต้นไม้ตัดสินใจ

จากภาพที่ 3 จะพบว่าจากโหนดราก จนถึงโหนดใบจะมีเพียงเส้นทางเดียวเท่านั้น ในเส้นทางนี้จะอธิบายถึงกฎที่ใช้สำหรับการจัดกลุ่มข้อมูล ซึ่งในแต่ละโหนดใบนั้นอาจจะให้ผลการตัดสินใจเหมือนกัน แต่จริงๆ แล้วอาจเกิดมาจากเงื่อนไขหรือกฎที่แตกต่างกันก็ได้

ต้นไม้ตัดสินใจมีหลายอัลกอริทึมที่ใช้ในการจำแนกข้อมูล เช่น ID3 (Iterative Dichotomizer 3), Chi-Squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 และ C5.0 algorithm เป็นต้น และแต่ละอัลกอริทึมก็จะให้โครงสร้างต้นไม้ตัดสินใจที่ต่างกัน (Larose, 2004) โดยอัลกอริทึมที่มีการนำมาใช้อย่างแพร่หลาย 3 อันดับแรก ได้แก่ ID3, C4.5 และ CART (Teknomo, 2009)

ID3 (Iterative Dichotomizer 3)

ID3 (Quinlan, 1986) เป็นอัลกอริทึมพื้นฐานแบบ Greedy Algorithm คือ มีขั้นตอนการแก้ปัญหาโดยพิจารณาว่าข้อมูลที่มีอยู่ในขณะนั้นมีทางเลือกใดที่ให้ผลตอบแทนดีที่สุด ซึ่งถ้าข้อมูลขณะนั้นมีพอเพียงก็จะทำให้ได้คำตอบที่ดีที่สุด โดยหลักการของอัลกอริทึม ID3 คือ จะคำนวณหาค่าเกณฑ์ (Gain) ซึ่งเป็นค่าที่บอกระดับความสามารถในการจำแนกกลุ่มของลักษณะ (Attribute) ที่ถูกเลือกเพื่อจัดกลุ่มของข้อมูลโดย Attribute ที่ให้ค่า Gain สูงสุดจะถูกนำมาเป็นโหนดราก การคำนวณค่า Gain อาศัยหลักการของทฤษฎีสารสนเทศ (Information Theory) ที่ว่าค่าสารสนเทศของข้อมูลขึ้นอยู่กับค่าความน่าจะเป็นของข้อมูล วัตถุประสงค์อยู่ในรูปของบิต (Bits) จากสูตร

$$\text{ค่าสารสนเทศของข้อมูล} = -\log_2(\text{ความน่าจะเป็นของข้อมูล})$$

ถ้าให้ชุดของข้อมูล S ประกอบด้วยค่าที่เป็นไปได้ คือ $\{s_1, s_2, \dots, s_n\}$ และให้ความน่าจะเป็นที่จะเกิดขึ้น s_i เป็น $P(s_i)$ สำหรับแต่ละค่า $s_i \in S$ ดังนั้นค่าสารสนเทศของ S เขียนแทนด้วย $I(S)$ จะคำนวณได้จากสูตร

$$I(S) = - \sum_{i=1}^n P(s_i) \log_2(P(s_i)) \quad (\text{หน่วยเป็น Bits})$$

ขั้นตอนการเลือก Attribute มาเป็นโหนดราก จะสมมุติให้ Attribute ใด Attribute หนึ่ง เป็นโหนดรากก่อน แล้วคำนวณค่า Gain จากผลต่างค่าสารสนเทศทั้งหมดของชุดข้อมูล $[I(S)]$ นั้น กับค่าสารสนเทศหลังจากเลือก Attribute นั้นเป็นราก $[I_x(t)]$ โดย $I_x(t)$ คำนวณจาก ค่าผลรวมของผลคูณระหว่างค่าสารสนเทศของแต่ละโหนดกับสัดส่วนของข้อมูลตัวอย่างในแต่ละกิ่งก้านต่อข้อมูลตัวอย่างทั้งหมดที่โหนดนั้นๆ

ถ้าให้ข้อมูลฝึกสอนมี T ค่าสังเกต และ Attribute ที่เลือกเป็นราก คือ X และมีค่าสังเกตหรือรายการข้อมูลทั้งหมด n ค่า รากหรือโหนดปัจจุบันจะแบ่งข้อมูลฝึกสอน T ออกเป็นเซตย่อยๆ $\{t_1, t_2, \dots, t_n\}$ แล้วจึงคำนวณค่าสารสนเทศหลังจากเลือก Attribute X ตามสูตร

$$I_x(T) = \sum_{i=1}^n \frac{|t_i|}{|T|} I(t_i)$$

ค่า Gain ของ Attribute X สามารถคำนวณได้จาก

$$\text{ค่า Gain (X)} = I(S) - I_x(T)$$

เมื่อ Attribute ตัวถัดไปถูกเลือกเป็นโหนดราก การคำนวณค่า Gain สำหรับ Attribute ก็จะคำนวณในลักษณะเดียวกับข้างต้น กระบวนการคำนวณค่า Gain ก็จะทำไปเรื่อยๆ จนกว่าจะใช้ Attribute ครบทุกตัวแล้วเลือก Attribute ที่ให้ค่า Gain สูงสุดมาเป็นโหนดราก

ลักษณะของข้อมูลสำหรับอัลกอริทึม ID3 ต้องเป็นข้อมูลเชิงกลุ่มเท่านั้น และไม่สามารถใช้กับข้อมูลที่มีการสูญหายได้ หากจำเป็นต้องใช้ข้อมูลเชิงปริมาณ ต้องทำการจัดกลุ่มข้อมูลทั้งหมดก่อน ป้องกันปัญหาความอคติของข้อมูล (Lior and Oded, 2008)

เพื่ออธิบายหลักการทำงานของอัลกอริทึม ID3 จะใช้ตัวอย่างปัจจัยที่ทำให้ผิวหนังไหม้ ซึ่งประกอบด้วย 4 Attribute คือ สีผม, สภาพอากาศ, สีผิว และการทาโลชั่น โดยมี Attribute ผิวหนัง เป็นเป้าหมายของการทำ Classification ดังตารางที่ 2

ตารางที่ 2 ตัวอย่างข้อมูลของปัจจัยที่ทำให้ผิวหนังไหม้

คนที่	สีผม	สภาพอากาศ	สีผิว	การทาโลชั่น	ผิวหนัง
1	บรอน	แดดจ้า	ขาว	ไม่ใช้	ไหม้
2	บรอน	แดดจ้า	ขาว	ใช้	ไหม้
3	ดำ	แดดจ้า	ขาว	ไม่ใช้	ไม่ไหม้
4	น้ำตาล	มีดครีမ်	ขาว	ไม่ใช้	ไม่ไหม้
5	น้ำตาล	ปกติ	ปกติ	ไม่ใช้	ไม่ไหม้
6	น้ำตาล	ปกติ	ปกติ	ใช้	ไหม้
7	ดำ	ปกติ	ปกติ	ใช้	ไม่ไหม้
8	บรอน	มีดครีမ်	ขาว	ไม่ใช้	ไหม้
9	บรอน	ปกติ	ปกติ	ไม่ใช้	ไม่ไหม้
10	น้ำตาล	มีดครีမ်	ปกติ	ไม่ใช้	ไม่ไหม้
11	บรอน	มีดครีမ်	ปกติ	ใช้	ไม่ไหม้
12	ดำ	มีดครีမ်	ขาว	ใช้	ไม่ไหม้
13	ดำ	แดดจ้า	ปกติ	ไม่ใช้	ไม่ไหม้
14	น้ำตาล	มีดครีမ်	ขาว	ใช้	ไหม้

ที่มา: ข้อมูลคนที่ 1-7 (ศิริพร, 2542) ข้อมูลคนที่ 8-14 ข้อมูลสมมุติ

จากตัวอย่างปัจจัยที่ทำให้ผิวหนังไหม้ในตารางที่ 2 เซตของข้อมูลฝึก T ประกอบด้วยข้อมูล 2 กลุ่ม คือ ตัวอย่างที่ผิวหนังไหม้ 9 คน และผิวหนังไหม้ 5 คน ดังนั้นค่าสารสนเทศทั้งหมดของข้อมูลชุดนี้จะคำนวณได้ดังนี้

$$\begin{aligned}
 I(S) &= [-(\text{ความถี่ของการปรากฏข้อมูลเป็นกลุ่มไม่ไหม้} / \text{จำนวนข้อมูลทั้งหมด}) \\
 &\quad \times \log_2 (\text{ความถี่ของการปรากฏข้อมูลเป็นกลุ่มไม่ไหม้} / \text{จำนวนข้อมูลทั้งหมด}) + \\
 &\quad [-(\text{ความถี่ของการปรากฏข้อมูลเป็นกลุ่มไหม้} / \text{จำนวนข้อมูลทั้งหมด}) \\
 &\quad \times \log_2 (\text{ความถี่ของการปรากฏข้อมูลเป็นกลุ่มไหม้} / \text{จำนวนข้อมูลทั้งหมด})]
 \end{aligned}$$

ดังนั้น

$$\begin{aligned} I(S) &= (-9/14)\log_2(9/14) + (-5/14)\log_2(5/14) \\ &= 0.940 \text{ bits} \end{aligned}$$

การจำแนกกลุ่มของข้อมูลออกเป็นผิวหนัง = ไม่ไหม้ หรือ
ผิวหนัง = ไหม้ จะต้องใช้ข้อมูลจาก Attribute อื่นประกอบการตัดสินใจ โดยคำนวณหาค่า Gain
ประกอบการเลือก Attribute ที่เหมาะสมสำหรับการจัดกลุ่ม

Attribute สีผม

$$\begin{aligned} I_{\text{สีผม}}(T) &= (5/14) \times [-(2/5) \times \log_2(2/5) - (3/5) \times \log_2(3/5)] \\ &\quad + (4/14) \times [-(4/4) \times \log_2(4/4) - (0/4) \times \log_2(0/4)] \\ &\quad + (5/14) \times [-(3/5) \times \log_2(3/5) - (2/5) \times \log_2(2/5)] \\ &= 0.693 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{Gain (สีผม)} &= 0.940 - 0.693 \\ &= 0.247 \text{ bits} \end{aligned}$$

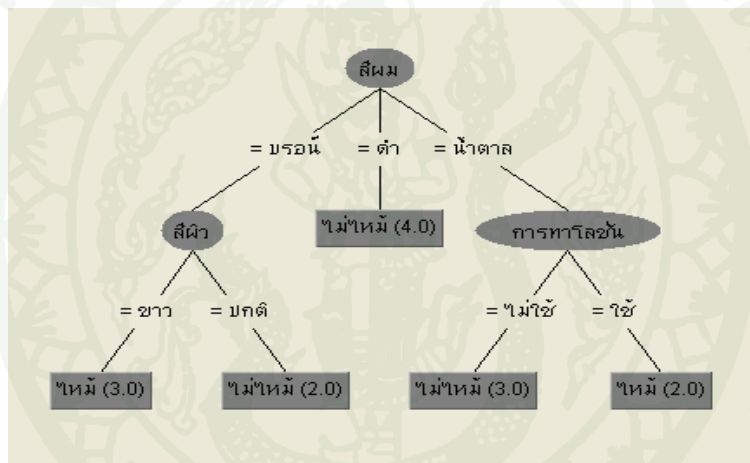
ในทำนองเดียวกันสามารถคำนวณหาค่าสารสนเทศหลังจากแบ่งตาม
Attribute สภาพอากาศ, สีผิว และการทาโลชั่น แล้วนำค่าสารสนเทศที่ได้มาคำนวณหาค่า Gain
ของ Attribute ต่างๆ ดังนี้

$$\begin{aligned} \text{Gain (สภาพอากาศ)} &= I(S) - I_{\text{สภาพอากาศ}}(T) \\ &= 0.940 - 0.911 \\ &= 0.029 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{Gain (สีผิว)} &= I(S) - I_{\text{สีผิว}}(T) \\ &= 0.940 - 0.788 \\ &= 0.152 \text{ bits} \end{aligned}$$

$$\begin{aligned}
 \text{Gain (การหาโลชั่น)} &= I(S) - I_{\text{การหาโลชั่น}}(T) \\
 &= 0.940 - 0.892 \\
 &= 0.048 \text{ bits}
 \end{aligned}$$

ซึ่ง Attribute ที่ให้ค่า Gain สูงที่สุด ได้แก่ สีผม ดังนั้น Attribute สีผม จึงถูกเลือกเป็นโหนดรากของต้นไม้ตัดสินใจ แต่เนื่องจาก Attribute สีผม ยังไม่สามารถจัดกลุ่มข้อมูลให้เป็นกลุ่มเดียวกันทั้งหมด จึงต้องสร้างต้นไม้ตัดสินใจต่อไป โดยพิจารณาเลือก Attribute ที่จะมาเป็นโหนดต่อจากโหนดราก ทำเช่นนี้ต่อไปโดยกระบวนการสร้างต้นไม้ตัดสินใจจะสิ้นสุดเมื่อโหนดใดเป็นกลุ่มของข้อมูลกลุ่มเดียวกันทั้งหมด โดยภาพต้นไม้ตัดสินใจของข้อมูลปัจจัยที่ทำให้ผิวหนังไหม้จะได้ดังภาพที่ 4



ภาพที่ 4 ต้นไม้ตัดสินใจที่ได้จากการใช้โปรแกรม Weka กับชุดข้อมูลปัจจัยที่ทำให้ผิวหนังไหม้

C4.5

C4.5 (Quinlan, 1992) เป็นอัลกอริทึมของต้นไม้ตัดสินใจที่พัฒนาเพิ่มเติม โดยใช้ ค่าอัตราส่วนเกน (Gain Ratio) เป็นตัวจำแนกในการตัดสินใจเลือก Attribute ที่จะใช้เป็นรากหรือโหนด ซึ่งการทำงานในขั้นตอนแรกคล้ายกับการทำงานของอัลกอริทึม ID3 โดยต้องหาค่าสารสนเทศของข้อมูล คือ ค่า Gain แล้วจึงหาค่าสารสนเทศของการแบ่งหรือจำแนกข้อมูล (Split Information) ตามลักษณะของ Attribute แต่ละตัวเพื่อมาคำนวณหาค่า Gain Ratio ซึ่งเป็นหลักเกณฑ์ในการสร้างต้นไม้ตัดสินใจสำหรับอัลกอริทึม C4.5

โดยค่าสารสนเทศของการแบ่งหรือจำแนกข้อมูลนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อแบ่งข้อมูลตัวอย่าง T เป็น n ชุดย่อยตาม Attribute x โดยค่านี้จะสูงสุดเมื่อ $|t_i|$ เป็น 1 เท่ากันในทุกกิ่ง และลดลงเมื่อค่า $|t_i|$ เพิ่มขึ้นเมื่อนำค่านี้ไปหารค่า Gain จะได้ค่า Gain Ratio ดังนี้

ค่าอัตราส่วนเกน (Gain Ratio)

$$\text{Gain Ratio} = \frac{\text{Gain}}{\text{Split Info}}$$

$$\text{เมื่อ Split Info} = - \sum_{i=1}^n \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|}$$

และ T = ชุดของข้อมูลตัวอย่าง

t_i = ชุดของข้อมูลตัวอย่างย่อยที่ i

n = จำนวนชุดตัวอย่างย่อย

ลักษณะของข้อมูลสำหรับอัลกอริทึม C4.5 สามารถใช้ได้ทั้งข้อมูลเชิงกลุ่มและข้อมูลเชิงปริมาณ และสามารถจัดการกับกรณีที่มีค่าสูญหายได้ เป็นส่วนที่พัฒนาเพิ่มเติมจาก ID3 (Lior and Oded, 2008)

เพื่ออธิบายหลักการทำงานของอัลกอริทึม C4.5 จะใช้ตัวอย่างปัจจัยที่ทำให้ผิวหนังไหม้จากตารางที่ 2 แต่เนื่องด้วยหลักการทำงานของอัลกอริทึม C4.5 จะมีบางส่วนเหมือนกับอัลกอริทึม ID3 นั่นคือ ค่า Gain สำหรับส่วนที่แตกต่างกับการทำงานอัลกอริทึม ID3 คือ ค่า Split Info จึงคำนวณหาค่า Split Info ดังนี้

Attribute สีผม

$$\text{Gain (สีผม)} = 0.247$$

$$\begin{aligned}\text{split info}_{\text{สีผม}} &= [-(5/14) \times \log_2(5/14)] + [-(4/14) \times \log_2(4/14)] \\ &\quad + [-(5/14) \times \log_2(5/14)] \\ &= 1.577 \text{ bits}\end{aligned}$$

เมื่อได้ค่า split info ของ Attribute สีผม แล้วนำค่าที่ได้มาหาค่า Gain Ratio ของ Attribute สีผม ดังนี้

$$\begin{aligned}\text{Gain Ratio} &= \frac{0.247}{1.577} \\ &= 0.156\end{aligned}$$

สำหรับค่า 0.156 จะแสดงถึงค่าความน่าจะเป็นในการแบ่งหรือจำแนกข้อมูลด้วย Attribute สีผม ที่มีการปรับแก้ความอคติจากค่า Gain

ในทำนองเดียวกันสามารถคำนวณหาค่า split info ของ Attribute ต่างๆ แล้วจึงนำมาหาค่า Gain Ratio ของแต่ละ Attribute โดยเลือก Attribute ที่ให้ค่า Gain Ratio สูงสุด เป็นโหนดรากของต้นไม้ตัดสินใจ ถ้ายังไม่สามารถจัดกลุ่มข้อมูลให้เป็นกลุ่มเดียวกันทั้งหมด จะต้องสร้างต้นไม้ตัดสินใจต่อไป โดยพิจารณาเลือก Attribute ที่จะมาเป็นโหนดต่อจากโหนดราก ทำเช่นนี้ต่อไปโดยกระบวนการสร้างต้นไม้ตัดสินใจ ซึ่งจะสิ้นสุดเมื่อโหนดใดเป็นกลุ่มของข้อมูลกลุ่มเดียวกันทั้งหมด

CART (Classification and Regression Trees)

CART (Breiman, 1984) เป็นสถิติอนพารามตริก ที่ใช้จำแนกกลุ่มออกเป็น 2 กลุ่ม โดยตัวแปรอิสระจะมีลักษณะเชิงกลุ่มหรือเชิงปริมาณก็ได้ ซึ่งหากตัวแปรตามเป็นตัวแปรเชิงกลุ่ม จะใช้เทคนิคการแบ่งกลุ่ม Classification Tree แต่ถ้าตัวแปรตามเป็นตัวแปรเชิงปริมาณ จะใช้เทคนิคการแบ่งกลุ่ม Regression Tree ซึ่งหลักการของวิธี CART จะเริ่มต้นด้วยการถือว่าสิ่งของทุกชิ้นอยู่กลุ่มเดียวกัน แล้วแบ่งกลุ่มออกเป็น 2 กลุ่ม โดยใช้ตัวแปรที่มีค่าสูงในกลุ่มหนึ่งและต่ำในกลุ่มหนึ่งเป็นตัวแบ่ง หากตัวแปรมีความเป็นไปได้มากกว่า 2 กลุ่ม จะพยายามจัดกลุ่มให้เหลือเพียง 2 กลุ่มเท่านั้น โดยจำนวนในแต่ละกลุ่มไม่จำเป็นต้องเท่ากัน จากนั้นก็แบ่งแต่ละกลุ่มย่อยออกเป็น 2 กลุ่ม แบบทวิภาค ทำการแบ่งย่อยอย่างนี้ไปเรื่อยๆ จนกระทั่งถึงจุดที่

เหมาะสมจึงหาค่าของตัวแปรที่ใช้แบ่งกลุ่ม อาจเป็นตัวแปรเชิงกลุ่มที่เรียงลำดับหรือไม่ได้เรียงลำดับก็ได้ ซึ่งเรียกแผนภาพที่ได้จากวิธีการนี้ว่า “ต้นไม้การจำแนก” โดยหลักเกณฑ์ที่ใช้ในการจำแนกกลุ่มของอัลกอริทึมนี้ คือ ค่า Goodness of Split ซึ่งคำนวณได้จากสูตร

ค่า Goodness of Split

$$\begin{aligned}\Delta i(s,t) &= i(t) - p_L[i(t_L)] - p_R[i(t_R)] \\ &= 1 - \sum_{j=1}^k p^2(j|t) - p_L[i(t_L)] - p_R[i(t_R)]\end{aligned}$$

$$\text{เมื่อ } i(t) = 1 - \sum_{j=1}^k p^2(j|t)$$

และ $\Delta i(s,t)$ = ค่า Goodness of Split เมื่อเลือก Attribute s เป็นโหนดราก
และ Attribute t เป็นโหนดลูก

j = จำนวนเหตุการณ์ที่สนใจ

t = จำนวนของชุดข้อมูล

k = จำนวนของ class

p_L = สัดส่วนของ Case ที่ node t ใน Child Node ด้านชุดข้อมูล t_L

p_R = สัดส่วนของ Case ที่ node t ใน Child Node ด้านชุดข้อมูล t_R

$$i(t_L) = 1 - \sum_{j=1}^k p^2(j|t_L)$$

$$i(t_R) = 1 - \sum_{j=1}^k p^2(j|t_R)$$

$$t_L = \text{ชุดของข้อมูลทางซ้ายของโหนด } t$$

$$t_R = \text{ชุดของข้อมูลทางขวาของโหนด } t$$

เพื่ออธิบายหลักการทำงานของอัลกอริทึม CART จะใช้ตัวอย่างปัจจัยที่ทำให้ผิวหนังไหม้จากตารางที่ 2 โดยหลักการทำงานของอัลกอริทึม CART เริ่มต้นด้วยการถือว่าทุกตัวแปรเป็นกลุ่มเดียวกัน โดยเลือก Attribute ผิวหนัง ทั้งนี้เพราะว่า Attribute ผิวหนังสามารถแบ่งข้อมูลเป็น 2 กลุ่มได้ โดยแบ่งเป็นกลุ่มที่ผิวหนังไหม้กับกลุ่มผิวหนังไม่ไหม้ จากนั้นจึงทำการคำนวณหาค่า Goodness of Split ระหว่าง Attribute ผิวหนังกับ Attribute อื่นๆ เพื่อหาค่า Goodness of Split ที่มากที่สุดสำหรับเลือก Attribute ในการสร้างต้นไม้ตัดสินใจดังนี้

Goodness of Split (ผิวหนัง,สีผม)

$$\begin{aligned} \Delta i \text{ (ผิวหนัง,สีผม)} &= 1 - \left[\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right] - \left[\left(\frac{5}{14} \right) \left(1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right] \right) \right] \\ &\quad - \left[\left(\frac{4}{14} \right) \left(1 - \left[\left(\frac{4}{4} \right)^2 + \left(\frac{0}{4} \right)^2 \right] \right) \right] - \left[\left(\frac{5}{14} \right) \left(1 - \left[\left(\frac{3}{5} \right)^2 + \left(\frac{2}{5} \right)^2 \right] \right) \right] \\ &= 0.1164 \end{aligned}$$

Goodness of Split (ผิวหนัง,อากาศ)

$$\begin{aligned} \Delta i \text{ (ผิวหนัง,อากาศ)} &= 1 - \left[\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right] - \left[\left(\frac{4}{14} \right) \left(1 - \left[\left(\frac{2}{4} \right)^2 + \left(\frac{2}{4} \right)^2 \right] \right) \right] \\ &\quad - \left[\left(\frac{6}{14} \right) \left(1 - \left[\left(\frac{4}{6} \right)^2 + \left(\frac{2}{6} \right)^2 \right] \right) \right] - \left[\left(\frac{6}{14} \right) \left(1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] \right) \right] \\ &= -0.0528 \end{aligned}$$

Goodness of Split (ผิวหนัง,สีผิว)

$$\begin{aligned}\Delta_i \text{ (ผิวหนัง,สีผิว)} &= 1 - \left[\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right] - \left[\left(\frac{7}{14} \right) \left(1 - \left[\left(\frac{3}{7} \right)^2 + \left(\frac{4}{7} \right)^2 \right] \right) \right] \\ &\quad - \left[\left(\frac{7}{14} \right) \left(1 - \left[\left(\frac{6}{7} \right)^2 + \left(\frac{1}{7} \right)^2 \right] \right) \right] \\ &= 0.0918\end{aligned}$$

Goodness of Split (ผิวหนัง,การทาโลชั่น)

$$\begin{aligned}\Delta_i \text{ (ผิวหนัง,การทาโลชั่น)} &= 1 - \left[\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right] - \left[\left(\frac{9}{14} \right) \left(1 - \left[\left(\frac{6}{9} \right)^2 + \left(\frac{3}{9} \right)^2 \right] \right) \right] \\ &\quad - \left[\left(\frac{5}{14} \right) \left(1 - \left[\left(\frac{3}{5} \right)^2 + \left(\frac{2}{5} \right)^2 \right] \right) \right] \\ &= 0.0020\end{aligned}$$

จะได้ว่า ค่า Goodness of Split (ผิวหนัง,สีผม) = 0.1164

ค่า Goodness of Split (ผิวหนัง,อากาศ) = - 0.0528

ค่า Goodness of Split (ผิวหนัง,สีผิว) = 0.0918

ค่า Goodness of Split (ผิวหนัง,การทาโลชั่น) = 0.0020

Attribute ที่เหมาะสมใช้แบ่งกลุ่มข้อมูลเป็นโหนดลูก คือ Attribute สีผม เนื่องจากให้ค่า Goodness of Split มากที่สุด จากนั้นทำการหา Attribute ตัวถัดไปเพื่อใช้ในการแบ่งกลุ่มข้อมูลสำหรับโหนดลูกถัดไป โดยการคำนวณค่า Goodness of Split ระหว่าง Attribute

สัมพันธ์กับ Attribute ที่เหลือทุกตัว แล้ว Attribute ที่ให้ค่า Goodness of Split สูงสุดจะถูกเลือก สำหรับใช้แบ่งกลุ่มข้อมูลต่อไป ดังนั้นกระบวนการสร้างต้นไม้ตัดสินใจจะสิ้นสุดเมื่อโหนดใดเป็นข้อมูลกลุ่มเดียวกันทั้งหมด

สรุป อัลกอริทึม ID3, C4.5 และ CART มีหลักการสร้างตัวจำแนกที่แตกต่างกัน คือ อัลกอริทึม ID3 จะใช้หลักเกณฑ์ค่า Gain (Quinlan, 1986) ส่วนอัลกอริทึม C4.5 จะมีวิธีการทำงานเหมือนกับอัลกอริทึม ID3 แต่จะมีการคำนวณค่าที่ใช้สำหรับการเลือก Attribute แตกต่างกันไป คือ อัลกอริทึม C4.5 จะใช้ค่า Gain Ratio แทนค่า Gain ที่ใช้ใน ID3 (Quinlan, 1992) สำหรับอัลกอริทึม CART จะใช้ Goodness of Split เป็นเกณฑ์ในการสร้างตัวจำแนก จากหลักการที่แตกต่างนี้จะส่งผลให้ต้นไม้ตัดสินใจมีลักษณะที่แตกต่างกัน (Breiman *et al.*, 1984)

วิธีการแบ่งชุดข้อมูลเพื่อการฝึกสอนและทดสอบ

1. วิธี Cross-validation

เป็นวิธีการแบ่งข้อมูลออกเป็นส่วนๆ เรียกว่า “Fold” เช่น วิธี 3 Fold Cross-validation จะทำการแบ่งข้อมูลออกเป็น 3 ส่วน ในครั้งแรกจะเก็บส่วนที่ 3 ไว้เพื่อเป็นชุดข้อมูลทดสอบ และส่วนที่ 1 และ 2 ใช้สำหรับเป็นชุดข้อมูลฝึกสอน ครั้งที่สองจะเก็บข้อมูลส่วนที่ 2 ไว้เป็นชุดข้อมูลทดสอบ สำหรับส่วนที่ 1 และ 3 ใช้เป็นชุดข้อมูลฝึกสอน และครั้งที่สุดท้ายจะใช้ข้อมูลส่วนที่ 1 ไว้เป็นชุดข้อมูลทดสอบ สำหรับข้อมูลส่วนที่ 2 และ 3 เป็นชุดข้อมูลฝึกสอน ซึ่งวิธี Cross-validation ข้อมูลทุกส่วนจะถูกใช้เป็นที่ชุดข้อมูลฝึกสอนและทดสอบสลับกันไป (Witten and Frank, 2005)

2. วิธี Percentage split

เป็นวิธีการแบ่งข้อมูลออกเป็น 2 ส่วนตามร้อยละที่กำหนด โดยแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ เช่น วิธี 60% split จะแบ่งข้อมูลชุดฝึกสอนเป็น 60% และที่เหลืออีก 40% จะใช้เป็นข้อมูลชุดทดสอบ (Witten and Frank, 2005)

การคัดเลือกตัวแปรอิสระ

การคัดเลือกตัวแปรอิสระในการวิเคราะห์อัลกอริทึม ID3, C4.5 และ CART เป็นวิธีการคัดเลือกตัวแปรอิสระที่มีความสำคัญต่อการจำแนกกลุ่มหรือประเภทข้อมูล การคัดเลือกแบ่งเป็น 2 แบบ คือ 1) ใช้ชุดข้อมูลฝึกสอนทั้งหมด (Use full training set) และ 2) ใช้ชุดข้อมูลฝึกสอนที่ได้จากการแบ่งแบบ 10 Folds Cross-validation โดยในแต่ละแบบใช้อัลกอริทึมการคัดเลือกตัวแปร 3 วิธี ได้แก่

1. วิธี Best First

เป็นวิธีการคัดเลือกตัวแปรอิสระ หรือเลือก attribute ของข้อมูลซึ่งมีหลายวิธีดังนี้ 1) การคัดเลือกเพิ่ม (Forward) คือจะเริ่มคัดเลือกตัวแปรอิสระเข้าทีละตัว 2) การคัดเลือกลด (Backward) วิธีนี้จะนำตัวแปรอิสระทุกตัวเข้าสมการก่อน แล้วจึงคัดเลือกรอกทีละตัว หรือ 3) ทำการเรียงตัวแปรตามรายการแล้วเลือกตัวแปร ณ ตำแหน่งตรงกลาง จากนั้นทำการค้นหาตัวแปรถัดไป ทั้งทางด้านซ้ายและด้านขวาของตัวแปรที่ถูกคัดเลือกก่อนหน้านี้ เพื่อคัดเข้าเป็นตัวแปรถัดไป หรืออาจจะถูกพิจารณาคัดออกในขั้นตอนของการคัดเลือกถัดไปก็ได้ (Witten and Frank, 2005)

2. วิธี Greedy Stepwise

เป็นวิธีการคัดเลือกตัวแปรอิสระ ที่ใช้หลักการคล้ายกับวิธี Stepwise ซึ่งเป็นการคัดเลือกตัวแปรทีละขั้นตอน โดยขั้นตอนแรกจะคัดเลือก attribute เข้าสมการ โดยวิธี Forward จากนั้นจึงทำการเลือกตัวแปรอิสระที่ผ่านเกณฑ์ที่กำหนดไว้เข้าสมการพร้อมทั้งพิจารณาว่ามีตัวแปรอิสระใดที่อยู่ในตัวแบบก่อนหน้านี้ควรจะถูกรีดออกหรือไม่ ทำเช่นนั้นจนกระทั่งไม่สามารถเลือกตัวแปรอิสระเข้าตัวแบบและไม่สามารถคัดตัวแปรอิสระออกจากตัวแบบได้ (Witten and Frank, 2005)

3. วิธี Genetic Search

เป็นวิธีการคัดเลือกตัวแปรอิสระ โดยใช้วิธีแบบ Genetic Algorithm อย่างง่าย ถือเป็นกระบวนการค้นหาคำตอบที่มีพื้นฐานอยู่บนทฤษฎีทางพันธุศาสตร์ และกลไกการคัดเลือกตามธรรมชาติ ระเบียบวิธีเลียนแบบพันธุศาสตร์ มีกระบวนการหลักสามส่วน คือ การคัดเลือก (Selection) การไขว้สายพันธุ์ (Crossover) และ การผ่าเหล่า (Mutation) ทำงานร่วมกันทำให้

ระเบียบวิธีนี้เป็นการหาคำตอบที่มีประสิทธิภาพสูงสุด โดยองค์ประกอบที่สำคัญ Genetic Algorithm มี 5 ส่วน คือ

3.1 Chromosome Encoding คือ ขั้นตอนสำหรับกำหนดตัวแปรในรูปแบบ Chromosome ซึ่งสามารถทำได้ในหลายรูปแบบ แล้วแต่ความเหมาะสมของลักษณะตัวแปร

3.2 Initial Population คือ การสุ่มเลือกประชากรต้นแบบเพื่อใช้เป็นจุดเริ่มต้นของขั้นตอนการวิวัฒนาการ เป็นขั้นตอนแรกก่อนเข้ากระบวนการ Genetic Algorithm โดยประชากรกลุ่มแรก หรือประชากรต้นกำเนิด ถูกสุ่มเลือกจากประชากรทั้งหมดที่มีอยู่ และในการสุ่มเลือกจะสุ่มตามจำนวนของประชากรที่ถูกกำหนดเป็น Parameter ของ Algorithm

3.3 Fitness Function เป็นวิธีสำหรับประเมินค่าความเหมาะสม เพื่อให้คะแนนแต่ละทางเลือกของคำตอบต่างๆ โครโมโซมทุกตัวจะมีค่าความเหมาะสมของตัวเองเพื่อใช้สำหรับพิจารณาว่า โครโมโซมตัวนั้น เหมาะหรือไม่ที่จะนำมาใช้สืบทอดพันธุกรรมสำหรับสร้างโครโมโซมรุ่นใหม่ โดยวิธีการสำหรับคิดค่าความเหมาะสมนั้นจะใช้สมการที่สอดคล้องกับแต่ละปัญหา เช่น กำหนดความเหมาะสมเท่ากับจำนวนของบิต 1 ทั้งหมดในโครโมโซม โครโมโซม A:100011100 ดังนั้น โครโมโซม A มีค่าความเหมาะสมเท่ากับ 4 เป็นต้น

3.4 Genetic Operator การดำเนินการต่างๆ ตามขั้นตอนของ Genetic Algorithm เพื่อให้การเกิดวิวัฒนาการไปสู่คำตอบที่ดีขึ้น ซึ่งได้แก่ Selection, Crossover และ Mutation

3.5 Parameter คือ ปัจจัยที่ส่งผลต่อการทำงานของ Genetic Algorithm เช่น ขนาดของประชากร, ความน่าจะเป็นของการ Crossover หรือ ความน่าจะเป็นของการ Mutation (Mitchell, 1998)

เกณฑ์ในการเปรียบเทียบประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ

Greenberg *et al* (2004) และ Witten and Frank (2005) แนะนำให้ใช้ค่าวัดความถูกต้อง (Accuracy), Kappa Statistic, อัตราผลบวกจริง (TP Rate), อัตราผลบวกเท็จ (FP Rate), ค่าพยากรณ์บวก (PPV), ค่าพยากรณ์ลบ (NPV) และค่าพื้นที่ใต้ ROC curve (AUC) เป็นเกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ ซึ่งคำนวณได้ดังนี้

ตารางที่ 3 การตรวจสอบความสามารถของตัวจำแนกที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจ
ในการวินิจฉัยการเป็นโรคของผู้ป่วย

Test result (Predicted class)	Truth (Actual class)		Total
	Disease(Yes)	No disease(No)	
Positive (Yes)	a True-Positive (TP)	b False-Positive (FP)	a+b All Positive
Negative (No)	c False- Negative (FN)	d True- Negative (TN)	c+d All Negative
Total	a+c All Disease	b+d All No disease	a+b+c+d All tested

ที่มา : Greenberg *et al.* (2004) ; Witten and Frank (2005)

1. ค่าวัดความถูกต้อง (Accuracy) หมายถึง ผลของการใช้อัลกอริทึมต้นไม้ตัดสินใจให้ผลการจำแนกที่ถูกต้องของการจำแนกกลุ่มที่เป็นโรค และกลุ่มที่ไม่เป็นโรค โดยวัดค่าความถูกต้องเป็นค่าร้อยละ ซึ่งอัลกอริทึมที่ให้ค่าร้อยละของค่าความถูกต้องสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Witten and Frank, 2005)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \times 100$$

2. Kappa statistics เป็นค่าสถิติที่ใช้สำหรับวัดความสอดคล้องระหว่างค่าการพยากรณ์และค่าที่สังเกตได้จริง มีค่าอยู่ระหว่าง -1 ถึง 1 (Cunningham, 2009) สามารถคำนวณได้จาก

$$\text{Kappa statistics} = \frac{(p_o - p_c)}{(1 - p_c)}$$

โดยที่

p_o = Observed agreement

$$p_o = \frac{(a+d)}{a+b+c+d}$$

p_c = Chance agreement

$$p_c = \frac{((a+c)(a+b) + (b+d)(c+d))}{(a+b+c+d)^2}$$

การแปลความหมายของ Kappa statistics (Landis and Koch, 1977) แสดงในตารางที่ 4

ตารางที่ 4 การแปลความหมายค่า Kappa statistics

ค่า Kappa statistics	หมายถึง
< 0.00	ความสอดคล้องไม่ดี
0.01 – 0.20	ความสอดคล้องค่อนข้างพอใช้
0.21 – 0.40	ความสอดคล้องพอใช้
0.40 – 0.60	ความสอดคล้องดีพอใช้
0.61 – 0.80	ความสอดคล้องดี
0.81 – 1	ความสอดคล้องดีเยี่ยม

ที่มา : Landis and Koch (1977)

3. อัตราผลบวกจริง (True Positive Rate) หรือ ค่าความไว (Sensitivity) หมายถึง ผลการทดสอบที่ให้ผลบวก แสดงว่า เป็นโรคตรงตามความเป็นจริง (a) เทียบกับจำนวนผู้ที่ เป็นโรคทั้งหมด (a+c) แสดงเป็นค่าร้อยละ และอัลกอริทึมที่ให้ค่าร้อยละของอัตราผลบวกจริงสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Greenberg *et al.*, 2004; Witten and Frank, 2005)

$$TP\ Rate = \frac{a}{a+c} \times 100$$

4. อัตราผลลบจริง (True Negative Rate) หรือ ค่าความจำเพาะ (Specificity) หมายถึง ผลการทดสอบที่ให้ผลลบ แสดงว่า ไม่เป็นโรคตรงตามความเป็นจริง (d) เทียบกับจำนวนผู้ที่ไม่

เป็นโรคทั้งหมด (b+d) แสดงเป็นค่าร้อยละ และอัลกอริทึมที่ให้ค่าร้อยละของอัตราผลบวจริงสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Greenberg *et al.*, 2004; Witten and Frank, 2005)

$$\text{TN Rate} = \frac{d}{b+d} \times 100$$

5. อัตราผลบวกเท็จ (False Positive Rate) หรือ 1 – ค่าความจำเพาะ หมายถึง ผลการทดสอบที่ให้ผลบวก คือ เป็นโรค แต่ผู้ที่ทดสอบไม่ได้เป็นโรคจริง (b) เทียบกับจำนวนผู้ที่ไม่เป็นโรคทั้งหมด (b+d) แสดงเป็นค่าร้อยละ และอัลกอริทึมที่ให้ค่าร้อยละของอัตราผลบวกเท็จต่ำกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Greenberg *et al.*, 2004; Witten and Frank, 2005)

$$\text{FP Rate} = \frac{b}{b+d} \times 100$$

6. ค่าพยากรณ์บวก (Positive Predictive Value) หมายถึง ผลการทดสอบที่ให้ผลบวก คือ เป็นโรคตรงตามความเป็นจริง (a) เทียบกับผลการทดสอบที่ให้ผลลัพธ์เป็นโรคทั้งหมด (a+b) แสดงเป็นค่าร้อยละ และอัลกอริทึมที่ให้ค่าร้อยละของค่าพยากรณ์บวกสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Greenberg *et al.*, 2004; Witten and Frank, 2005)

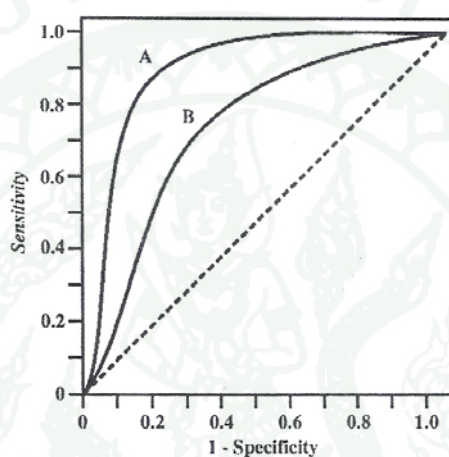
$$\text{PPV} = \frac{a}{a+b} \times 100$$

7. ค่าพยากรณ์ลบ (Negative Predictive Value) หมายถึง ผลการทดสอบที่ให้ผลลบ คือ ไม่เป็นโรค (d) เมื่อเทียบกับผลการทดสอบที่ให้ผลลัพธ์ไม่เป็นโรคทั้งหมด (c+d) แสดงเป็นค่าร้อยละ และอัลกอริทึมที่ให้ค่าร้อยละของค่าพยากรณ์ลบสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า (Greenberg *et al.*, 2004; Witten and Frank, 2005)

$$\text{NPV} = \frac{d}{c+d} \times 100$$

8. Area under the Receiver Operator Characteristic curve

Thomas (2004) และ Witten and Frank (2005) ได้อธิบาย Receiver Operator Characteristic (ROC) curve ว่าเป็นการแสดงความสัมพันธ์ระหว่าง True Positive Rate (Sensitivity) กับ False Positive Rate ($1 - \text{Specificity}$) ซึ่งค่าพื้นที่ใต้ ROC curve (AUC) สามารถใช้เปรียบเทียบประสิทธิภาพของอัลกอริทึมได้โดยไม่ต้องสนใจ โดยค่าร้อยละของ AUC สูงกว่าจะเป็นตัวจำแนกกลุ่มหรือประเภทข้อมูลที่ดีกว่าดังภาพที่ 5



ภาพที่ 5 ตัวอย่างการเปรียบเทียบค่า AUC

ที่มา: Greenberg *et al.* (2004)

จากภาพที่ 5 ถ้าค่า AUC ยิ่งมากแสดงถึงประสิทธิภาพของอัลกอริทึมจะมากขึ้นด้วย เช่น จากภาพพื้นที่ใต้โค้งของอัลกอริทึม B มีค่าน้อยกว่าอัลกอริทึม A แสดงว่าอัลกอริทึม A จะมีประสิทธิภาพมากกว่าอัลกอริทึม B (Greenberg *et al.*, 2004) และแปลความหมายของ AUC ดังตารางที่ 5

ตารางที่ 5 การแปลความหมายค่า AUC

ค่า AUC (%)	หมายถึง
< 51	ไม่ดี
51 – 60	ต้องปรับปรุง
61 – 70	ค่อนข้างพอใช้
71 – 80	ดีพอใช้
81 – 90	ดี
91 – 100	ดีเยี่ยม

ที่มา: Thomas (2004)

ส่วนที่ 2 แนวทางการซักประวัติผู้ป่วยโรคทางระบบหายใจ และความรู้เบื้องต้นเกี่ยวกับโรคระบบทางเดินหายใจ

แนวทางการซักประวัติผู้ป่วยโรคทางระบบหายใจ

การตรวจร่างกายโรคระบบหายใจเป็นการประเมินหน้าที่และการทำงานของระบบหายใจว่าขณะนั้นระบบหายใจอยู่ในเกณฑ์ปกติหรือผิดปกติ (Bickley and Barbara, 2002) การซักประวัติผู้ป่วยโรคระบบทางเดินหายใจที่จะนำไปสู่การวินิจฉัยที่ถูกต้องมีแนวทางดังนี้

1. อาการนำหรืออาการสำคัญที่เป็นสาเหตุให้ผู้ป่วยมาพบแพทย์
2. ระยะเวลาที่มีอาการของโรค
3. การดำเนินของโรค เช่น เป็นโรคติดต่อกันมาเป็นระยะเวลานานหรือเป็นๆ หายๆ
4. เหตุการณ์หรืออาการร่วมที่เกิดขึ้นและมีส่วนช่วยในการวินิจฉัยแยกโรค ซึ่งอาจจะเป็นสาเหตุของการเกิดปัญหาในผู้ป่วย

5. ประวัติที่จะช่วยบ่งชี้ว่าอาจมีการติดเชื้อร่วมด้วย เช่น อาการเหนื่อยหอบ อาการหายใจมีเสียงหวีดๆ ในโรคปอดอักเสบ อาการเจ็บคอ มีไข้ ในโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน เป็นต้น

6. ความรุนแรงของโรคที่เป็น มีผลต่อระบบหายใจของผู้ป่วยมากน้อยเพียงใด เช่น ทำให้เกิดมีหายใจลำบาก (Dyspnea) เมื่ออยู่เฉยๆ หรือเป็นเฉพาะเมื่อออกกำลังกาย เป็นต้น

7. ประวัติการรักษาก่อนหน้านี้และการตอบสนองต่อการรักษาที่ได้รับนั้น

8. ประวัติครอบครัว เช่น ประวัติของอาการหรืออาการแสดงอย่างเดียวกัน ในครอบครัว ประวัติโรคประจำตัวที่มีอยู่ทั้งของตนเองและครอบครัวตลอดจน คนใกล้ชิด ประวัติภาวะภูมิแพ้ วัณโรค เป็นต้น

9. ประวัติเกี่ยวกับภาวะแวดล้อม เช่น การสูบบุหรี่ของบุคคลในครอบครัว หรือ อาชีพของผู้ป่วย

10. ประวัติของอาการสำคัญทางระบบหายใจ ได้แก่ อาการไอ อาการหายใจลำบาก อาการหายใจมีเสียงหวีดๆ อาการเจ็บคอ เป็นต้น

สำหรับโรงพยาบาลพระนครศรีอยุธยา มีแนวทางการซักประวัติเพื่อจำแนกโรคดังนี้

1. ข้อมูลทั่วไปของผู้ป่วย ได้แก่ ชื่อ อายุ ที่อยู่ อาชีพ และโรคประจำตัว เป็นต้น

2. ตรวจร่างกายเบื้องต้น ได้แก่ น้ำหนัก ส่วนสูง อุณหภูมิร่างกาย และความดันโลหิต

3. ซักอาการสำคัญที่เป็นสาเหตุให้ผู้ป่วยมาพบแพทย์ สำหรับผู้ป่วยโรคระบบทางเดินหายใจ จะมีลักษณะอาการที่สำคัญ เช่น มีไข้ ไอ มีเสมหะ เสดด เจ็บคอ คัดจมูก หายใจเหนื่อยหอบ หายใจมีกลิ่นเหม็น จมูกไม่ได้กลิ่น ปวดศีรษะ เสียงแหบ แน่นหน้าอก เจ็บหน้าอก ปวดจมูก ปวดกระบอกตา และอาเจียน เป็นต้น

4. ระยะเวลาในการเกิดอาการผิดปกติ

หลังจากทราบประวัติในเบื้องต้น พยาบาลผู้ทำการคัดกรองจะทำการส่งประวัติไปให้แพทย์เป็นผู้ทำการวินิจฉัยต่อไป โดยแพทย์จะซักถามตามลักษณะของอาการที่เป็นสาเหตุให้ผู้ป่วยต้องมาพบแพทย์ เช่น หากมาด้วยอาการไอ แพทย์อาจทำการซักประวัติถึงจำนวนวันที่มีอาการ ช่วงเวลาที่มีอาการ เป็นต้น (พรทิพย์, 2552)

ความรู้เบื้องต้นเกี่ยวกับโรกระบบทางเดินหายใจ

การศึกษาครั้งนี้มุ่งความสนใจไปที่กลุ่มโรกระบบทางเดินหายใจ 3 โรค ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ซึ่งรายละเอียดของแต่ละโรคมีดังนี้

โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน

สุรเกียรติ (2544) ได้อธิบายความรู้เกี่ยวกับโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI) หรือไข้หวัดเป็นโรคที่พบได้บ่อยที่สุดทั้งในเด็กและผู้ใหญ่ ทั้งนี้เชื้อไวรัสที่เป็นสาเหตุของไข้หวัด (เชื้อหวัด) จะผลัดเปลี่ยนหมุนเวียนกันทำให้เกิดการอักเสบของทางเดินหายใจส่วนต้น (จมูกและคอ) ครึ่งละชนิด เมื่ออายุมากขึ้นร่างกายจะมีภูมิคุ้มกันต่อเชื้อหวัดชนิดต่าง ๆ มากขึ้นก็จะป่วยเป็นไข้หวัดห่างขึ้นและมีอาการรุนแรงน้อยลงไป โรคนี้สามารถติดต่อกันได้ง่ายโดยการอยู่ใกล้ชิดกัน มักจะพบมากในช่วงฤดูฝน ฤดูหนาว หรือในช่วงที่มีอากาศเปลี่ยนแปลง

สาเหตุ เกิดจากเชื้อไวรัส ซึ่งมีประมาณ 200 ชนิด การเกิดโรคขึ้นในแต่ละครั้งจะเกิดจากเชื้อหวัดเพียงชนิดเดียว เมื่อเป็นแล้วก็จะภูมิคุ้มกันต่อเชื้อหวัดชนิดนั้นในการเจ็บป่วยครั้งใหม่ก็จะเกิดจากเชื้อหวัดชนิดใหม่ หมุนเวียนเช่นนี้ไปเรื่อยๆ เชื้อหวัดมีอยู่ในน้ำมูก น้ำลาย และเสมหะของผู้ป่วยติดต่อโดยการไอ จาม หรือหายใจรดกัน นอกจากนี้เชื้อหวัดยังอาจติดต่อโดยการสัมผัสมือ กล่าวคือ เชื้อหวัดอาจติดมือของผู้ป่วยซึ่งเมื่อสัมผัสถูกมือของคนอื่นเชื้อหวัดก็จะติดมือของคนๆ นั้น และเมื่อใช้นิ้วมือขยี้ตา หรือแคะจมูกเชื้อก็จะเข้าสู่ร่างกายของคนๆ นั้น จนกลายเป็นไข้หวัด มีระยะฟักตัว (ระยะตั้งแต่ผู้ป่วยรับเชื้อเข้าไปจนกระทั่งมีอาการเกิดขึ้น)

1 - 3 วัน

อาการ มีไข้ตัวร้อนเป็นพักๆ ครั่นเนื้อครั่นตัว อ่อนเพลีย ปวดศีรษะเล็กน้อย คัดจมูก มีน้ำมูกใส จาม คอแห้ง หรือเจ็บคอเล็กน้อย ไอแห้งๆ หรือไอมีเสมหะเล็กน้อย ลักษณะสีขาว บางครั้งอาจรู้สึกเจ็บแสบลิ้นปี่เวลาไอ ในผู้ใหญ่อาจไม่มีไข้ มีเพียงคัดจมูก น้ำมูกใส ในเด็กมักเกิดอาการไข้ขึ้นมาทันทีทันใด บางครั้งอาจมีไข้สูงและชัก ท้องเดิน หรือถ่ายเป็นมูกร่วมด้วย ถ้ามีอาการเกิน 4 วัน อาจมีน้ำมูกข้นเหลืองหรือเขียว หรือมีเสมหะเป็นสีเหลืองหรือเขียว

สิ่งที่ตรวจพบ ไข้ น้ำมูก เยื่อจมูกบวมและแดง คอแดงเล็กน้อย ในเด็กอาจพบต่อมทอนซิลโตแต่ไม่แดงมาก และไม่มีหนอง

โรคไซนัสอักเสบเฉียบพลัน

สุรเกียรติ (2544) ได้ให้ความรู้เกี่ยวกับโรคไซนัสอักเสบเฉียบพลันว่า ไซนัส (sinus) หมายถึง โพรงอากาศเล็กๆ ในกะโหลกซึ่งอยู่รอบๆ จมูก และมีทางเชื่อมต่อกับโพรงจมูก ดังนั้น จึงอาจมีเชื้อโรคลุกลามจากโพรงจมูกเข้าไปในโพรงไซนัสได้ ตามปกติทางเชื่อมดังกล่าวจะเปิดโล่งให้มีการระบายของน้ำเมือกที่สร้างขึ้นในโพรงไซนัสได้สะดวกจึงไม่เกิดการอักเสบ แต่ถ้าหากทางเชื่อมดังกล่าวเกิดการอุดตันขึ้นมา เช่น เป็นหวัด ผนังกันจมูกคด มีเนื้องอกในจมูก ได้รับบาดเจ็บจากการนั่งเครื่องบิน หรือดื่มน้ำ ทำให้น้ำเมือกในโพรงไซนัสไม่สามารถระบายได้ ก็จะทำให้เชื้อโรคที่อยู่ในโพรงไซนัสสามารถเจริญงอกงามทำให้เกิดการอักเสบ และเป็นหนองขังภายในโพรงไซนัสได้

สาเหตุ เชื้อที่เป็นสาเหตุ ที่พบบ่อย ได้แก่ ปิตาสเตรปโตค็อกคัส, สแตฟฟีโลค็อกคัส, นิวโมค็อกคัส และฮีโมฟิลุสอินฟลูเอนซา เป็นต้น นอกจากนี้ยังอาจเกิดจากการลุกลามของเชื้อโรค จากบริเวณรากฟันที่เป็นหนองเข้าไปในโพรงไซนัสโดยตรงก็ได้ ไซนัสอักเสบมักพบเป็นโรคแทรกซ้อนของไข้หวัด, หวัดจากการแพ้, เยื่อจมูกอักเสบ, เนื้องอกในจมูก, ผนังกันจมูกคด และรากฟันเป็นหนอง เป็นต้น

อาการ มีอาการปวดมึนๆ หนักๆ ตรงบริเวณหัวตา หน้าผาก โหนกแก้มหรือรอบๆ กระบอกตา อาจรู้สึกคล้ายปวดฟัน บริเวณขากรรไกรบน อาการปวดอาจเป็นมากในเวลาเช้าหรือบ่าย เวลากำศีรษะหรือเปลี่ยนท่า ผู้ป่วยจะมีอาการคัดจมูก พุดเสียงขึ้นจมูก มีน้ำมูกข้น

เหลืองหรือเขียว เจ็บคอ มีเสมหะเหลืองหรือเขียวในลำคอ และอาจหายใจมีกลิ่นเหม็น ในรายที่เป็นไซนัสอักเสบเฉียบพลันมักมีไข้ร่วมด้วย

สิ่งตรวจพบ เชื้ออณูจุลชีววิทยา คอแดงเล็กน้อย ที่สำคัญจะพบว่า ถ้าเคาะหรือกดแรงๆ ตรงบริเวณหัวตา หน้าผาก หรือใต้ตาจะรู้สึกเจ็บ อาจมีไข้ในรายที่เป็นเฉียบพลัน

โรคปอดอักเสบ

สุรเกียรติ (2544) ได้อธิบายว่า ปอดอักเสบหรือปอดบวม หมายถึง การอักเสบของปอดซึ่งถือเป็นภาวะร้ายแรงชนิดหนึ่ง ภาษาอังกฤษเรียกว่า “นิวโมเนีย” (Pneumonia) มักพบในคนที่ไม่แข็งแรง (มีภูมิคุ้มกันโรคต่ำ) โดยอาจพบเป็นโรคแทรกซ้อนของไข้หวัด ไข้หวัดใหญ่ ต่อมทอนซิลอักเสบ หัด อีสุกอีใส และไทริน เป็นต้น

สาเหตุ เกิดจากเชื้อโรคหรือสารเคมีเข้าไปทำให้เกิดการอักเสบของปอด เชื้อโรคที่สำคัญได้แก่ 1. เชื้อแบคทีเรีย ซึ่งพบเป็นสาเหตุส่วนใหญ่ของโรคนี้ เช่น เชื้อนิวโมค็อกคัส (Pneumococcus) สแตฟฟีโลค็อกคัส (Staphylococcus) สเตรปโตค็อกคัส (Streptococcus) และเคล็บซิลลา (Klebsiella) 2. เชื้อไวรัส เช่น หัด ไข้หวัดใหญ่ อีสุกอีใส และไมโคพลาสมา (Mycoplasma pneumoniae) เป็นต้น 3. เชื้อรา พบได้ค่อนข้างน้อยแต่รุนแรง และ 4. สารเคมีที่พบบ่อย เช่น น้ำมันก๊าดซึ่งผู้ป่วยสูดดมเข้าไปในปอด

อาการ มักเกิดขึ้นทันทีทันใดด้วยอาการไข้สูง หนาวสั่น และหายใจหอบ ในระยะแรกอาจมีอาการไอแห้งๆ ไม่มีเสมหะต่อมาเสมหะจะข้นออกเป็นสีเหลืองสีเขียว สีสนิมเหล็ก หรือมีเลือดปน อาจมีอาการเจ็บแปลบในหน้าอกเวลาหายใจเข้าหรือเวลาไอแรงๆ บางครั้งอาจปวดร้าวไปที่หัวไหล่ สีข้างหรือท้อง ในเด็กเล็กอาจมีอาการปวดท้อง ท้องอืด ท้องเดิน อาเจียน กระสับกระส่ายหรือชัก

สิ่งตรวจพบ ไข้สูง 39 - 40 องศาเซลเซียส หน้าแดง ริมฝีปากแดง ลิ้นเป็นฝ้า หายใจตั้งแต่ท้องอกมากกว่า 40 ครั้งต่อนาที ซึ่งโครงบวม รุนแรงมาก ถ้าเป็นมากอาจมีอาการตัวเขียว (ริมฝีปากเขียว ลิ้นเขียว เล็บเขียว) หรือภาวะขาดน้ำ บางรายอาจมีริมฝีปาก ปอด อาจเคาะทึบ (Dullness) ใช้เครื่องฟังตรวจปอดอาจมีเสียงหายใจค่อย (Diminished breath sound) หรือมีเสียงกรอบแกรบ (Crepitation) ซึ่งมักจะได้ยินตรงใต้สะบักทั้ง 2 ข้าง

ส่วนที่ 3 ผลงานวิจัยที่เกี่ยวข้อง

Soman and Bobbie (2005) ได้ทำการศึกษาการจำแนกประเภทผู้ป่วยที่เป็นโรคหัวใจ เติ้นฝึกปกติด้วยอัลกอริทึม One R, C4.5 และ Naïve Bayes โดยแบ่งข้อมูลที่ใช้ในการฝึกสอน และข้อมูลทดสอบ เป็น 4 เงื่อนไข ประกอบด้วย 1. ใช้ข้อมูลฝึกสอน 100% แล้วนำข้อมูลชุดเดียวกันมาเป็นข้อมูลทดสอบ ผลปรากฏว่า อัลกอริทึม C4.5 (J48) มีความถูกต้องในการทำนาย สูงสุด (91.81%) 2. ชุดข้อมูลฝึกสอน 50% และข้อมูลทดสอบ 50% ผลการทดลองปรากฏว่า อัลกอริทึม Naïve Bayes มีความถูกต้องในการทำนายสูงสุด (70.80%) 3. ชุดข้อมูลฝึกสอน 70% และชุดข้อมูลทดสอบ 30% ผลการทดลองปรากฏว่าอัลกอริทึม Naïve Bayes มีความถูกต้องใน การทำนายสูงสุด (75%) 4. ชุดข้อมูลฝึกสอน 80% และชุดข้อมูลทดสอบ 20% ผลการทดลอง ปรากฏว่าอัลกอริทึม Naïve Bayes มีความถูกต้องในการทำนายสูงสุด (74.73%) จากการแบ่งข้อมูล ชุดฝึกสอนและชุดทดสอบ สำหรับอัลกอริทึม C4.5 จะได้ผลการทำนายสูงสุดในกรณีที่ใช้ข้อมูล ฝึกสอน 100%

Kamila (2007) ได้ทำการศึกษา การเลือกใช้อัลกอริทึมเหมืองข้อมูลที่เหมาะสมสำหรับ ระบบสนับสนุนการตัดสินใจทางการแพทย์ โดยทำการเปรียบเทียบผลของอัลกอริทึมเหมืองข้อมูล อันได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ (ID3, C4.5) อัลกอริทึมโครงข่ายประสาทเทียม และ อัลกอริทึม Naïve Bayes กับกลุ่มตัวอย่าง คือ ผู้ป่วยโรคเบาหวานจำนวน 768 คน ผู้ป่วย โรคมะเร็งเต้านมจำนวน 699 คน ผู้ป่วยโรคตับอักเสบจำนวน 155 คน ผู้ป่วยโรคหัวใจ จำนวน 303 คน และผู้ป่วยโรคผิวหนังจำนวน 366 คน จากผลการศึกษาพบว่า ในกลุ่มผู้ป่วยโรคเบาหวาน โรคผิวหนัง และโรคหัวใจ อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด ได้แก่ อัลกอริทึม Naïve Bayes ให้ค่าความถูกต้อง 76.30 % 97.40% และ 55.50% ตามลำดับ และสำหรับกลุ่มผู้ป่วย โรคมะเร็งเต้านมและโรคตับอักเสบ อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด ได้แก่ อัลกอริทึม C4.5 ให้ค่าความถูกต้อง 96.00% และ 79.10 % ตามลำดับ

Porcel *et al.* (2008) ได้ทำการศึกษา ระบบ Decision Tree เพื่อหาจำแนกผู้ป่วยโรค วัณโรคและผู้ป่วยโรคเยื่อหุ้มปอดอักเสบ ทำการศึกษากลุ่มตัวอย่างทั้งหมด 238 คน โดยแบ่งเป็น ผู้ป่วยโรควัณโรคจำนวน 64 คน และผู้ป่วยโรคเยื่อหุ้มปอดอักเสบจำนวน 174 คน ทำการศึกษา อัลกอริทึม C4.5 ที่เป็นอัลกอริทึมหนึ่งในระบบ Decision Tree และเลือกใช้วิธีการแบ่งข้อมูลแบบ 10-Folds Cross-validation สำหรับการแบ่งชุดข้อมูลเพื่อให้ระบบเรียนรู้ มีตัวแปรทั้งหมด 12 ตัวแปร ผลการศึกษาพบว่า ได้ค่าความไว 92.20% ค่าความจำเพาะ 98.30% และได้ค่า ROC

0.976 ซึ่งจากผลการวิเคราะห์ Decision Tree สามารถช่วยในการจำแนกกลุ่มผู้ป่วยโรคฉี่หนู และผู้ป่วยโรคเชื้อเห็บปอดอักเสบได้อย่างดี

พฤษดี (2551) ได้ทำการศึกษาระบบจำแนกประเภทแบบทดสอบสำหรับผู้ทดสอบ สุขภาพจิต ด้วยเทคนิคต้นไม้ตัดสินใจกับกลุ่มตัวอย่าง 550 ตัวอย่าง ผลการวิเคราะห์ สามารถ จำแนกโมเดลได้ทั้งสิ้นจำนวน 15 แบบจำลอง และแบบจำลองที่ได้จากการศึกษาครั้งนี้ได้ถูกนำมา เป็นองค์ความรู้เพื่อการศึกษาต่อไป

วรารณ (2551) ได้ทำการศึกษาวิเคราะห์ปัจจัยเสี่ยงและการจัดจำแนกกลุ่มของการติดเชื้อ ในผู้ป่วยวัณโรคปอดที่เกิดโรคกลับด้วยการวิเคราะห์การถดถอยโลจิสติก และการวิเคราะห์ Classification Tree โดยทำการศึกษาในกลุ่มผู้ป่วยจำนวน 196 ราย ผลการศึกษาพบว่า การ วิเคราะห์การถดถอยโลจิสติกมีค่าความไว 24.10 ค่าความจำเพาะ 95.50 ค่าพยากรณ์บวก 70 และ ค่าพยากรณ์ลบ 74.10 สำหรับการวิเคราะห์ Classification Tree ให้ค่าความไว 16.70 ค่า ความจำเพาะ 98.50 ค่าพยากรณ์บวก 83.30 และค่าพยากรณ์ลบ 72.80 เมื่อพิจารณาค่าความไว ค่าความจำเพาะ และค่าพยากรณ์ลบ จะให้ค่าไม่แตกต่างกันมาก แต่สำหรับค่าพยากรณ์บวกจะ พบว่าการวิเคราะห์ Classification Tree ให้ค่าที่สูงกว่าการวิเคราะห์การถดถอยโลจิสติก

จากการศึกษาแนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้องดังกล่าวข้างต้น พบว่ามีผู้นำวิธี ต้นไม้ตัดสินใจไปประยุกต์ใช้กับงานด้านการแพทย์ และเลือกใช้อัลกอริทึมของต้นไม้ตัดสินใจที่ แตกต่างกัน ผู้วิจัยจึงสนใจที่จะศึกษาความสามารถและประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ โดยทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึม ID3, C4.5 และ CART ในการ จำแนกผู้ป่วย 3 โรค คือ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไข้น้ำสออักเสบเฉียบพลัน และโรคปอดอักเสบ เพื่อนำผลที่ได้ไปสนับสนุนระบบการคัดกรอง และอาจใช้เป็น แนวทางการคัดกรองผู้ป่วยในเบื้องต้นเพื่อการวินิจฉัยต่อไป

อุปกรณ์และวิธีการ

อุปกรณ์

ในการวิเคราะห์และประมวลผลข้อมูลใช้โปรแกรม Microsoft Excel โปรแกรม Weka และโปรแกรม SPSS กับเครื่องคอมพิวเตอร์ของภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน

วิธีการ

การศึกษาความสามารถและประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ 3 อัลกอริทึม คือ ID3 C4.5 และ CART โดยมีขั้นตอนการวิจัยดังต่อไปนี้

1. การเตรียมข้อมูล

ข้อมูลที่ใช้ศึกษาเป็นข้อมูลเวชระเบียนผู้ป่วยนอกของโรงพยาบาลพระนครศรีอยุธยา เฉพาะผู้ที่เข้ารับการรักษาเป็นครั้งแรกระหว่างปี พ.ศ. 2547-2549 ด้วยกลุ่มโรคระบบทางเดินหายใจ ใน 3 โรคต่อไปนี้ คือ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และ โรคปอดอักเสบ ตั้งแต่แสดงอาการจนกระทั่งได้รับการวินิจฉัยในช่วง 2 สัปดาห์ มีจำนวน 7,324 คน ในที่นี้ตัวแปรตาม คือ ผลการวินิจฉัยจากแพทย์ที่ระบุว่าเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน หรือโรคปอดอักเสบ ส่วนตัวแปรอิสระที่เกี่ยวข้องกับผู้ป่วยมี 22 ตัวแปร ได้แก่ อายุ (ปี) อุณหภูมิร่างกาย (องศาเซลเซียส) เขตที่อยู่อาศัย อาชีพ อาการมีน้ำมูก มีเสมหะ มีไข้ คัดจมูก แน่นหน้าอก ปวดกระบอกตา ปวดศีรษะ ปวดจมูก ปวดหู เวียนศีรษะ เจ็บคอ เจ็บหน้าอก หายใจเหนื่อยหอบ/หายใจลำบาก หายใจมีเสียงหวีดๆ หายใจมีกลิ่นเหม็น ไอ และอาเจียน รายละเอียดตัวแปร และรหัสข้อมูลแสดงในตารางที่ 6

ตารางที่ 6 รายละเอียดของตัวแปรและการกำหนดรหัสที่ศึกษา

ตัวแปร	คำอธิบาย
ตัวแปรตาม (ตัวแปรเป้าหมาย)	
Diseases	โรคระบบทางเดินหายใจ ได้แก่
- ACUTE URI	- โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
- ACUTE SINUSITIS	- โรคไซนัสอักเสบเฉียบพลัน
- PNEUMONIA	- โรคปอดอักเสบ
ตัวแปรอิสระ	
ตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ ได้แก่	
อายุ	อายุ (ปี)
อุณหภูมิร่างกาย	อุณหภูมิร่างกาย (องศาเซลเซียส)
ตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพ ได้แก่	
เพศ	เพศ (1 = ชาย, 2 = หญิง)
ที่อยู่อาศัย	ที่อยู่อาศัย (1 = ในเขตเทศบาล, 2 = นอกเขตเทศบาล, 3 = อื่นๆ)
อาชีพ	อาชีพ (1 = ในความปกครอง, 2 = นักเรียน/นักศึกษา, 3 = ข้าราชการ, 4 = พนักงานรัฐวิสาหกิจ, 5 = พนักงานบริษัทเอกชน, 6 = รับจ้าง, 7 = ธุรกิจส่วนตัว, 8 = กสิกรรม, 9 = อื่นๆ)
อาการไข้	อาการไข้ (0 = ไม่มีอาการไข้, 1 = มีอาการไข้)
อาการปวดศีรษะ	อาการปวดศีรษะ (0 = ไม่มีอาการปวดศีรษะ, 1 = มีอาการปวดศีรษะ)
อาการเวียนศีรษะ	อาการเวียนศีรษะ (0 = ไม่มีอาการเวียนศีรษะ, 1 = มีอาการเวียนศีรษะ)
อาการเจ็บคอ	อาการเจ็บคอ (0 = ไม่มีอาการเจ็บคอ, 1 = มีอาการเจ็บคอ)
มีเสมหะ	มีเสมหะ (0 = ไม่มีเสมหะ, 1 = มีเสมหะ)
อาการไอ	อาการไอ (0 = ไม่มีอาการไอ, 1 = มีอาการไอ)
มีน้ำมูก	มีน้ำมูก (0 = ไม่มีน้ำมูก, 1 = มีน้ำมูกใส, 2 = มีน้ำมูกสีขาวขุ่น, 3 = มีน้ำมูกสีเหลือง, 4 = มีน้ำมูกสีเขียว)
อาการคัดจมูก	อาการคัดจมูก (0 = ไม่มีอาการคัดจมูก, 1 = มีอาการคัดจมูก)
อาการอาเจียน	อาการอาเจียน (0 = ไม่มีอาการอาเจียน, 1 = มีอาการอาเจียน)

ตารางที่ 6 (ต่อ)

ตัวแปร	คำอธิบาย
อาการหายใจเหนื่อยหอบ/ หายใจลำบาก	อาการหายใจเหนื่อยหอบ/หายใจลำบาก (0 = ไม่มีอาการหายใจเหนื่อยหอบ/หายใจลำบาก, 1 = มีอาการหายใจเหนื่อยหอบ/หายใจลำบาก)
อาการหายใจมีเสียงหวีดๆ	อาการหายใจมีเสียงหวีดๆ (0 = ไม่มีอาการหายใจมีเสียงหวีดๆ, 1 = มีอาการหายใจมีเสียงหวีดๆ)
อาการหายใจมีกลิ่นเหม็น	อาการหายใจมีกลิ่นเหม็น (0 = ไม่มีอาการหายใจมีกลิ่นเหม็น, 1 = มีอาการหายใจมีกลิ่นเหม็น)
อาการแน่นหน้าอก	อาการแน่นหน้าอก (0 = ไม่มีอาการแน่นหน้าอก, 1 = มีอาการแน่นหน้าอก)
อาการเจ็บหน้าอก	อาการเจ็บหน้าอก (0 = ไม่มีอาการเจ็บหน้าอก, 1 = มีอาการเจ็บหน้าอก)
อาการปวดกระบอกตา	อาการปวดกระบอกตา (0 = ไม่มีอาการปวดกระบอกตา, 1 = มีอาการปวดกระบอกตา)
อาการปวดจมูก	อาการปวดจมูก (0 = ไม่มีอาการปวดจมูก, 1 = มีอาการปวดจมูก)
อาการปวดหู	อาการปวดหู (0 = ไม่มีอาการปวดหู, 1 = มีอาการปวดหู)

2. วิเคราะห์ข้อมูลเบื้องต้น

ศึกษาลักษณะพื้นฐานของข้อมูลด้วยค่าสถิติ จำนวน ร้อยละ ค่าเฉลี่ย และส่วนเบี่ยงเบน
มาตรฐาน

3. การประยุกต์อัลกอริทึมต้นไม้ตัดสินใจ

ทำการวิเคราะห์ความสามารถของอัลกอริทึมต้นไม้ตัดสินใจ 3 อัลกอริทึม คือ ID3, C4.5 และ CART โดยการจำแนกโรคระบบทางเดินหายใจ 3 โรค

3.1 การคัดเลือกตัวแปร

พิจารณาการคัดเลือกตัวแปรเป็น 2 กรณี คือ

3.1.1 กรณีไม่มีการคัดเลือกตัวแปร

3.1.2 กรณีมีการคัดเลือกตัวแปร

สำหรับกรณีมีการคัดเลือกตัวแปร ใช้โปรแกรม Weka กับอัลกอริทึมการคัดเลือกตัวแปร 3 วิธี ได้แก่ 1) วิธี Best First 2) วิธี Greedy Stepwise และ 3) วิธี Genetic Search กับชุดข้อมูลฝึกสอนที่ได้จัดเตรียมไว้ 2 แบบ คือ 1) Use full training set คือ การใช้ข้อมูลทั้งหมดเป็นชุดข้อมูลฝึกสอน และ 2) 10 Folds Cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วนโดยใช้ข้อมูล 9 ส่วนเป็นชุดข้อมูลฝึกสอนและส่วนที่เหลือเป็นชุดข้อมูลทดสอบ โดยทำการเปลี่ยนชุดข้อมูลทดสอบจนครบทุกส่วน

3.2 การสร้างตัวจำแนก

ขั้นตอนนี้ใช้โปรแกรม Weka สำหรับอัลกอริทึม ID3 และ C4.5 และใช้โปรแกรม SPSS สำหรับอัลกอริทึม CART กับชุดข้อมูลฝึกสอน (Training Set) และชุดข้อมูลทดสอบ (Testing Set) ที่มาจากการแบ่งข้อมูล 2 แบบ คือ

แบบที่ 1 ใช้หลักการแบ่งข้อมูลแบบ Percentage split โดยกำหนดอัตราส่วนระหว่างชุดข้อมูลฝึกสอนกับชุดข้อมูลทดสอบ ดังนี้

1.1) 50% split หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = 50% : 50%

1.2) 60% split หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = 60% : 40%

1.3) 70% split หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = 70% : 30%

แบบที่ 2 ใช้หลักในการแบ่งข้อมูล แบบ k- Folds Cross-validation โดยแบ่งชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ ดังนี้

2.1) 5 – Folds Cross-validation

2.2) 10 – Folds Cross-validation

2.3) 15 – Folds Cross-validation

แสดงตัวอย่างการแบ่งข้อมูลแบบ 5 – Folds Cross-validation คือ จะทำการแบ่งข้อมูลออกเป็น 5 ส่วนในครั้งแรกข้อมูลส่วนที่ 1, 2, 3 และ 4 จะใช้เป็นชุดข้อมูลฝึกสอน สำหรับส่วนที่ 5 จะถูกใช้เป็นชุดข้อมูลทดสอบ ครั้งที่สอง จะเก็บข้อมูลส่วนที่ 4 ไว้เป็นชุดข้อมูลทดสอบ สำหรับส่วนที่ 1, 2, 3 และ 5 ไว้เป็นชุดข้อมูลฝึกสอน ทำเช่นนี้ต่อไปจนทุกส่วนถูกใช้เป็นชุดข้อมูลทดสอบ การสร้างตัวจำแนกจะได้จากการที่ให้ระบบเรียนรู้จากชุดข้อมูลฝึกสอน

Fold ที่ 1

Training Set	Training Set	Training Set	Training Set	Testing Set
--------------	--------------	--------------	--------------	-------------

Fold ที่ 2

Training Set	Training Set	Training Set	Testing Set	Training Set
--------------	--------------	--------------	-------------	--------------

Fold ที่ 3

Training Set	Training Set	Testing Set	Training Set	Training Set
--------------	--------------	-------------	--------------	--------------

Fold ที่ 4

Training Set	Testing Set	Training Set	Training Set	Training Set
--------------	-------------	--------------	--------------	--------------

Fold ที่ 5

Testing Set	Training Set	Training Set	Training Set	Training Set
-------------	--------------	--------------	--------------	--------------

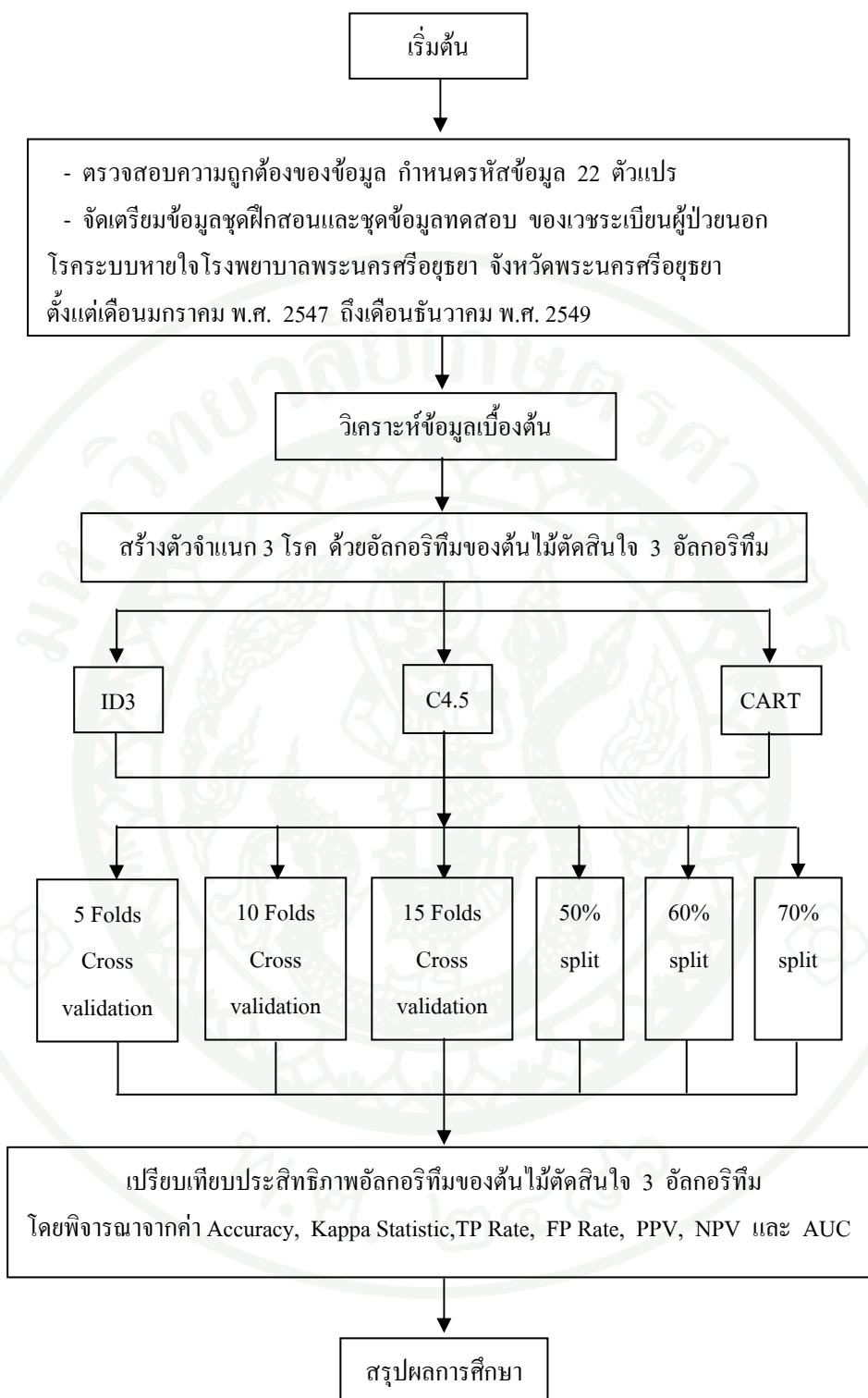
ภาพที่ 6 แสดงการแบ่งข้อมูลเพื่อการเรียนรู้การสร้างตัวจำแนกแบบ 5 – Folds Cross-validation

3.3 การทดสอบตัวจำแนก

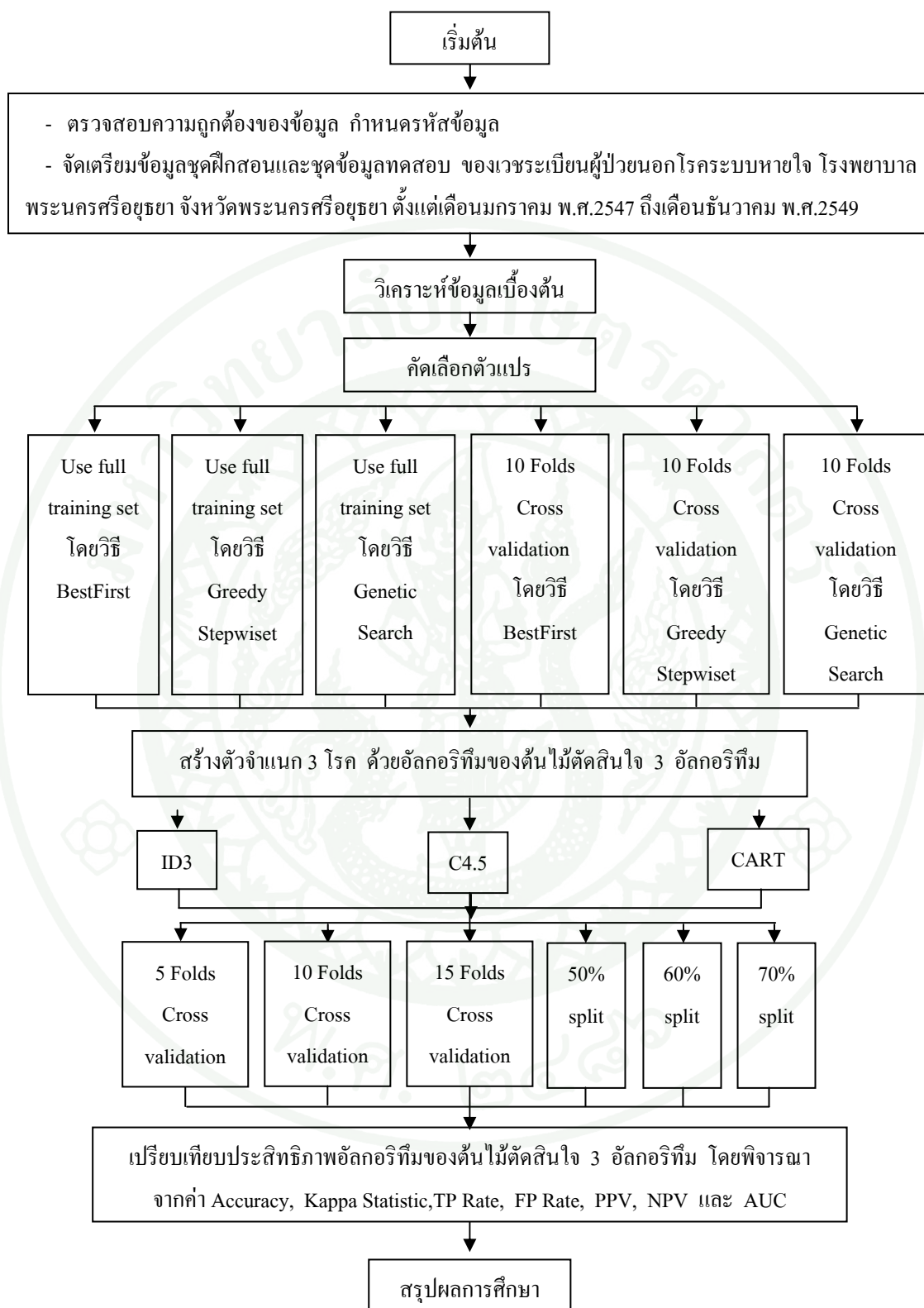
เพื่อทดสอบความถูกต้องของตัวจำแนกที่ได้จากขั้นตอนการสร้างตัวจำแนกกับชุดข้อมูลฝึกสอน จึงนำตัวจำแนกดังกล่าวมาใช้กับชุดข้อมูลทดสอบ โดยนำข้อมูลเข้าสู่ระบบเพื่อให้ระบบทำนาย และพิจารณาเปรียบเทียบประสิทธิภาพของตัวจำแนกที่ได้จากแต่ละอัลกอริทึมต้นไม้มัดสินใจ ด้วยค่าวัดประสิทธิภาพต่อไปนี้ ได้แก่ ค่า Accuracy, Kappa Statistic, TP Rate, FP Rate, PPV, NPV และ AUC

3.4 กรอบแนวทาง การสร้างตัวจำแนก

แสดงในภาพที่ 7 และ 8



ภาพที่ 7 ขั้นตอนการพัฒนาตัวจำแนกโรคระบบทางเดินหายใจ ด้วยอัลกอริทึมต้นไม้ตัดสินใจ D3, C4.5 และ CART กรณีไม่มีการคัดเลือกตัวแปร



ภาพที่ 8 ขั้นตอนการพัฒนาตัวจำแนกโรคระบบทางเดินหายใจ ด้วยอัลกอริทึมต้นไม้ตัดสินใจ ID3, C4.5 และ CART กรณีมีการคัดเลือกตัวแปร

สถานที่และระยะเวลาทำการวิจัย

สถานที่ทำการวิจัยครั้งนี้ คือ ภาควิชาสถิติ คณะวิทยาศาสตร์
มหาวิทยาลัยเกษตรศาสตร์ ระยะเวลาในการวิจัยเริ่มตั้งแต่ เดือนพฤศจิกายน พ.ศ. 2550 ถึง เดือน
ตุลาคม พ.ศ. 2552

ประโยชน์ที่คาดว่าจะได้รับ

ได้แบบจำลองต้นไม้ตัดสินใจที่ง่าย สามารถนำไปช่วยสนับสนุนการตรวจคัดกรองการ
จำแนกผู้ป่วยโรคระบบทางเดินหายใจของโรงพยาบาลพระนครศรีอยุธยา และขยายผลไปใช้กับ
โรงพยาบาลที่สนใจ หรือเป็นแนวทางการพัฒนาการคัดกรองผู้ป่วยโรคอื่นได้

ผลและวิจารณ์

ผล

จากวัตถุประสงค์ที่ต้องการศึกษาและประยุกต์วิธีสืบค้นความรู้จากฐานข้อมูล โดยใช้ อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรกระบบทางเดินหายใจ สำหรับจำแนกผู้ป่วยโรกระบบทางเดินหายใจ 3 โรค ของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา และเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ 3 อัลกอริทึม คือ ID3, C4.5 และ CART ในการจำแนกผู้ป่วยทั้ง 3 โรคที่ระบุ

ผลการศึกษาสรุปเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลพื้นฐานของผู้ป่วย ส่วนที่ 2 ประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ ส่วนที่ 3 ผลที่ได้จากการประยุกต์อัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ป่วย 3 โรค

1. ข้อมูลพื้นฐานของผู้ป่วย

จากการวิเคราะห์ข้อมูลเวชระเบียนผู้ป่วยกลุ่มโรกระบบทางเดินหายใจที่เข้ารับการรักษาครั้งแรกด้วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI) โรคไซนัสอักเสบเฉียบพลัน (Acute Sinusitis) และโรคปอดอักเสบ (Pneumonia) ของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา ระหว่างปี พ.ศ. 2547 – 2549 พบว่าผู้ป่วยนอกที่เข้ารับการรักษาครั้งแรกกลุ่มโรกระบบทางเดินหายใจ มีจำนวน 7,327 คน เป็นเพศชายร้อยละ 51.40 และเพศหญิงร้อยละ 48.60 มีอายุส่วนใหญ่อยู่ในกลุ่ม 25 – 59 ปี จำนวน 3,649 คน หรือร้อยละ 49.80 รองลงมาคือกลุ่มอายุ 15 – 24 ปี จำนวน 2,485 คน คิดเป็นร้อยละ 33.90 ส่วนใหญ่ผู้ป่วยร้อยละ 61.00 อาศัยอยู่นอกเขตเทศบาล ที่เหลืออาศัยอยู่ในเขตเทศบาล และอื่น ๆ โดยมากประกอบอาชีพรับจ้าง ร้อยละ 42.90 รองลงมา เป็นนักเรียน/นักศึกษา ร้อยละ 27.10 ผู้ป่วยกลุ่มโรกระบบทางเดินหายใจร้อยละ 47.90 เป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน รองลงมาคือร้อยละ 29.10 เป็นโรคไซนัสอักเสบเฉียบพลัน และที่เหลือเป็นโรคปอดอักเสบ ดังแสดงในตารางที่ 7

ตารางที่ 7 จำนวนและร้อยละของผู้ป่วยกลุ่มโรคระบบทางเดินหายใจ ที่เข้ารับการรักษารั้งแรก
ด้วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบ
เฉียบพลัน และโรคปอดอักเสบของโรงพยาบาลพระนครศรีอยุธยา จังหวัด
พระนครศรีอยุธยา ปี พ.ศ. 2547 – 2549

(n = 7,327)

ตัวแปร	จำนวน (คน)	ร้อยละ
เพศ		
ชาย	3,766	51.40
หญิง	3,561	48.60
อายุ (ปี)		
น้อยกว่า 1 ปี	224	3.10
1 – 14	798	10.90
15 – 24	2,485	33.90
25 – 59	3,649	49.80
60 ขึ้นไป	171	2.30
อายุเฉลี่ย($\bar{X} \pm S.D.$) = 26.37 ± 12.85		
ที่อยู่อาศัย		
ในเขตเทศบาล	2,382	32.50
นอกเขตเทศบาล	4,466	61.00
อื่น ๆ	479	6.50
อาชีพ		
ในความปกครอง	408	5.60
นักเรียน/นักศึกษา	1,984	27.10
ข้าราชการ	490	6.70
พนักงานรัฐวิสาหกิจ	295	4.00
พนักงานบริษัทเอกชน	355	4.80
รับจ้าง	3,140	42.90
ธุรกิจส่วนตัว	286	3.90

ตารางที่ 7 (ต่อ)

(n = 7,327)		
ตัวแปร	จำนวน (คน)	ร้อยละ
กสิกรรม	232	3.20
อื่นๆ	137	1.90
โรคระบบทางเดินหายใจ		
โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน	3,513	47.90
โรคไซนัสอักเสบเฉียบพลัน	2,132	29.10
โรคปอดอักเสบ	1,682	23.00
อุณหภูมิร่างกาย(องศาเซลเซียส)		
ปกติ	5,058	69.00
ผิดปกติ	2,269	31.00

โดยผู้ป่วยที่เข้ามารับการรักษาด้วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน มาด้วยอาการมีไข้เป็นอันดับหนึ่ง จำนวน 2,257 คน คิดเป็นร้อยละ 64.20 อาการมีน้ำมูกเป็นอันดับสอง จำนวน 2,053 คน คิดเป็นร้อยละ 58.40 และอันดับสามมาด้วยอาการไอ จำนวน 2,002 คน คิดเป็นร้อยละ 57.00

สำหรับผู้ป่วยที่เข้ามารับการรักษาด้วยโรคไซนัสอักเสบเฉียบพลัน มาด้วยอาการมีน้ำมูกเป็นอันดับหนึ่ง จำนวน 1,038 คน คิดเป็นร้อยละ 48.70 รองลงมาคือ อาการมีเสมหะ จำนวน 1,237 คน คิดเป็นร้อยละ 60.80 และอันดับสามมาด้วยอาการปวดจมูก จำนวน 933 คน คิดเป็นร้อยละ 43.80

และสำหรับผู้ป่วยที่เข้ามารับการรักษาด้วยโรคปอดอักเสบ มาด้วยอาการไอ เป็นอันดับหนึ่ง จำนวน 1,185 คน คิดเป็นร้อยละ 70.50 รองลงมาคือ อาการหายใจเหนื่อยหอบ/หายใจลำบาก จำนวน 830 คน คิดเป็นร้อยละ 49.80 และอันดับสามมาด้วยอาการมีน้ำมูก จำนวน 650 คน คิดเป็นร้อยละ 38.60 รายละเอียดดังตารางที่ 8

ตารางที่ 8 จำนวนและร้อยละ ของลักษณะอาการของผู้ป่วยที่เข้ามารับการรักษาครั้งแรกด้วย
โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน
และโรคปอดอักเสบของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา
ระหว่างปี พ.ศ. 2547 – 2549

ลักษณะอาการ	Acute URI (n = 3,513)	Acute Sinusitis (n = 2,132)	Pneumonia (n = 1,682)
มีอาการไข้	2,257 (64.20)	127 (6.00)	482 (28.70)
มีอาการปวดศีรษะ	874 (24.90)	472 (22.10)	145 (8.60)
มีอาการเวียนศีรษะ	399 (11.40)	16 (0.80)	17 (1.00)
มีอาการเจ็บคอ	1,998 (56.90)	357 (16.70)	17 (1.00)
มีเสมหะ	1,160 (33.00)	1,038 (48.70)	576 (34.20)
มีอาการไอ	2,002 (57.00)	647 (30.30)	1,185 (70.50)
มีอาการอาเจียน	49 (1.40)	290 (13.60)	437 (26.00)
มีน้ำมูก	2,053 (58.40)	1,237 (60.80)	650 (38.60)
- มีน้ำมูกใส	1,779 (50.60)	770 (36.10)	165 (9.80)
- มีน้ำมูกสีขาวยุ่น	71 (2.00)	111 (5.20)	378 (22.50)

ตารางที่ 8 (ต่อ)

ลักษณะอาการ	Acute URI (n = 3,513)	Acute Sinusitis (n = 2,132)	Pneumonia (n = 1,682)
- มีน้ำมูกสีเหลือง	62 (1.80)	290 (13.60)	79 (4.70)
- มีน้ำมูกสีเขียว	141 (4.00)	126 (5.90)	28 (1.70)
มีอาการคัดจมูก	353 (10.00)	443 (20.80)	203 (12.10)
มีอาการเหนื่อยหอบ/หายใจลำบาก	0 (0.00)	323 (15.20)	830 (49.30)
มีอาการหายใจมีเสียงหวีด	0 (0.00)	0 (0.00)	484 (28.80)
มีอาการหายใจมีกลิ่นเหม็น	0 (0.00)	514 (24.10)	0 (0.00)
มีอาการแน่นหน้าอก	0 (0.00)	0 (0.00)	234 (13.90)
มีอาการเจ็บหน้าอก	0 (0.00)	0 (0.00)	347 (20.60)
มีอาการปวดกระบอกตา	0 (0.00)	267 (12.50)	0 (0.00)
มีอาการปวดจมูก	0 (0.00)	933 (43.80)	0 (0.00)
มีอาการปวดหู	0 (0.00)	483 (22.70)	0 (0.00)

หมายเหตุ ตัวเลขที่อยู่ในวงเล็บ () หมายถึง ค่าร้อยละเมื่อเทียบกับจำนวนผู้ป่วย (n) ของแต่ละโรค

2. ประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจ

ในการเปรียบเทียบประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจได้ทำการเปรียบเทียบ 3 อัลกอริทึม ได้แก่ ID3, C4.5 และ CART โดยพิจารณาเป็น 2 กรณี ได้แก่ กรณีที่ไม่มีการคัดเลือกตัวแปร คือ ใช้ทุกตัวแปรของข้อมูล ซึ่งมีทั้งหมด 22 ตัวแปร และกรณีที่มีการคัดเลือกตัวแปร ใช้อัลกอริทึมการคัดเลือกตัวแปร 3 วิธี คือ วิธี Best First, วิธี Greedy Stepwise และวิธี Genetic Search กับข้อมูลฝึกสอนแบบ Use full training set และแบบ 10 Folds Cross-validation

2.1 กรณีไม่มีการคัดเลือกตัวแปร

ผลการศึกษากรณีที่ไม่มีคัดเลือกตัวแปร สำหรับอัลกอริทึม ID3, C4.5 และ CART ซึ่งมีการแบ่งชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบเป็น 6 แบบ ได้แก่ 5 Folds Cross-validation, 10 Folds Cross-validation, 15 Folds Cross-validation, 50% split, 60% split และ 70% split ได้ผลการศึกษาโดยจำแนกตามโรคดังนี้

ตารางที่ 9 เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูล

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
ID3						
Accuracy	99.00%	99.40%	99.41%	98.75%	98.70%	99.13%
Kappa Statistic	0.9822	0.9849	0.9852	0.9802	0.9800	0.9844
TP Rate	99.12%	99.17%	99.23%	98.73%	99.11%	99.40%
FP Rate	0.45%	0.75%	0.75%	0.95%	1.05%	0.80%
PPV	99.03%	99.26%	99.23%	99.19%	98.74%	98.91%
NPV	99.18%	99.24%	99.29%	98.86%	99.24%	99.49%
AUC	99.20%	99.30%	99.35%	99.95%	99.00%	99.15%
C4.5						
Accuracy	99.17%	99.18%	99.18%	98.61%	99.18%	99.42%
Kappa Statistic	0.9833	0.9836	0.9836	0.9721	0.9835	0.9881
TP Rate	98.98%	99.01%	99.09%	98.05%	98.82%	99.31%
FP Rate	0.80%	0.65%	0.80%	0.95%	0.80%	0.60%
PPV	99.29%	99.01%	99.20%	99.01%	99.41%	99.40%
NPV	99.06%	99.34%	99.16%	98.25%	98.99%	99.41%
AUC	99.60%	99.50%	99.60%	99.20%	99.50%	99.70%
CART						
Accuracy	96.72%	96.72%	96.72%	97.07%	97.20%	96.86%
Kappa Statistic	0.9344	0.9344	0.9344	0.9405	0.9438	0.9369
TP Rate	93.80%	93.80%	93.80%	90.04%	89.32%	85.48%
FP Rate	8.81%	8.81%	8.81%	10.64%	11.63%	5.84%

ตารางที่ 9 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
PPV	90.00%	90.00%	90.00%	88.48%	86.50%	94.29%
NPV	94.52%	94.52%	94.52%	90.81%	90.83%	85.18%
AUC	99.10%	99.10%	99.10%	98.90%	99.20%	99.20%

จากตารางที่ 9 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 15 Folds Cross-validation มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อใช้กับการแบ่งข้อมูลแบบ 70% split

และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม C4.5 ที่ใช้การแบ่งข้อมูลแบบ 70% split ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โดยให้ค่า Accuracy เท่ากับ 99.42% ค่า TP Rate เท่ากับ 99.31% ค่า FP Rate เท่ากับ 0.60% ค่า PPV เท่ากับ 99.40% ค่า NVP เท่ากับ 99.41% และค่า AUC เท่ากับ 99.70%

ตารางที่ 10 เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับจำแนกผู้ป่วยโรคไตเรื้อรังแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูล

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
ID3						
Accuracy	99.32%	99.37%	99.37%	98.99%	99.08%	99.32%
Kappa Statistic	0.9864	0.9871	0.9858	0.9798	0.9807	0.9872
TP Rate	99.24%	99.29%	99.20%	99.26%	99.33%	99.41%
FP Rate	0.75%	0.70%	0.75%	1.20%	1.15%	0.75%
PPV	98.83%	98.87%	98.78%	97.91%	98.00%	98.82%
NPV	99.67%	99.71%	99.67%	99.69%	99.70%	99.73%
AUC	99.30%	99.40%	99.30%	98.75%	98.80%	99.15%
C4.5						
Accuracy	99.18%	99.33%	99.36%	98.99%	98.46%	99.09%
Kappa Statistic	0.9801	0.9837	0.9844	0.9759	0.9636	0.9786
TP Rate	99.19%	99.29%	99.20%	99.91%	99.42%	99.40%
FP Rate	1.15%	0.95%	0.85%	1.40%	1.30%	1.35%
PPV	97.98%	98.41%	98.59%	97.64%	95.56%	97.64%
NPV	99.67%	99.71%	99.67%	99.57%	99.75%	99.73%
AUC	99.30%	99.20%	99.30%	98.90%	98.50%	99.10%
CART						
Accuracy	94.36%	94.36%	94.36%	94.93%	94.80%	94.97%
Kappa Statistic	0.9344	0.9344	0.9344	0.9349	0.9276	0.9345
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	7.36%	7.36%	7.36%	6.57%	6.87%	6.61%

ตารางที่ 10 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
PPV	80.86%	80.86%	80.86%	81.86%	82.47%	82.67%
NPV	100%	100%	100%	100%	100%	100%
AUC	99.40%	99.40%	99.40%	99.20%	99.30%	99.50%

จากตารางที่ 10 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 10 Folds Cross-validation และ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตอักเสบเฉียบพลัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตเฉียบพลัน เมื่อใช้กับการแบ่งข้อมูลแบบ 15 Folds Cross-validation

และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตเฉียบพลัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม ID3 ที่ใช้การแบ่งข้อมูลแบบ 10 Folds Cross-validation ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตอักเสบเฉียบพลัน โดยให้ค่า Accuracy เท่ากับ 99.37% ค่า TP Rate เท่ากับ 99.29% ค่า FP Rate เท่ากับ 0.70% ค่า PPV เท่ากับ 98.87% ค่า NVP เท่ากับ 99.71% และค่า AUC เท่ากับ 99.40%

ตารางที่ 11 เปรียบเทียบประสิทธิภาพของอัลกอริทึม กรณีไม่มีการเลือกตัวแปร สำหรับจำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูล

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
ID3						
Accuracy	99.19%	99.26%	99.37%	99.10%	99.31%	99.18%
Kappa Statistic	0.9803	0.9811	0.9822	0.9752	0.9818	0.9784
TP Rate	98.39%	98.39%	98.57%	98.08%	99.60%	98.06%
FP Rate	1.90%	0.90%	0.85%	1.25%	0.85%	1.00%
PPV	98.57%	98.69%	98.69%	98.08%	98.53%	98.63%
NPV	99.52%	99.52%	99.57%	99.43%	99.55%	99.41%
AUC	99.05%	99.15%	99.15%	98.85%	99.25%	99.00%
C4.5						
Accuracy	99.09%	99.22%	99.29%	98.72%	98.87%	99.27%
Kappa Statistic	0.9744	0.9779	0.9798	0.9635	0.9682	0.9758
TP Rate	98.49%	98.98%	99.16%	97.47%	98.36%	98.63%
FP Rate	1.40%	1.35%	1.25%	1.90%	1.85%	1.10%
PPV	97.56%	97.62%	97.74%	96.89%	96.76%	99.24%
NPV	99.56%	99.70%	99.75%	99.26%	99.51%	99.58%
AUC	99.50%	99.50%	99.50%	98.60%	98.80%	99.30%
CART						
Accuracy	97.39%	97.39%	97.39%	97.65%	97.85%	97.91%
Kappa Statistic	0.9252	0.9252	0.9253	0.9328	0.9388	0.9406
TP Rate	95.60%	95.60%	95.60%	83.56%	88.08%	91.29%
FP Rate	4.57%	4.57%	4.57%	6.02%	7.29%	5.18%

ตารางที่ 11 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
PPV	84.30%	84.30%	84.30%	79.57%	74.21%	82.85%
NPV	97.63%	97.63%	97.63%	95.32%	97.03%	94.54%
AUC	99.40%	99.40%	99.40%	99.50%	99.60%	99.60%

จากตารางที่ 11 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 15 Folds Cross-validation มีประสิทธิภาพสูงสุด ในการจำแนกผู้ป่วยโรคปอดอักเสบ เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อใช้กับการแบ่งข้อมูลแบบ 15 Folds Cross-validation

และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 5 Folds, 10 Folds และ 15 Folds Cross-validation ได้ค่าวัดประสิทธิภาพเท่ากัน และมีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 15 Folds Cross-validation ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคปอดอักเสบ โดยให้ค่า Accuracy เท่ากับ 99.37% ค่า TP Rate เท่ากับ 98.57% ค่า FP Rate เท่ากับ 0.85% ค่า PPV เท่ากับ 98.69% ค่า NVP เท่ากับ 99.57% และค่า AUC เท่ากับ 99.15%

2.2 กรณีมีการคัดเลือกตัวแปร

ผลจากการศึกษาในการทดสอบกรณีคัดเลือกตัวแปร ด้วยอัลกอริทึม 3 วิธี คือ Best First, Greedy Stepwise และ Genetic Search จากข้อมูลที่จัดเตรียมไว้ 2 กรณี คือ กรณี Use full training set และกรณี 10 Folds Cross-validation ได้ผลการคัดเลือกตัวแปรจำแนกตามโรค แสดงในตารางที่ 12 – 14 ดังนี้

ตารางที่ 12 การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี

Algorithms	Use full training set		10 Folds Cross-validation	
Best First	1. อาการไข้	2. อาการปวดศีรษะ	1. อาการไข้	2. อาการปวดศีรษะ
	3. อาการเวียนศีรษะ	4. อาการเจ็บคอ	3. อาการเวียนศีรษะ	4. อาการเจ็บคอ
	5. อาการอาเจียน	6. อาการหายใจเหนื่อยหอบ	5. อาการอาเจียน	6. อาการหายใจเหนื่อยหอบ
	7. หายใจมีกลิ่นเหม็น	8. อาการปวดจมูก	7. หายใจมีกลิ่นเหม็น	8. อาการปวดจมูก
	9. อาการปวดหู		9. อาการปวดหู	
Greedy Stepwise	1. อาการไข้	2. อาการปวดศีรษะ	1. อาการไข้	2. อาการปวดศีรษะ
	3. อาการเวียนศีรษะ	4. อาการเจ็บคอ	3. อาการเวียนศีรษะ	4. อาการเจ็บคอ
	5. อาการอาเจียน	6. อาการหายใจเหนื่อยหอบ	5. อาการอาเจียน	6. อาการหายใจเหนื่อยหอบ
	7. หายใจมีกลิ่นเหม็น	8. อาการปวดจมูก	7. หายใจมีกลิ่นเหม็น	8. อาการปวดจมูก
	9. อาการปวดหู		9. อาการปวดหู	
Genetic Search	1. อาการไข้	2. อาการปวดศีรษะ	1. อาการไข้	2. อาการเวียนศีรษะ
	3. อาการเวียนศีรษะ	4. อาการเจ็บคอ	3. อาการเจ็บคอ	4. อาการอาเจียน
	5. อาการอาเจียน	6. อาการหายใจเหนื่อยหอบ	5. อาการหายใจเหนื่อยหอบ	6. อาการปวดจมูก
	7. อาการหายใจมีเสียงหวีดๆ	8. อาการหายใจมีกลิ่นเหม็น	7. อาการปวดหู	
	9. อาการปวดจมูก	10. อาการปวดหู		

จากตารางที่ 12 พบว่าจำนวนตัวแปรที่คัดเลือกได้ มีจำนวนใกล้เคียงกัน โดยวิธี Genetic Search กับการแบ่งชุดข้อมูลเพื่อเรียนรู้แบบ 10 Folds Cross-validation ได้จำนวนตัวแปรน้อยที่สุดเพียง 7 ตัวแปร ได้แก่ อาการไข้ อาการปวดศีรษะ อาการเวียนศีรษะ อาการเจ็บคอ อาการอาเจียน อาการหายใจเหนื่อยหอบ/หายใจลำบาก อาการปวดจุก และอาการปวดหู

ตารางที่ 13 การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคไซนัสอักเสบเฉียบพลัน จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี

Algorithms	Use full training set		10 Folds Cross-validation	
Best First	1. อาการไข้	2. อาการเวียนศีรษะ	1.อาการไข้	2. อาการเวียนศีรษะ
	3. อาการคัดจมูก	4. อาการหายใจมีเสียงหวีดๆ	3. อาการหายใจมีเสียงหวีดๆ	4. อาการหายใจมีกลิ่นเหม็น
	5. อาการหายใจมีกลิ่นเหม็น	6. อาการปวดกระบอกตา	5. อาการปวดกระบอกตา	6. อาการปวดจุก
	7. อาการปวดจุก	8. อาการปวดหู	7. อาการปวดหู	
Greedy Stepwise	1. อาการไข้	2. อาการเวียนศีรษะ	1.อาการไข้	2. อาการเวียนศีรษะ
	3. อาการคัดจมูก	4. อาการหายใจมีเสียงหวีดๆ	3. อาการหายใจมีเสียงหวีดๆ	4. อาการหายใจมีกลิ่นเหม็น
	5. อาการหายใจมีกลิ่นเหม็น	6. อาการปวดกระบอกตา	5. อาการปวดกระบอกตา	6. อาการปวดจุก
	7. อาการปวดจุก	8. อาการปวดหู	7. อาการปวดหู	
Genetic Search	1. อาการไข้	2. อาการเวียนศีรษะ	1.อาการไข้	2. อาการเวียนศีรษะ
	3. อาการคัดจมูก	4. อาการหายใจมีเสียงหวีดๆ	3. อาการหายใจมีเสียงหวีดๆ	4. อาการหายใจมีกลิ่นเหม็น
	5. อาการหายใจมีกลิ่นเหม็น	6. อาการปวดกระบอกตา	5. อาการปวดกระบอกตา	6. อาการปวดจุก
	7. อาการปวดจุก	8. อาการปวดหู	7. อาการปวดหู	

จากตารางที่ 13 พบว่าจำนวนตัวแปรที่คัดเลือกได้มีจำนวนใกล้เคียงกัน ซึ่งอัลกอริทึมการคัดเลือกตัวแปรทั้ง 3 วิธี เมื่อใช้กับการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบ 10 Folds Cross-validation ได้จำนวนตัวแปรน้อยที่สุดเพียง 7 ตัวแปร ได้แก่ อาการไข้ อาการเวียนศีรษะ อาการหายใจมีเสียงหวีดๆ อาการหายใจมีกลิ่นเหม็น อาการปวดกระบอกตา อาการปวดจุก และอาการปวดหู

ตารางที่ 14 การคัดเลือกตัวแปรด้วยอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search สำหรับโรคปอดอักเสบ จำแนกตามการแบ่งข้อมูลที่ใช้ 2 กรณี

Algorithms	Use full training set		10 Folds Cross-validation	
Best First	1. อาการเจ็บคอ	2. มีน้ำมูก	1. อาการเจ็บคอ	2. อาการอาเจียน
	3. อาการอาเจียน	4. อาการหายใจเหนื่อยหอบ	3. อาการหายใจเหนื่อยหอบ	4. อาการหายใจมีเสียงหวีดๆ
	5. อาการหายใจมีเสียงหวีดๆ	6. อาการแน่นหน้าอก	5. อาการแน่นหน้าอก	6. อาการเจ็บหน้าอก
	7. อาการเจ็บหน้าอก	8. อาการปวดจมูก	7. อาการปวดจมูก	8. อาการปวดหู
	9. อาการปวดหู			
Greedy Stepwise	1. อาการเจ็บคอ	2. มีน้ำมูก	1. อาการเจ็บคอ	2. อาการอาเจียน
	3. อาการอาเจียน	4. อาการหายใจเหนื่อยหอบ	3. อาการหายใจเหนื่อยหอบ	4. อาการหายใจมีเสียงหวีดๆ
	5. อาการหายใจมีเสียงหวีดๆ	6. อาการแน่นหน้าอก	5. อาการแน่นหน้าอก	6. อาการเจ็บหน้าอก
	7. อาการเจ็บหน้าอก	8. อาการปวดจมูก	7. อาการปวดจมูก	8. อาการปวดหู
	9. อาการปวดหู			
Genetic Search	1. อาการเจ็บคอ	2. มีน้ำมูก	1. อาการเจ็บคอ	2. อาการอาเจียน
	3. อาการอาเจียน	4. อาการหายใจเหนื่อยหอบ	3. อาการหายใจเหนื่อยหอบ	4. อาการหายใจมีเสียงหวีดๆ
	5. อาการหายใจมีเสียงหวีดๆ	6. อาการแน่นหน้าอก	5. อาการแน่นหน้าอก	6. อาการเจ็บหน้าอก
	7. อาการเจ็บหน้าอก	8. อาการปวดจมูก	7. อาการปวดจมูก	8. อาการปวดหู
	9. อาการปวดหู			

จากตารางที่ 14 พบว่าจำนวนตัวแปรที่คัดเลือกได้มีจำนวนใกล้เคียงกัน ซึ่งอัลกอริทึมการคัดเลือกตัวแปรทั้ง 3 วิธี เมื่อใช้กับการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบ 10 Folds Cross-validation ได้จำนวนตัวแปรน้อยที่สุดเพียง 8 ตัวแปร ได้แก่ อาการเจ็บคอ อาการอาเจียน อาการหายใจ

เหนื่อยหอบ/หายใจลำบาก อาการหายใจมีเสียงหวีดๆ อาการแน่นหน้าอก อาการเจ็บหน้าอก อาการปวดจุก และอาการปวดหู

ตารางที่ 15 สรุปจำนวนตัวแปรที่คัดเลือกได้จากอัลกอริทึม Best First, Greedy Stepwise และ Genetic Search จำแนกตามโรค และการแบ่งชุดข้อมูล

โรคระบบทางเดินหายใจ	Best First	Greedy Stepwise	Genetic Search
โรคการติดเชื้อทางเดินหายใจ			
ส่วนบนแบบเฉียบพลัน			
Use full training set	9	9	10
10 Folds Cross-validation	9	9	7
โรคไซนัสอักเสบเฉียบพลัน			
Use full training set	8	8	8
10 Folds Cross-validation	7	7	7
โรคปอดอักเสบ			
Use full training set	9	9	9
10 Folds Cross-validation	8	8	8

ผลการคัดเลือกตัวแปร พบว่าในแต่ละโรคได้จำนวนตัวแปรที่ใกล้เคียงกัน ขึ้นต่อไปจะเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ โดยใช้ตัวแปรที่คัดเลือกได้จากแต่ละโรค ดังนี้

โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน ใช้ 7 ตัวแปร คือ อาการไข้ อาการปวดศีรษะ อาการเวียนศีรษะ อาการเจ็บคอ อาการอาเจียน อาการหายใจเหนื่อยหอบ/หายใจลำบาก อาการปวดจุก และอาการปวดหู

โรคไซนัสอักเสบเฉียบพลัน ใช้ 7 ตัวแปร คือ อาการไข้ อาการเวียนศีรษะ อาการหายใจมีเสียงหวีดๆ อาการหายใจมีกลิ่นเหม็น อาการปวดกระบอกตา อาการปวดจุก และอาการปวดหู

และสำหรับโรคปอดอักเสบ ใช้ 8 ตัวแปร คือ อาการเจ็บคอ อาการอาเจียน อาการหายใจเหนื่อยหอบ/หายใจลำบาก อาการหายใจมีเสียงหวีดๆ อาการแน่นหน้าอก อาการเจ็บหน้าอก อาการปวดจมูก และอาการปวดหู

ตัวแปรที่ผ่านการคัดเลือก ถูกนำมาพัฒนาตัวจำแนกผู้ป่วยรายโรคโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ ID3, C4.5 และ CART กับชุดข้อมูลเรียนรู้ที่มาจากการจัดเตรียม 6 แบบ คือ 5 Folds Cross-validation, 10 Folds Cross-validation, 15 Folds Cross-validation, 50% split, 60% split และ 70% split และทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมทั้ง 3 ดังตารางที่ 16 - 18

ตารางที่ 16 เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่ใช้การคัดเลือกตัวแปรสำหรับการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ

Algorithms	5 Folds Cross- validation	10 Folds Cross- validation	15 Folds Cross- validation	50% split	60% split	70% split
ID3						
Accuracy	91.51%	91.51%	91.48%	91.24%	91.64%	92.31%
Kappa Statistic	0.8300	0.8300	0.8295	0.8242	0.8320	0.8455
TP Rate	90.75%	90.75%	90.75%	90.44%	90.24%	90.60%
FP Rate	8.50%	8.50%	8.50%	8.80%	8.35%	7.60%
PPV	91.63%	91.63%	91.63%	91.01%	91.78%	92.85%
NPV	91.40%	91.40%	91.40%	91.44%	91.52%	91.86%
AUC	95.70%	95.60%	95.50%	95.50%	95.70%	96.30%

ตารางที่ 16 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
C4.5						
Accuracy	91.47%	91.51%	91.48%	91.24%	91.50%	92.32%
Kappa Statistic	0.8292	0.8300	0.8295	0.8242	0.8292	0.8455
TP Rate	90.74%	90.73%	90.68%	90.44%	90.27%	92.40%
FP Rate	8.55%	8.45%	8.50%	8.80%	8.50%	7.60%
PPV	91.55%	91.66%	91.60%	94.72%	95.04%	95.63%
NPV	94.21%	94.18%	94.13%	90.10%	90.39%	90.18%
AUC	95.50%	95.40%	95.30%	95.40%	95.50%	96.20%
CART						
Accuracy	90.20%	90.20%	90.20%	88.35%	89.10%	90.67%
Kappa Statistic	0.8292	0.8292	0.8292	0.8242	0.8292	0.8427
TP Rate	88.81%	88.81%	88.81%	86.89%	87.15%	89.68%
FP Rate	8.45%	8.45%	8.45%	10.24%	8.94%	8.47%
PPV	91.03%	91.03%	91.03%	89.19%	90.77%	90.29%
NPV	89.43%	89.43%	89.43%	87.57%	87.53%	90.99%
AUC	95.80%	95.80%	95.80%	95.70%	95.80%	96.20%

จากตารางที่ 16 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อใช้กับการแบ่งข้อมูลแบบ 70% split

และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม C4.5 ที่ใช้การแบ่งข้อมูลแบบ 70% split ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โดยให้ค่า Accuracy เท่ากับ 92.32% ค่า TP Rate เท่ากับ 92.40% ค่า FP Rate เท่ากับ 7.60% ค่า PPV เท่ากับ 95.63% ค่า NPV เท่ากับ 90.18% และค่า AUC เท่ากับ 96.20%

ตารางที่ 17 เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่ใช้การเลือกตัวแปรสำหรับการจำแนกผู้ป่วยโรคไขสันหลังอักเสบเฉียบพลัน ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ

Algorithms	5 Folds Cross- validation	10 Folds Cross- validation	15 Folds Cross- validation	50% split	60% split	70% split
ID3						
Accuracy	94.36%	94.36%	94.36%	93.70%	93.62%	93.95%
Kappa	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
Statistic						
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	9.70%	9.70%	9.70%	10.45%	10.40%	9.80%
PPV	80.63%	80.63%	80.63%	79.06%	79.24%	80.41%
NPV	100%	100%	100%	100%	100%	100%
AUC	96.10%	96.10%	96.10%	95.60%	95.80%	96.00%

ตารางที่ 17 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
C4.5						
Accuracy	94.36%	94.36%	94.36%	93.70%	93.62%	93.95%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	9.70%	9.70%	9.70%	10.45%	10.40%	9.80%
PPV	80.63%	80.63%	80.63%	79.06%	79.25%	80.41%
NPV	100%	100%	100%	100%	100%	100%
AUC	89.70%	89.70%	89.70%	89.50%	89.60%	90.20%
CART						
Accuracy	94.36%	94.36%	94.36%	94.69%	93.92%	94.18%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	7.36%	7.36%	7.36%	7.08%	7.99%	7.47%
PPV	80.63%	80.63%	80.63%	81.57%	79.78%	79.11%
NPV	100%	100%	100%	100%	100%	100%
AUC	95.00%	95.00%	95.00%	94.70%	94.90%	95.30%

จากตารางที่ 17 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 5 Folds, 10 Folds และ 15 Folds Cross-validation ได้ค่าวัดประสิทธิภาพเท่ากัน และมีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตเรื้อรังแบบเฉียบพลัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตเรื้อรังแบบเฉียบพลัน เมื่อใช้กับการแบ่งข้อมูลแบบ 5 Folds, 10 Folds และ 15 Folds Cross-validation

และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 50% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 50% split ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคไตเรื้อรังอักเสบเฉียบพลัน โดยให้ค่า Accuracy เท่ากับ 94.69% ค่า TP Rate เท่ากับ 100% ค่า FP Rate เท่ากับ 7.08% ค่า PPV เท่ากับ 81.57% ค่า NPV เท่ากับ 100% และค่า AUC เท่ากับ 94.70%

ตารางที่ 18 เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่มีการคัดเลือกตัวแปรสำหรับการจำแนกผู้ป่วยโรคไตเรื้อรังอักเสบ ตามการแบ่งชุดข้อมูลเพื่อการเรียนรู้แบบต่างๆ

Algorithms	5 Folds Cross- validation on	10 Folds Cross- validation	15 Folds Cross- validation	50% split	60% split	70% split
ID3						
Accuracy	94.63%	94.63%	94.63%	94.41%	94.61%	94.68%
Kappa Statistic	0.8412	0.8412	0.8412	0.8362	0.8357	0.8455
TP Rate	93.99%	93.99%	93.99%	91.24%	93.19%	93.23%
FP Rate	9.85%	9.85%	9.85%	9.45%	9.45%	9.30%
PPV	81.87%	81.87%	81.87%	83.49%	83.09%	83.24%
NPV	98.44%	98.44%	98.44%	97.63%	98.12%	98.16%
AUC	96.00%	96.00%	96.00%	96.40%	96.50%	96.50%

ตารางที่ 18 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
C4.5						
Accuracy	94.61%	94.61%	94.61%	94.41%	94.68%	94.70%
Kappa Statistic	0.8403	0.8403	0.8403	0.8362	0.8446	0.8455
TP Rate	93.98%	93.98%	93.98%	91.24%	93.38%	93.42%
FP Rate	9.95%	9.95%	9.95%	9.45%	9.45%	9.30%
PPV	81.75%	81.75%	81.75%	83.49%	82.94%	83.24%
NPV	98.44%	98.44%	98.44%	97.63%	98.22%	98.16%
AUC	90.90%	90.90%	90.90%	91.70%	91.70%	92.20%
CART						
Accuracy	93.69%	93.69%	93.69%	91.90%	91.66%	92.58%
Kappa Statistic	0.8416	0.8416	0.8416	0.8464	0.8467	0.8469
TP Rate	92.07%	92.07%	92.07%	88.90%	87.29%	89.02%
FP Rate	5.90%	5.90%	5.90%	7.35%	7.27%	6.56%
PPV	79.37%	79.37%	79.37%	75.06%	74.60%	76.81%
NPV	97.96%	97.96%	97.96%	97.10%	96.76%	97.21%
AUC	95.90%	95.90%	95.90%	96.50%	96.60%	96.70%

จากตารางที่ 18 อัลกอริทึม ID3 ที่ใช้กับการแบ่งข้อมูลแบบ 70% split มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคปอดอักเสบ เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

ส่วนอัลกอริทึม C4.5 ก็มีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อใช้กับการแบ่งข้อมูลแบบ 70% split

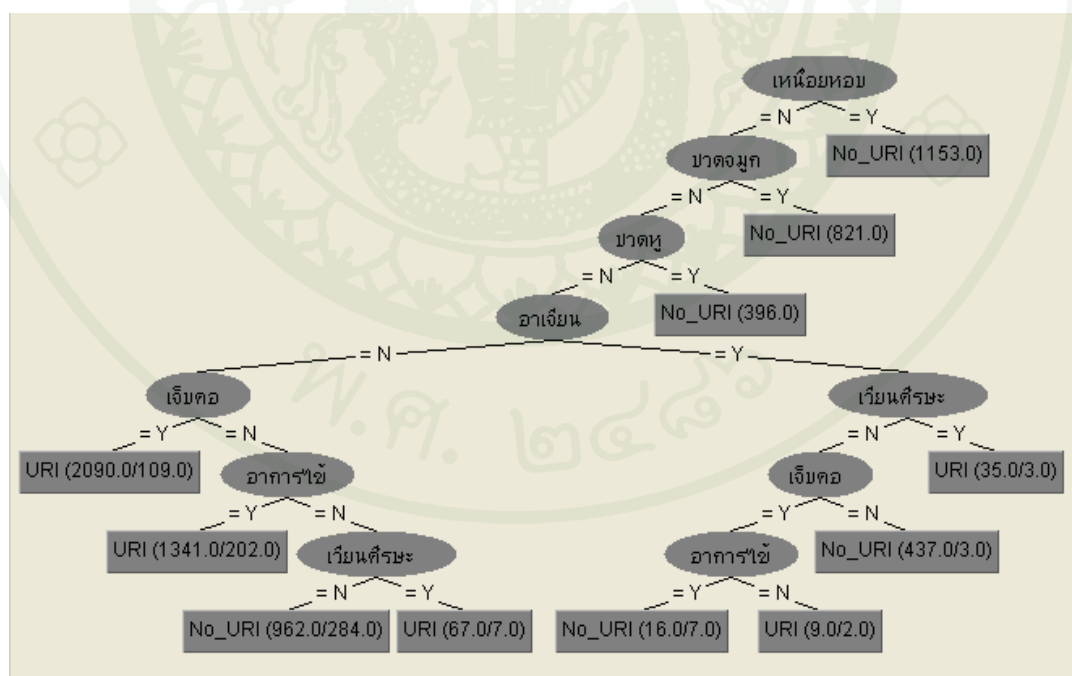
และอัลกอริทึม CART ที่ใช้กับการแบ่งข้อมูลแบบ 5 Folds, 10 Folds และ 15 Folds Cross-validation ได้ค่าวัดประสิทธิภาพเท่ากัน และมีประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคเดียวกัน เมื่อเทียบกับการแบ่งข้อมูลแบบอื่น

เมื่อเปรียบเทียบทั้งสามอัลกอริทึม พบว่าอัลกอริทึม C4.5 ที่ใช้การแบ่งข้อมูลแบบ 70% split ให้ค่าวัดประสิทธิภาพสูงสุดในการจำแนกผู้ป่วยโรคปอดอักเสบ โดยให้ค่า Accuracy เท่ากับ 94.70% ค่า TP Rate เท่ากับ 93.42% ค่า FP Rate เท่ากับ 9.30% ค่า PPV เท่ากับ 83.24% ค่า NVP เท่ากับ 98.16% และค่า AUC เท่ากับ 92.20%

3. ผลที่ได้จากการประยุกต์อัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ป่วย 3 โรค

จากอัลกอริทึมต้นไม้ตัดสินใจที่ให้ค่าประสิทธิภาพสูงสุด ได้แผนภาพต้นไม้ตัดสินใจพร้อมกฎการจำแนกข้อมูลในแต่ละโรคดังนี้

โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน



ภาพที่ 9 ต้นไม้ตัดสินใจสำหรับการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลันด้วยอัลกอริทึม C4.5

สรุปเป็นกฎการจำแนกข้อมูลผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
ได้ดังนี้

1. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (2090.0/109.0)
2. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ แต่มีไข้ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (1341.0/202.0)
3. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ ไม่มีไข้ แต่มีอาการเวียนศีรษะ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (67.0/7.0)
4. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู มีเวียนศีรษะ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (35.0/3.0)
5. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ ไม่มีไข้ และไม่เวียนศีรษะ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (962.0/248.0)
6. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่เวียนศีรษะ แต่มีอาการอาเจียน เจ็บคอ และมีไข้ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (9.0/2.0)
7. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่เวียนศีรษะ และไม่เจ็บคอ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (437.0/30.0)
8. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก และไม่ปวดหู ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (396.0)

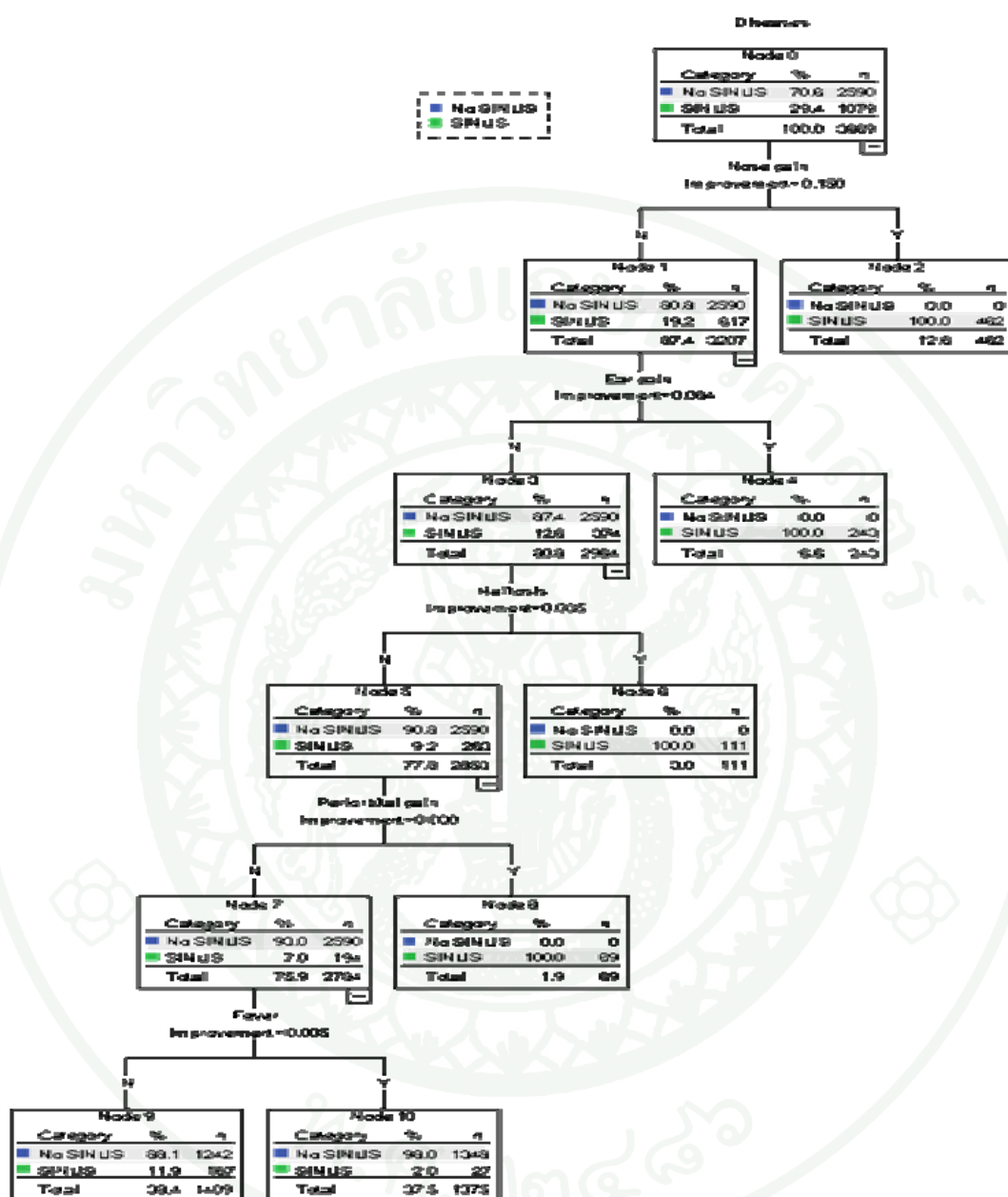
9. ไม่มีอาการหายใจเหนื่อยหอบ และไม่ปวดจุก ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (810.0)

10. มีอาการหายใจเหนื่อยหอบ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (1153.0)

ตัวอย่าง ผู้ที่มีอาการเหนื่อยหอบจะไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (1153.0) หมายถึง ผู้ที่มีอาการเหนื่อยหอบ มีทั้งหมด 1,153 คน และทั้ง 1,153 คน ไม่เป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน หรือผู้ที่ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจุก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีโอกาสเป็นโรคการติดเชื้อทางเดิน

หายใจส่วนบนแบบเฉียบพลัน (2090.0/109.0) หมายถึง มีผู้ที่ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจุก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีทั้งหมด 2,090 คน แต่จะมีผู้ที่เป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลันทั้งสิ้น 109 คน เป็นต้น

โรคไข้น้ำสออักเสบเฉียบพลัน

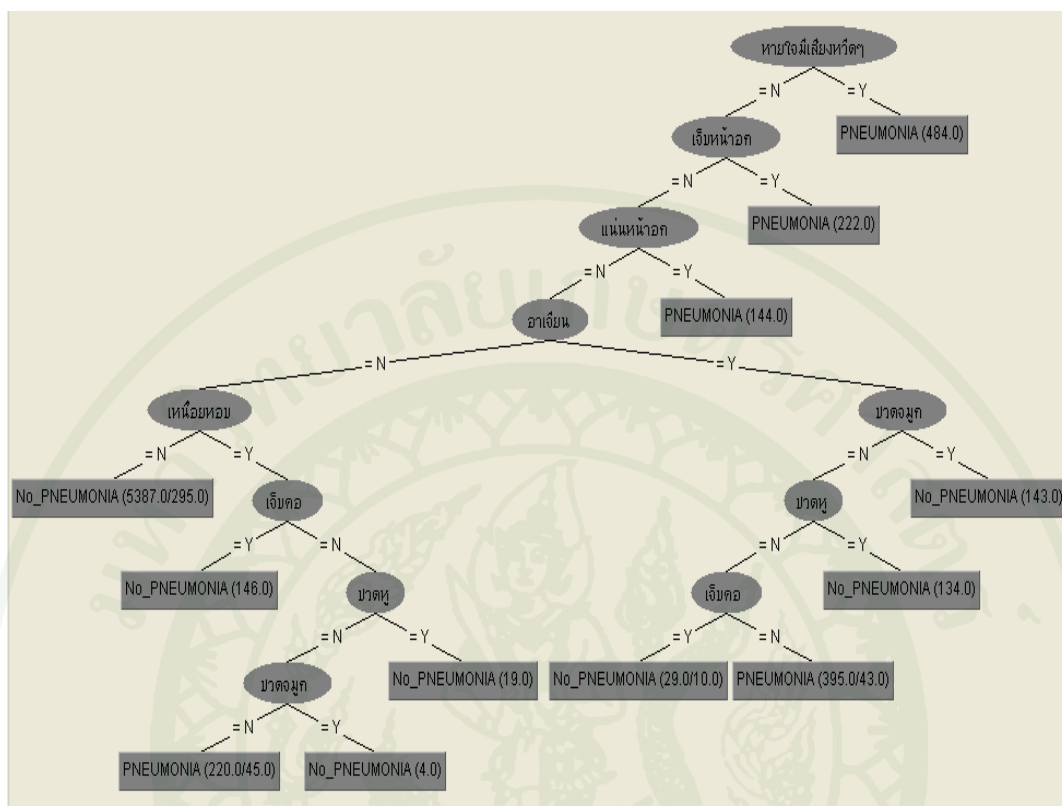


ภาพที่ 10 ต้นไม้การจำแนกผู้ป่วยโรคไข้น้ำสออักเสบเฉียบพลันด้วยอัลกอริทึม CART

สรุปเป็นกฎการจำแนกผู้ป่วยโรคไตเรื้อรังอีกเสบเฉียบพลัน ได้ดังนี้

1. มีอาการปวดจุมก มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (402)
2. ไม่มีอาการปวดจุมก แต่มีอาการปวดหู มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (240)
3. ไม่มีอาการปวดจุมก ไม่ปวดหู แต่มีอาการหายใจมีกลิ่นเหม็น มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (111.0)
5. ไม่มีอาการปวดจุมก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น แต่มีอาการปวดกระบอกตา มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (89.0)
6. ไม่มีอาการปวดจุมก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น ไม่ปวดกระบอกตา ไม่มีอาการปวดกระบอกตา แต่มีไข้ มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (1375.0/27.0)
7. ไม่มีอาการปวดจุมก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น ไม่ปวดกระบอกตา ไม่มีอาการปวดกระบอกตา และไม่มีไข้ ไม่มีโอกาสเป็นโรคไตเรื้อรังอีกเสบเฉียบพลัน (1409.0/1242.0)

โรคปอดอักเสบ



ภาพที่ 11 ต้นไม้ตัดสินใจสำหรับการจำแนกผู้ป่วยโรคปอดอักเสบด้วยอัลกอริทึม C4.5

สรุปเป็นกฎการจำแนกผู้ป่วยโรคปอดอักเสบ ได้ดังนี้

1. มีอาการหายใจมีเสียงหวีดๆ มีโอกาสเป็นโรคปอดอักเสบ (484.0)
2. ไม่มีอาการหายใจมีเสียงหวีดๆ แต่มีอาการเจ็บหน้าอก มีโอกาสเป็นโรคปอดอักเสบ (222.0)
3. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก แต่มีอาการแน่นหน้าอก มีโอกาสเป็นโรคปอดอักเสบ (144.0)
4. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก แต่มีอาการอาเจียนและปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ (143.0)

5. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก แต่มีอาการอาเจียนและปวดหู ไม่มีโอกาสเป็นโรคปอดอักเสบ (134.0)
6. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก ไม่ปวดหู ไม่เจ็บคอ แต่มีอาการอาเจียน มีโอกาสเป็นโรคปอดอักเสบ (395.0/43.0)
7. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก ไม่ปวดหู ไม่เจ็บคอ แต่มีอาการอาเจียนและเจ็บคอ ไม่มีโอกาสเป็นโรคปอดอักเสบ (29.0/10.0)
8. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน และไม่หายใจเหนื่อยหอบ ไม่มีโอกาสเป็นโรคปอดอักเสบ (5387.0)
9. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน แต่มีอาการหายใจเหนื่อยหอบและเจ็บคอ ไม่มีโอกาสเป็นโรคปอดอักเสบ (146.0)
10. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ แต่มีอาการหายใจเหนื่อยหอบและปวดหู ไม่มีโอกาสเป็นโรคปอดอักเสบ (19.0)
11. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ ไม่ปวดหู แต่มีอาการปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ (4.0)
12. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ ไม่ปวดหู และไม่ปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ (220.0/45.0)

วิจารณ์

1. เมื่อพิจารณาประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจที่ดีที่สุด ที่ใช้ในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ ทั้งกรณีที่ไม่มีการคัดเลือกตัวแปรและมีการคัดเลือกตัวแปร จะเห็นว่าอัลกอริทึม C4.5 ที่ใช้กับการแบ่งข้อมูลเพื่อการเรียนรู้แบบ 70% split ในการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ เมื่อใช้ทั้งหมด 22 ตัวแปร จะให้ประสิทธิภาพการจำแนกสูงกว่าเมื่อมีการใช้เพียง 7 ตัวแปร แต่อย่างไรก็ตาม การใช้เพียง 7 ตัวแปรสำหรับการจำแนกผู้ป่วยแต่ละโรค ได้ประสิทธิภาพสูงถึงร้อยละ 92 และ 95 ตามลำดับ ซึ่งเป็นการประหยัดคำถามในการซักประวัติผู้ป่วย สำหรับการคัดกรองผู้ป่วย ทำนองเดียวกัน การจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน ใช้เพียง 8 ตัวแปรที่คัดเลือกได้ ก็สามารถใช้อัลกอริทึม CART ในการช่วยจำแนกผู้ป่วยได้ถูกต้องถึงร้อยละ 95 อย่างไรก็ตาม ผลการศึกษามาจากกรณีศึกษาเฉพาะข้อมูลโรงพยาบาลพระนครศรีอยุธยา ช่วงปี พ.ศ. 2547 ถึง พ.ศ. 2549 เท่านั้น

2. กลุ่มโรคที่ทำการศึกษา ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบ มีตัวแปรที่ผ่านการคัดเลือกและมีอิทธิพลต่อการจำแนกโรคแต่ละโรคได้อย่างชัดเจน เช่น ตัวแปรอาการหายใจมีกลิ่นเหม็น จะพบเฉพาะในผู้ป่วยที่เป็นโรคไซนัสอักเสบเฉียบพลันเท่านั้น จึงทำให้ได้ประสิทธิภาพของการจำแนกมีค่าสูง

3. ลักษณะของข้อมูลสำหรับอัลกอริทึม ID3 ต้องเป็นข้อมูลเชิงกลุ่มและต้องไม่มีค่าของข้อมูลสูญหายเท่านั้น หากจำเป็นต้องใช้ข้อมูลเชิงปริมาณ ต้องทำการจัดกลุ่มข้อมูลก่อน หากไม่ทำการจัดกลุ่มข้อมูลจะเกิดปัญหาความอคติของข้อมูล มีผลต่อการวิเคราะห์การจำแนก สำหรับลักษณะของข้อมูลที่น่ามาใช้กับอัลกอริทึม C4.5 และ CART สามารถใช้ได้ทั้งข้อมูลเชิงกลุ่มและข้อมูลเชิงปริมาณ โดยอัลกอริทึม C4.5 สามารถใช้ข้อมูลที่มีค่าสูญหายได้

4. การแบ่งชุดข้อมูลให้ระบบเรียนรู้ทั้ง 2 วิธี ได้แก่ วิธี Cross-validation และวิธี Percentage split พบว่าการแบ่งข้อมูลแบบ 15 Folds และ 70% split จะทำให้ระบบมีข้อมูลฝึกสอนสำหรับการเรียนรู้มากขึ้น มีผลต่อการสร้างตัวจำแนกผู้ป่วยใน 3 โรคที่มีประสิทธิภาพสูง

5. การศึกษาด้วยอัลกอริทึมต้นไม้ตัดสินใจ ไม่เหมาะสมสำหรับโรคที่พบได้น้อย (Rare diseases) เนื่องจากอัลกอริทึมต้นไม้ตัดสินใจเป็นกระบวนการเรียนรู้ที่ใช้ค้นหาารูปแบบ

ความสัมพันธ์จากฐานข้อมูลขนาดใหญ่ จึงเหมาะสมกับการศึกษาในกลุ่มโรคที่มีค่าความชุก (Prevalence) สูง ซึ่งโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไฉนส์อักเสบเฉียบพลัน และโรคปอดอักเสบ เป็นโรคที่มีค่าความชุกสูงเป็นอันดับ 1-3 ของโรงพยาบาลพระนครศรีอยุธยา

6. ผลการศึกษาจากกรณีที่ไม่มีการคัดเลือกตัวแปร แม้ว่าค่าประสิทธิภาพที่ได้จะให้ค่าสูงกว่ากรณีคัดเลือกตัวแปรแต่ก็ให้ค่าสูงกว่าเพียงเล็กน้อยเท่านั้น แต่เมื่อเทียบกับจำนวนตัวแปรที่ใช้ในการจำแนกแต่ละโรค จะพบว่ากรณีการคัดเลือกตัวแปรใช้เพียง 7 ตัวแปรในการคัดกรองผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ และใช้ 8 ตัวแปรในการคัดกรองผู้ป่วยโรคไฉนส์อักเสบเฉียบพลัน การศึกษานี้จึงเสนอให้ใช้กรณีที่มีการคัดเลือกตัวแปร

สรุปและข้อเสนอแนะ

สรุป

การศึกษาการประยุกต์อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรกระบบทางเดินหายใจ มีวัตถุประสงค์ของการศึกษาเพื่อประยุกต์การใช้อัลกอริทึม 3 อัลกอริทึม คือ ID3, C4.5 และ CART กับ การวินิจฉัยโรกระบบทางเดินหายใจ และเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจข้างต้น ในการจำแนกผู้ป่วยกลุ่มโรกระบบทางเดินหายใจ ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน โรคไซนัสอักเสบเฉียบพลัน และโรคปอดอักเสบของโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา ทำการศึกษาเฉพาะผู้ที่เข้ามารับการรักษาในครั้งแรก ตั้งแต่ปี พ.ศ. 2547 - 2549 จำนวน 7,327 คน โดยตรวจสอบความถูกต้องของข้อมูล กำหนดรหัสข้อมูลตัวแปร ซึ่งตัวแปรที่นำมาพิจารณา 22 ตัวแปร ประกอบด้วย อายุ อุณหภูมิร่างกาย เขตที่อยู่อาศัย อาชีพ อาการต่างๆ เช่น มีน้ำมูก มีไข้ คัดจมูก แน่นหน้าอก ปวดกระบอกตา ปวดศีรษะ หายใจเหนื่อยหอบ และไอ เป็นต้น การศึกษานี้จึงพิจารณาใช้ตัวแปรทั้งหมดกับใช้ตัวแปรบางส่วน โดยผ่านอัลกอริทึมการคัดเลือกตัวแปร 3 วิธี ได้แก่ 1) วิธี Best First 2) วิธี Greedy Stepwise และ 3) วิธี Genetic Search แล้วนำตัวแปรที่ผ่านการคัดเลือกมาสร้างตัวจำแนก โดยแบ่งข้อมูลเพื่อการเรียนรู้ 6 แบบ ได้แก่ 50% split 60% split 70% split 5-Folds Cross-validation 10-Folds Cross-validation และ 15-Folds Cross-validation ทำการเปรียบเทียบประสิทธิภาพของตัวจำแนกที่ได้จากแต่ละอัลกอริทึมต้นไม้ตัดสินใจ ด้วยค่าวัดประสิทธิภาพต่อไปนี้ ได้แก่ ค่า Accuracy Kappa Statistic TP Rate FP Rate PPV NPV และ AUC สรุปผลการศึกษาได้ดังนี้

1. การประยุกต์การใช้อัลกอริทึมต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบทางเดินหายใจกับการวินิจฉัยโรค

จากการประยุกต์อัลกอริทึมต้นไม้ตัดสินใจ ทำให้ได้แนวทางสำหรับสนับสนุนระบบการคัดกรองผู้ป่วยโรคระบบทางเดินหายใจในเบื้องต้นเพื่อการวินิจฉัยต่อไป ดังภาพต้นไม้การจำแนกผู้ป่วยทั้ง 3 โรค (ภาพที่ 9-11) และได้กฎการจำแนกผู้ป่วยในแต่ละโรคดังนี้

โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน ได้กฎการจำแนกผู้ป่วยดังนี้

1. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
2. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ แต่มีไข้ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
3. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ ไม่มีไข้ แต่มีอาการเวียนศีรษะ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
4. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู มีเวียนศีรษะ มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
5. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ ไม่มีไข้ และไม่เวียนศีรษะ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
6. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่เวียนศีรษะ แต่มีอาการอาเจียน เจ็บคอ และมีไข้ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน
7. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่เวียนศีรษะ และไม่เจ็บคอ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน

8. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก และไม่ปวดหู ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน

9. ไม่มีอาการหายใจเหนื่อยหอบ และไม่ปวดจมูก ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน

10. มีอาการหายใจเหนื่อยหอบ ไม่มีโอกาสเป็นโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน

โรคไซนัสอักเสบเฉียบพลัน ได้ถูกการจำแนกผู้ป่วยดังนี้

1. มีอาการปวดจมูก มีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน
2. ไม่มีอาการปวดจมูก แต่มีอาการปวดหู มีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน
3. ไม่มีอาการปวดจมูก ไม่ปวดหู แต่มีอาการหายใจมีกลิ่นเหม็น มีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน
5. ไม่มีอาการปวดจมูก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น แต่มีอาการปวดกระบอกตามีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน
6. ไม่มีอาการปวดจมูก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น ไม่ปวดกระบอกตา ไม่มีอาการปวดกระบอกตา แต่มีไข้ มีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน
7. ไม่มีอาการปวดจมูก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น ไม่ปวดกระบอกตา ไม่มีอาการปวดกระบอกตา และไม่มีไข้ ไม่มีโอกาสเป็นโรคไซนัสอักเสบเฉียบพลัน

โรคปอดอักเสบ ได้ถูกการจำแนกผู้ป่วยดังนี้

1. มีอาการหายใจมีเสียงหวีดๆ มีโอกาสเป็นโรคปอดอักเสบ

2. ไม่มีอาการหายใจมีเสียงหวีดๆ แต่มีอาการเจ็บหน้าอก มีโอกาสเป็นโรคปอดอักเสบ
3. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก แต่มีอาการแน่นหน้าอก มีโอกาสเป็นโรคปอดอักเสบ
4. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก แต่มีอาการอาเจียนและปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ
5. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก แต่มีอาการอาเจียนและปวดหู ไม่มีโอกาสเป็นโรคปอดอักเสบ
6. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก ไม่ปวดหู ไม่เจ็บคอ แต่มีอาการอาเจียน มีโอกาสเป็นโรคปอดอักเสบ
7. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจุก ไม่ปวดหู ไม่เจ็บคอ แต่มีอาการอาเจียนและเจ็บคอ ไม่มีโอกาสเป็นโรคปอดอักเสบ
8. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน และไม่หายใจเหนื่อยหอบ ไม่มีโอกาสเป็นโรคปอดอักเสบ
9. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน แต่มีอาการหายใจเหนื่อยหอบและเจ็บคอ ไม่มีโอกาสเป็นโรคปอดอักเสบ
10. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ แต่มีอาการหายใจเหนื่อยหอบและปวดหู ไม่มีโอกาสเป็นโรคปอดอักเสบ
11. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ ไม่ปวดหู แต่มีอาการปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ
12. ไม่มีอาการหายใจมีเสียงหวีดๆ ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่อาเจียน ไม่เจ็บคอ ไม่ปวดหู และไม่ปวดจุก ไม่มีโอกาสเป็นโรคปอดอักเสบ

2. เปรียบเทียบประสิทธิภาพของอัลกอริทึมในการจำแนกผู้ป่วยกลุ่มโรคระบบทางเดินหายใจ

ตารางที่ 19 เปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม่ตัดสินใจ กรณีไม่มีการคัดเลือกตัวแปรและกรณีมีการคัดเลือกตัวแปร ในการจำแนกผู้ป่วยกลุ่มโรคระบบทางเดินหายใจใน 3 โรค

ค่าวัด ประสิทธิภาพ	กรณีไม่คัดเลือกตัวแปร			กรณีคัดเลือกตัวแปร		
	Acute	Acute	Pneumonia	Acute	Acute	Pneumonia
	URI	Sinusitis	ID 3	URI	Sinusitis	C4.5
	C4.5	ID 3	15 Folds	C4.5	CART	70% split
	70% split	10 Folds		70% split	50% split	
Accuracy	99.42%	99.37%	99.37%	92.32%	94.69%	94.70%
Kappa Statistic	0.9881	0.9871	0.9253	0.8455	0.8407	0.8455
TP Rate	99.31%	99.29%	95.57%	92.40%	100%	93.42%
FP Rate	0.60%	0.70%	0.85%	7.60%	7.08%	9.30%
PPV	99.40%	98.87%	98.69%	95.63%	81.57%	83.24%
NPV	99.41%	99.71%	99.57%	90.18%	100%	98.16%
AUC	99.70%	99.40%	99.15%	96.20%	94.70%	92.20%

จากตารางที่ 19 พบว่าทั้งกรณีไม่มีการคัดเลือกตัวแปร และกรณีมีการคัดเลือกตัวแปรได้ค่าวัดประสิทธิภาพสูงกว่าร้อยละ 90 แต่กรณีคัดเลือกตัวแปรใช้จำนวนตัวแปรที่น้อยกว่า จึงควรใช้กรณีคัดเลือกตัวแปร โดยให้ผลการจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ ด้วยอัลกอริทึม C4.5 กับข้อมูลที่มีการแบ่งแบบ 70% split มีประสิทธิภาพสูงสุด ส่วนการจำแนกผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน ด้วยอัลกอริทึม CART กับข้อมูลที่มีการแบ่งแบบ 50% split มีประสิทธิภาพสูงสุด ดังตารางที่ 20

ตารางที่ 20 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการคัดเลือกตัวแปร กับจำนวนตัวแปรในการจำแนกผู้ป่วยกลุ่มโรคระบบทางเดินหายใจใน 3 โรค

ค่าวัดประสิทธิภาพ	โรคระบบทางเดินหายใจ		
	Acute URI	Acute Sinusitis	Pneumonia
	C4.5	CART	C4.5
	70% split	50% split	70% split
Accuracy	92.32%	94.69%	94.70%
Kappa Statistic	0.8455	0.8407	0.8455
TP Rate	92.40%	100%	93.42%
FP Rate	7.60%	7.08%	9.30%
PPV	95.63%	81.57%	83.24%
NPV	90.18%	100%	98.16%
AUC	96.20%	94.70%	92.20%
จำนวนตัวแปร	7	8	7

ข้อเสนอแนะ

1. ผลการศึกษาที่ได้ ควรเสนอให้โรงพยาบาลพระนครศรีอยุธยา ลองนำไปทดลองใช้เป็นแนวทางการคัดกรองหรือจำแนกผู้ป่วย โรคระบบทางเดินหายใจทั้ง 3 โรค ควบคู่กับแนวทางการคัดกรองผู้ป่วยที่ใช้ในปัจจุบัน เพื่อเป็นการยืนยันว่า อัลกอริทึมต้นไม้ตัดสินใจที่ได้ใช้สร้างเป็นกฎหรือเงื่อนไขการจำแนกผู้ป่วยในแต่ละโรคได้ และสามารถนำไปใช้ได้จริงหรือไม่

2. การศึกษานี้ได้ทำการศึกษาเพียง 3 โรค ได้แก่ โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI) โรคไซนัสอักเสบเฉียบพลัน (Acute Sinusitis) และ โรคปอดอักเสบ (Pneumonia) ในการศึกษาต่อไปอาจพิจารณาเพิ่มชนิดของโรคในกลุ่มอื่น หรือเพิ่มโรคที่อยู่ในกลุ่มเดียวกัน โดยอาจทำการศึกษากลุ่มโรคที่มีลักษณะอาการคล้ายคลึงกันซึ่งยากต่อการคัดกรองในเบื้องต้น โรคที่นำมาทำการศึกษาคควรเป็นโรคที่มีค่าความชุกของโรคสูง

3. ผู้ที่สนใจประยุกต์อัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรค อาจจะเพิ่มอัลกอริทึมต้นไม้ตัดสินใจแบบอื่นในการศึกษา เช่น อัลกอริทึม CHAID และ C5.0 เป็นต้น ทำการเปรียบเทียบกับวิธีการเหมือนข้อมูลแบบอื่น เช่น Bayesian Classification, Neural Network และ Genetic Algorithm เป็นต้น แต่การใช้อัลกอริทึมใดต้องศึกษาเงื่อนไขลักษณะข้อมูลจะต้องมีลักษณะเป็นข้อมูลเชิงกลุ่มหรือข้อมูลเชิงปริมาณของแต่ละอัลกอริทึม พร้อมทั้งทำการเปรียบเทียบกับวิธีการทางสถิติอื่น เช่น การวิเคราะห์จำแนกประเภท (Discriminant Analysis) การวิเคราะห์การถดถอยโลจิสติกแบบทวี เป็นต้น ซึ่งนอกจากนี้ควรนำตัวแปรอื่น ที่ไม่ใช่ลักษณะอาการของโรค เข้ามาศึกษาด้วย เช่น ระดับมลพิษในอากาศ ลักษณะทางกรรมพันธุ์ พฤติกรรมในชีวิตประจำวันที่อาจส่งผลต่อการเกิดโรค เป็นต้น เพื่อนำผลจากการศึกษา มาอธิบายปัจจัยการเกิดโรคในปัจจุบัน

4. ผู้ที่ทำหน้าที่วิเคราะห์ข้อมูลทางสถิติ ต้องทราบข้อจำกัดข้อมูลและข้อตกลงเกี่ยวกับสถิติวิเคราะห์ต่างๆ ทำนองเดียวกันหากต้องการประยุกต์วิธีอัลกอริทึมต้นไม้ตัดสินใจ ก็ต้องศึกษาข้อจำกัด ลักษณะข้อมูลที่จะใช้กับอัลกอริทึมต่างๆ เช่นเดียวกัน

เอกสารและสิ่งอ้างอิง

การนิคมอุตสาหกรรมแห่งประเทศไทย. รายงานนิคมอุตสาหกรรมในประเทศไทย. แหล่งที่มา:

http://www.ieat.go.th/index_thtest.php?lang=en&lang=en&CLmIEAT2=12,

27 มกราคม 2550.

ศิริพรรณ วิญญูนันท์กุล. 2542. การปรับปรุงวิธีการสร้างแผนภาพตัดสินใจทวิภาคโดยการเรียนรู้ต้นไม้ตัดสินใจ. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

พรทิพย์ รัตนวิชัย. 2552. แนวทางการคัดกรองโรกระบบทางเดินหายใจเบื้องต้น โรงพยาบาลพระนครศรีอยุธยา. โรงพยาบาลพระนครศรีอยุธยา, พระนครศรีอยุธยา. (อัสสำเนา)

บุษดี ศิริแสงตระกูล. 2551. ระบบจำแนกประเภทแบบทดสอบสำหรับผู้ทดสอบสุขภาพจิตด้วยเทคนิค Decision Tree. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยขอนแก่น.

จันทร์จิรา วิรัช. 2544. ผลของโปรแกรมการฟื้นฟูสมรรถภาพปอดต่ออาการหายใจลำบากและคุณภาพชีวิตของผู้ป่วยโรคปอดอุดกั้นเรื้อรัง. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเชียงใหม่.

วิมเนศ ทองมี. 2545. วิธีการตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยใช้ฟuzzy logic. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

วราภรณ์ พินา. 2551. การวิเคราะห์ปัจจัยเสี่ยงและการจัดจำแนกกลุ่มของการติดเชื้อในผู้ป่วยวัณโรคปอดที่เกิดโรคกลับ. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเชียงใหม่.

สุรเกียรติ์ อาชานานุภาพ. 2544. ตำราการตรวจรักษาโรคทั่วไป (หลักการวินิจฉัยและรักษาโรค / 280 โรคและการดูแลรักษา). ครั้งที่ 3. สำนักพิมพ์หมอชาวบ้าน, กรุงเทพฯ.

สำนักกระบวนวิชา. 2550. สรุปรายงานการเฝ้าระวังโรคประจำปี พ.ศ 2547 – 2548.

สรรพพงศ์ (นามแฝง). 2551. Hospital Health Care. **มติชนรายวัน** 31(11125):5

Anthony, F.S., B. Eugene, K.L. Dennis and L. Joseph. 2008. **Harrison Principles of Internal Medicine**. 17th ed. McGraw-Hill Professional, New York.

Bickley, L. S. and B. Bates. 2002. **Bates' Guide to Physical Examination and History Taking**. 8th ed. Lippincott Williams & Wilkins, New York.

Chan, P.K. and S.J. Stolfo. 1995. Learning arbiter and combiner trees from partitioned data for scaling machine learning. **In Proceedings of the First International Conference on Knowledge Discovery and Data Mining**, 1:39-44

Daniel, L.T. 2004. **Discovering Knowledge in Data: An Introduction to Data Mining**. Wiley-Interscience, Massachusetts.

Fayyad, M., P.S. Usama and U. Ramasamy. 1996. **Advances in Knowledge Discovery and Data Mining**. MIT Press, Massachusetts.

Greenberg, R.S., D.R. Stephen and B.R. John. 2004. **Medical Epidemiology**. 3th ed. McGraw-Hill, New York.

Holly, Y. 2001. **Shared User-Computer Control of a Robotic Wheelchair System**. M.S. Thesis, Massachusetts Institute of Technology.

Jiawei, H. and K. Micheline. 2000. **Data Mining Concepts and Techniques**. Morgan Kaufmann, New York.

Johannes, G.R. and V. Ganti. 1998. Rainforest: A framework for fast decision tree construction of large datasets, pp.416-427. **Proceedings of the 24th International Conference on Very Large Data Bases**. Morgan Kaufmann Publishers Inc, San Francisco.

- Jose, P., M.C. Aleman and A. Esquerda. 2008. A decision tree for differentiating tuberculous from malignant pleural effusions. **Respiratory Medicine** 8(120): 1159-1164.
- Kamila, A. 2007. **Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems**. M.S. Thesis, Swedish University.
- Kardi, T. 2009. **Tutorial on Decision Tree**. Available Source: <http://people.revoledu.com/kardi/tutorial/DecisionTree/what-is-decision-tree.htm>, October 11, 2009.
- Lei, T.P. 1990. **Toward an Intelligent Classification-tree Approach to Problem-solving**. M.S. Thesis, University of Illinois.
- Leo, B., S. J. Charles and F. Jerome. 1984. **Classification and Regression Trees**. University of Californiat, CA.
- Lior, R. and O. Maimon. 2008. **Data Mining with Decision Trees Theory and Applications**. World Scientific Publishing Company, New Jersey.
- Melanie, M. 1998. **An Introduction to Genetic Algorithms**. The MIT Press, Massachusetts.
- Mel, R. 2002. **The science of bad breath**. Scientific American 286: 72-79.
- Michael, C. 2009. **More than Just the Kappa Coefficient: A Program to Fully Characterize Lnter-Rater Reliability between Two Raters**. University of Pittsburgh, PA.

- Peter, C., P. Hadjnian, S. Rolf, V. Jaap and Zanasi. 1997. **Discovering Data Mining: from concept to implementation**. Prentice Hall PTR, New York.
- Ponnuswamy, V. 2001. **Classification of Software Quality with Tree Modeling Using C4.5 Algorithm**. M.S. Thesis, Florida Atlantic University.
- Richard, L.J. and G.G. Koch. 1997. The measurement of observer agreement for categorical data. **Biometrics** 33:159-174
- Ross, Q. J. 1986. Induction of decision trees. **Machine Learning** 1(1):81-106.
- Ross, Q. J. 1992. **C4.5: Programs For Machine Learning**. Morgan Kaufmann, LA.
- Soman, T. and B.O. Patrick. 2005. **Classification of Arrhythmia Using Machine Learning Techniques**. Available Source:
http://cse.spsu.edu/pbobbie/SharedFile/ECGDiagnosis_ICOSSE_2005_VFinal.pdf,
 January 27, 2007.
- Tape, T.G. 2004. **Interpreting Diagnostic Tests**. Available Source:
<http://gim.unmc.edu/dxtests>, August 25, 2007.
- Usama, F.M., P.S. Gregory and S. Padhraic. 1996. **From Data Mining to Knowledge Discovery in Databases**. MIT Press, Menlo Park.
- Vassilios, S. 1992. **Expert System Based Security Assessment and Enhancement**. M.S. Thesis, Case Western Reserve University.
- Witten, H.I. and F. Eibe. 2005. **Data Mining Practical Machine Learning Tools and Techniques**. 2th ed. Morgan Kaufmann, LA.

Yunling, W. 2004. **A comparative study of classification algorithms for network Intrusion detection**. M.S. Thesis, Florida Atiantic University.





กรณีมีการคัดเลือกตัวแปร

ตารางผนวกที่ 1 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการคัดเลือกตัวแปร ด้วยวิธี Genetic Search กับข้อมูลแบบ Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูล ทดสอบ

Algorithms	5 Folds Cross- validation	10 Folds Cross- validation	15 Folds Cross- validation	50% split	60% split	70% split
ID3						
Accuracy	91.51%	91.51%	91.48%	91.24%	91.64%	92.31%
Kappa Statistic	0.8300	0.8300	0.8295	0.8242	0.8320	0.8455
TP Rate	90.75%	90.75%	90.75%	90.44%	90.24%	90.60%
FP Rate	8.50%	8.50%	8.50%	8.80%	8.35%	7.60%
PPV	91.63%	91.63%	91.63%	91.01%	91.78%	92.85%
NPV	91.40%	91.40%	91.40%	91.44%	91.52%	91.86%
AUC	95.70%	95.60%	95.50%	95.50%	95.70%	96.30%
C4.5						
Accuracy	91.47%	91.51%	91.48%	91.24%	91.50%	92.32%
Kappa Statistic	0.8292	0.8300	0.8295	0.8242	0.8292	0.8455
TP Rate	90.74%	90.73%	90.68%	90.44%	90.27%	92.40%
FP Rate	8.55%	8.45%	8.50%	8.80%	8.50%	7.60%
PPV	91.55%	91.66%	91.60%	94.72%	95.04%	95.63%
NPV	94.21%	94.18%	94.13%	90.10%	90.39%	90.18%
AUC	95.50%	95.40%	95.30%	95.40%	95.50%	96.20%

ตารางผนวกที่ 1 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
CART						
Accuracy	90.20%	90.20%	90.20%	88.35%	89.10%	90.67%
Kappa Statistic	0.8292	0.8292	0.8292	0.8242	0.8292	0.8427
TP Rate	88.81%	88.81%	88.81%	86.89%	87.15%	89.68%
FP Rate	8.45%	8.45%	8.45%	10.24%	8.94%	8.47%
PPV	91.03%	91.03%	91.03%	89.19%	90.77%	90.29%
NPV	89.43%	89.43%	89.43%	87.57%	87.53%	90.99%
AUC	95.80%	95.80%	95.80%	95.70%	95.80%	96.20%

ตารางผนวกที่ 2 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรแบบ Best First และ Greedy Stepwise เมื่อใช้การแบ่งข้อมูลแบบ Use full training set หรือ Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
ID3						
Accuracy	92.07%	91.97%	91.81%	91.51%	91.88%	91.91%
Kappa Statistic	0.8411	0.8393	0.8360	0.8304	0.8376	0.8381
TP Rate	91.81%	91.33%	91.11%	88.14%	87.98%	87.78%
FP Rate	7.95%	8.00%	8.20%	8.30%	7.85%	7.80%
PPV	91.63%	92.00%	91.89%	94.73%	95.41%	95.63%
NPV	92.48%	91.95%	91.74%	88.65%	88.68%	88.76%

ตารางผนวกที่ 2 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
AUC	97.00%	96.90%	96.90%	96.70%	97.00%	97.20%
C4.5						
Accuracy	92.04%	92.08%	92.06%	91.51%	91.74%	91.91%
Kappa Statistic	0.8406	0.8414	0.8409	0.8304	0.8848	0.8381
TP Rate	91.83%	91.83%	91.76%	87.78%	87.99%	88.13%
FP Rate	8.00%	7.90%	7.95%	8.30%	8.05%	7.80%
PPV	92.54%	91.66%	91.66%	94.72%	95.04%	95.63%
NPV	92.50%	92.48%	92.42%	88.65%	88.93%	88.76%
AUC	96.10%	96.10%	96.00%	96.20%	96.30%	96.70%
CART						
Accuracy	90.20%	90.20%	90.20%	88.40%	87.66%	88.93%
Kappa Statistic	0.8389	0.8389	0.8389	0.8304	0.8348	0.8371
TP Rate	88.81%	88.81%	88.81%	87.25%	86.49%	86.44%
FP Rate	8.45%	8.45%	8.45%	10.48%	11.28%	8.50%
PPV	91.03%	91.03%	91.03%	89.06%	87.38%	91.29%
NPV	89.43%	89.43%	89.43%	87.78%	87.91%	86.74%
AUC	96.90%	96.90%	96.90%	96.70%	96.90%	97.10%

ตารางผนวกที่ 3 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรด้วยวิธี Genetic Search จาก Use full training set สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
ID3						
Accuracy	92.96%	92.86%	92.70%	92.28%	92.66%	92.68%
Kappa Statistic	0.8588	0.8569	0.8537	0.8455	0.8531	0.8534
TP Rate	93.54%	93.03%	92.81%	89.49%	89.38%	89.17%
FP Rate	7.10%	7.15%	7.35%	4.95%	4.16%	3.93%
PPV	91.63%	92.00%	91.89%	94.72%	95.41%	95.63%
NPV	94.18%	93.65%	93.45%	90.09%	90.32%	90.18%
AUC	97.30%	97.20%	97.20%	97.00%	97.30%	97.50%
C4.5						
Accuracy	92.93%	92.97%	92.94%	92.28%	92.53%	92.68%
Kappa Statistic	0.8582	0.8591	0.8585	0.8455	0.8503	0.8534
TP Rate	93.57%	93.55%	93.50%	89.49%	89.41%	89.17%
FP Rate	7.15%	7.05%	7.10%	7.60%	7.30%	7.10%
PPV	91.55%	91.66%	91.60%	94.72%	95.04%	95.63%
NPV	94.21%	94.18%	94.13%	90.10%	90.39%	90.18%
AUC	96.60%	96.50%	96.40%	96.50%	96.70%	97.00%

ตารางผนวกที่ 3 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
CART						
Accuracy	91.09%	91.09%	91.09%	88.13%	88.65%	90.50%
Kappa Statistic	0.8565	0.8565	0.8565	0.8455	0.8503	0.8524
TP Rate	90.44%	90.44%	90.44%	86.48%	87.89%	84.00%
FP Rate	8.31%	8.31%	8.31%	10.32%	10.62%	9.00%
PPV	91.03%	91.03%	91.03%	88.75%	88.89%	90.97%
NPV	91.14%	91.14%	91.14%	87.57%	88.41%	90.03%
AUC	97.20%	97.20%	97.20%	96.60%	97.20%	97.40%

ตารางผนวกที่ 4 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการคัดเลือกตัวด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูลที่แบ่งแบบ Cross – validation สำหรับจำแนกผู้ป่วยโรคไตเรื้อรังแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
ID3						
Accuracy	94.36%	94.36%	94.36%	93.70%	93.62%	93.95%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	9.70%	9.70%	9.70%	10.45%	10.40%	9.80%
PPV	80.63%	80.63%	80.63%	79.06%	79.24%	80.41%
NPV	100%	100%	100%	100%	100%	100%

ตารางผนวกที่ 4 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
AUC	96.10%	96.10%	96.10%	95.60%	95.80%	96.00%
C4.5						
Accuracy	94.36%	94.36%	94.36%	93.70%	93.62%	93.95%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	9.70%	9.70%	9.70%	10.45%	10.40%	9.80%
PPV	80.63%	80.63%	80.63%	79.06%	79.25%	80.41%
NPV	100%	100%	100%	100%	100%	100%
AUC	89.70%	89.70%	89.70%	89.50%	89.60%	90.20%
CART						
Accuracy	94.36%	94.36%	94.36%	94.69%	93.92%	94.18%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	7.36%	7.36%	7.36%	7.08%	7.99%	7.47%
PPV	80.63%	80.63%	80.63%	81.57%	79.78%	79.11%
NPV	100%	100%	100%	100%	100%	100%
AUC	95.00%	95.00%	95.00%	94.70%	94.90%	95.30%

ตารางผนวกที่ 5 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีมีการเลือกตัวแปรด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูล Cross – validation สำหรับจำแนกผู้ป่วยโรคไตเรื้อรังแบบเฉียบพลัน แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
ID3						
Accuracy	95.16%	95.16%	95.16%	94.71%	93.78%	94.95%
Kappa Statistic	0.8571	0.8571	0.8571	0.8427	0.8416	0.8532
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	8.52%	8.52%	8.52%	9.45%	10.32%	9.72%
PPV	82.43%	82.43%	82.43%	79.56%	79.25%	80.41%
NPV	100%	100%	100%	100%	100%	100%
AUC	96.30%	96.30%	96.30%	95.75%	95.90%	96.74%
C4.5						
Accuracy	94.36%	94.36%	94.36%	93.70%	93.62%	93.95%
Kappa Statistic	0.8551	0.8551	0.8551	0.8407	0.8410	0.8502
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	9.70%	9.70%	9.70%	10.45%	10.40%	9.80%
PPV	80.63%	80.63%	80.63%	79.06%	79.24%	80.41%
NPV	100%	100%	100%	100%	100%	100%
AUC	89.70%	89.70%	89.70%	89.50%	89.60%	90.20%

ตารางผนวกที่ 5 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
CART						
Accuracy	94.36%	94.36%	94.36%	94.70%	93.92%	94.18%
Kappa Statistic	0.8357	0.8357	0.8357	0.8377	0.8420	0.8478
TP Rate	100%	100%	100%	100%	100%	100%
FP Rate	7.36%	7.36%	7.36%	7.08%	7.99%	7.42%
PPV	80.63%	80.63%	80.63%	81.57%	79.78%	79.11%
NPV	100%	100%	100%	100%	100%	100%
AUC	96.60%	96.60%	96.60%	95.60%	95.80%	96.00%

ตารางผนวกที่ 6 ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ กรณีที่มีการเลือกตัวแปรด้วยวิธี Best First, Greedy Stepwise และ Genetic Search จากข้อมูล Cross – validation สำหรับจำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูลทดสอบ

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
ID3						
Accuracy	94.63%	94.63%	94.63%	94.41%	94.61%	94.68%
Kappa Statistic	0.8412	0.8412	0.8412	0.8362	0.8357	0.8455
TP Rate	93.99%	93.99%	93.99%	91.24%	93.19%	93.23%
FP Rate	9.85%	9.85%	9.85%	9.45%	9.45%	9.30%
PPV	81.87%	81.87%	81.87%	83.49%	83.09%	83.24%
NPV	98.44%	98.44%	98.44%	97.63%	98.12%	98.16%
AUC	96.00%	96.00%	96.00%	96.40%	96.50%	96.50%

ตารางผนวกที่ 6 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
C4.5						
Accuracy	94.61%	94.61%	94.61%	94.41%	94.68%	94.70%
Kappa Statistic	0.8403	0.8403	0.8403	0.8362	0.8446	0.8455
TP Rate	93.98%	93.98%	93.98%	91.24%	93.38%	93.42%
FP Rate	9.95%	9.95%	9.95%	9.45%	9.45%	9.30%
PPV	81.75%	81.75%	81.75%	83.49%	82.94%	83.24%
NPV	98.44%	98.44%	98.44%	97.63%	98.22%	98.16%
AUC	90.90%	90.90%	90.90%	91.70%	91.70%	92.20%
CART						
Accuracy	93.69%	93.69%	93.69%	91.90%	91.66%	92.58%
Kappa Statistic	0.8416	0.8416	0.8416	0.8464	0.8467	0.8469
TP Rate	92.07%	92.07%	92.07%	88.90%	87.29%	89.02%
FP Rate	5.90%	5.90%	5.90%	7.35%	7.27%	6.56%
PPV	79.37%	79.37%	79.37%	75.06%	74.60%	76.81%
NPV	97.96%	97.96%	97.96%	97.10%	96.76%	97.21%
AUC	95.90%	95.90%	95.90%	96.50%	96.60%	96.70%

ตารางผนวกที่ 7 ประสิทธิภาพของอัลกอริทึมเริ่มต้นไม่ตัดสินใจ ด้วย Best First, Greedy Stepwise และ Genetic Search จากการใช้ Use full training set โดย สำหรับ จำแนกผู้ป่วยโรคปอดอักเสบ แยกตามวิธีการแบ่งชุดข้อมูล

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross-validation	Cross-validation	Cross-validation	split	split	split
ID3						
Accuracy	95.85%	95.84%	95.84%	95.74%	95.87%	95.91%
Kappa Statistic	0.8821	0.8819	0.8818	0.8795	0.8842	0.8852
TP Rate	91.56%	91.40%	91.45%	90.28%	91.04%	91.55%
FP Rate	6.15%	6.05%	6.10%	5.90%	5.75%	5.90%
PPV	90.25%	90.37%	90.31%	91.15%	91.18%	90.84%
NPV	97.52%	97.47%	97.48%	97.10%	97.29%	97.45%
AUC	97.60%	97.60%	97.70%	97.70%	97.80%	97.90%
C4.5						
Accuracy	95.86%	95.88%	95.88%	95.52%	95.80%	95.91%
Kappa Statistic	0.8828	0.8832	0.8832	0.8738	0.8821	0.8853
TP Rate	91.91%	91.28%	91.27%	89.43%	91.14%	91.55%
FP Rate	6.00%	5.95%	5.95%	6.05%	6.00%	5.90%
PPV	90.61%	90.73%	90.73%	91.15%	90.74%	90.84%
NPV	97.43%	97.41%	97.41%	96.82%	97.33%	97.45%
AUC	96.80%	96.80%	96.70%	95.20%	96.50%	97.10%

ตารางผนวกที่ 7 (ต่อ)

	5 Folds	10 Folds	15 Folds	50%	60%	70%
Algorithms	Cross- validation	Cross- validation	Cross- validation	split	split	split
CART						
Accuracy	90.04%	90.04%	90.04%	89.47%	95.97%	95.95%
Kappa Statistic	0.8836	0.8836	0.8836	0.8838	0.8868	0.8866
TP Rate	94.42%	94.42%	94.42%	90.25%	91.57%	91.57%
FP Rate	10.67%	10.67%	10.67%	10.77%	2.71%	2.72%
PPV	71.39%	71.39%	71.39%	71.74%	91.03%	91.04%
NPV	98.27%	98.27%	98.27%	96.80%	97.47%	97.45%
AUC	94.60%	94.60%	94.60%	97.70%	97.70%	97.90%

ตัวอย่างผนวกที่ 1 ผลลัพธ์ที่ได้จากโปรแกรม Weka สำหรับอัลกอริทึม ID3 กรณีไม่มีการคัดเลือกตัวแปร กับการแบ่งชุดข้อมูลแบบ 10 Folds Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI)

=== Run information ===

Scheme: weka.classifiers.trees.Id3

Relation: URIandNoURI

Instances: 7327

Attributes: 23

โรค

เพศ

อายุ

ที่อยู่

อาชีพ

อุณหภูมิ

อาการไข้

ปวดศีรษะ

เวียนศีรษะ

เจ็บคอ

มีเสมหะ

ไอ

มีน้ำมูก

คัดจมูก

อาเจียน

เหนื่อยหอบ

หายใจมีเสียงหวีดๆ

หายใจมีกลิ่นเหม็น

แน่นหน้าอก

เจ็บหน้าอก

ปวดกระบอกตา

ปวดจุก

ปวดหู

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Id3

เจ็บคอ = Y

| เหนื่อยหอบ = N

| | อายุ = a4

| | | มีน้ำมูก = Y3

| | | | อุณหภูมิ = temadnor

| | | | | เพศ = F: URI

| | | | | เพศ = M: No_URI

| | | | | อุณหภูมิ = temnor: No_URI

| | | มีน้ำมูก = Y1

| | | | ปวดจุก = N

| | | | | มีเสมหะ = Y

| | | | | อาชีพ = Occ6

| | | | | | ปวดศีรษะ = N

| | | | | | ที่อยู่ = L1: URI

| | | | | | ที่อยู่ = L2

| | | | | | | อุณหภูมิ = temadnor: URI

| | | | | | | อุณหภูมิ = temnor

| | | | | | | | เพศ = F: No_URI

| | | | | | | | เพศ = M: No_URI

| | | | | | | | ที่อยู่ = L3: URI

| | | | | | | | ปวดศีรษะ = Y: URI

| | | | | | | | อาชีพ = Occ1: null

| | | | | | | | อาชีพ = Occ3: URI

| | | | | | | | อาชีพ = Occ2: null

| | | | | | | | อาชีพ = Occ4: null

| | | | | อาชีพ = Occ8: URI
 | | | | | อาชีพ = Occ5: URI
 | | | | | อาชีพ = Occ9: null
 | | | | | อาชีพ = Occ7: null
 | | | | | มีเสมหะ = N: URI
 | | | | | ปวดจมูก = Y: No_URI
 | | | | | มีน้ำมูก = N
 | | | | | หายใจมีกลิ่นเหม็น = N
 | | | | | เพศ = F: URI
 | | | | | เพศ = M
 | | | | | อาชีพ = Occ6
 | | | | | มีเสมหะ = Y
 | | | | | ที่อยู่ = L1: URI
 | | | | | ที่อยู่ = L2
 | | | | | อุณภูมิ = temadnor: No_URI
 | | | | | อุณภูมิ = temnor
 | | | | | อาเจียน = N: URI
 | | | | | อาเจียน = Y: URI
 | | | | | ที่อยู่ = L3: URI
 | | | | | มีเสมหะ = N: URI
 | | | | | อาชีพ = Occ1: null
 | | | | | อาชีพ = Occ3: URI
 | | | | | อาชีพ = Occ2: null
 | | | | | อาชีพ = Occ4: null
 | | | | | อาชีพ = Occ8: URI
 | | | | | อาชีพ = Occ5
 | | | | | อาการไข้ = Y: No_URI
 | | | | | อาการไข้ = N: URI
 | | | | | อาชีพ = Occ9: URI
 | | | | | อาชีพ = Occ7: null
 | | | | | หายใจมีกลิ่นเหม็น = Y: No_URI

| | | มีน้ำมูก = Y2
 | | | ไอ = N: URI
 | | | ไอ = Y
 | | | | ที่อยู่ = L1: URI
 | | | | ที่อยู่ = L2: No_URI
 | | | | ที่อยู่ = L3: null
 | | | มีน้ำมูก = Y4
 | | | ปวดหู = N: URI
 | | | ปวดหู = Y: No_URI
 | | อายุ = a2
 | | เพศ = F
 | | | อาชีพ = Occ6: null
 | | | อาชีพ = Occ1
 | | | | อุณภูมิ = temadnor
 | | | | มีเสมหะ = Y: URI
 | | | | มีเสมหะ = N: No_URI
 | | | | อุณภูมิ = temnor: URI
 | | | | อาชีพ = Occ3: null
 | | | | อาชีพ = Occ2: URI
 | | | | อาชีพ = Occ4: null
 | | | | อาชีพ = Occ8: null
 | | | | อาชีพ = Occ5: null
 | | | | อาชีพ = Occ9: null
 | | | | อาชีพ = Occ7: null
 | | เพศ = M
 | | | ที่อยู่ = L1
 | | | | ไอ = N: No_URI
 | | | | ไอ = Y: URI
 | | | | ที่อยู่ = L2: No_URI
 | | | | ที่อยู่ = L3: URI
 | | อายุ = a3

| | | ปวดจุก = N: URI
 | | | ปวดจุก = Y: No_URI
 | | อายุ = a5
 | | | มีเสมหะ = Y: No_URI
 | | | มีเสมหะ = N: URI
 | | อายุ = a1
 | | | อาการไอ = Y: No_URI
 | | | อาการไอ = N: URI
 | | เหนื่อยหอบ = Y: No_URI
 | | | เจ็บคอ = N
 | | | | อาการไอ = Y
 | | | | มีน้ำมูก = Y3
 | | | | ไอ = N
 | | | | | มีเสมหะ = Y: No_URI
 | | | | | มีเสมหะ = N: URI
 | | | | | ไอ = Y: No_URI
 | | | | | มีน้ำมูก = Y1
 | | | | | หายใจมีเสียงหวีดๆ = N
 | | | | | หายใจมีกลิ่นเหม็น = N
 | | | | | อาเจียน = N
 | | | | | ปวดกระบอกตา = N
 | | | | | | อายุ = a4: URI
 | | | | | | อายุ = a2
 | | | | | | มีเสมหะ = Y: URI
 | | | | | | มีเสมหะ = N: No_URI
 | | | | | | อายุ = a3: URI
 | | | | | | อายุ = a5: URI
 | | | | | | อายุ = a1: null

| | | | | ปวดกระบอกตา = Y: No_URI
 | | | | | อาเจียน = Y
 | | | | | อายุ = a4: null
 | | | | | อายุ = a2: No_URI
 | | | | | อายุ = a3: null
 | | | | | อายุ = a5: URI
 | | | | | อายุ = a1: null
 | | | | หายใจมีกลิ่นเหม็น = Y: No_URI
 | | | หายใจมีเสียงหวีดๆ = Y: No_URI
 | | มีน้ำมูก = N
 | | | เหนื่อยหอบ = N
 | | | อาเจียน = N
 | | | | อาชีพ = Occ6
 | | | | | มีเสมหะ = Y
 | | | | | อายุ = a4
 | | | | | ไอ = N: No_URI
 | | | | | ไอ = Y
 | | | | | | อุณหภูมิ = temadnor: No_URI
 | | | | | | อุณหภูมิ = temnor: URI
 | | | | | อายุ = a2: null
 | | | | | อายุ = a3: URI
 | | | | | อายุ = a5: null
 | | | | | อายุ = a1: null
 | | | | | มีเสมหะ = N
 | | | | | ปวดกระบอกตา = N
 | | | | | ปวดจมูก = N: URI
 | | | | | ปวดจมูก = Y: No_URI
 | | | | | ปวดกระบอกตา = Y: No_URI
 | | | | | อาชีพ = Occ1: URI
 | | | | | อาชีพ = Occ3
 | | | | | ไอ = N: No_URI

| | | | | ไอ = Y
 | | | | | อายุ = a4: URI
 | | | | | อายุ = a2: null
 | | | | | อายุ = a3: No_URI
 | | | | | อายุ = a5: null
 | | | | | อายุ = a1: null
 | | | | | อาชีพ = Occ2
 | | | | | อายุ = a4
 | | | | | ไอ = N: No_URI
 | | | | | ไอ = Y: URI
 | | | | | อายุ = a2
 | | | | | อุณภูมิ = temadnor
 | | | | | เพศ = F
 | | | | | ที่อยู่ = L1: URI
 | | | | | ที่อยู่ = L2: URI
 | | | | | ที่อยู่ = L3: URI
 | | | | | เพศ = M: No_URI
 | | | | | อุณภูมิ = temnor: URI
 | | | | | อายุ = a3
 | | | | | มีเสมหะ = Y
 | | | | | ปวดศีรษะ = N: No_URI
 | | | | | ปวดศีรษะ = Y: URI
 | | | | | มีเสมหะ = N: URI
 | | | | | อายุ = a5: null
 | | | | | อายุ = a1: null
 | | | | | อาชีพ = Occ4
 | | | | | อุณภูมิ = temadnor: No_URI
 | | | | | อุณภูมิ = temnor: URI
 | | | | | อาชีพ = Occ8: null
 | | | | | อาชีพ = Occ5: URI
 | | | | | อาชีพ = Occ9: No_URI

| | | | อาชีพ = Occ7: URI
 | | | | อาเจียน = Y
 | | | | เวียนศีรษะ = N
 | | | | อายุ = a4: No_URI
 | | | | อายุ = a2: No_URI
 | | | | อายุ = a3
 | | | | เพศ = F: No_URI
 | | | | เพศ = M: URI
 | | | | อายุ = a5: null
 | | | | อายุ = a1: null
 | | | | เวียนศีรษะ = Y
 | | | | เพศ = F: URI
 | | | | เพศ = M: No_URI
 | | | เห็น้อยหอบ = Y: No_URI
 | | มีน้ำมูก = Y2: No_URI
 | | มีน้ำมูก = Y4: No_URI
 | | อาการไข้ = N
 | | ปวดจมูก = N
 | | | เห็น้อยหอบ = N
 | | | มีน้ำมูก = Y3
 | | | | อาชีพ = Occ6: No_URI
 | | | | อาชีพ = Occ1: null
 | | | | อาชีพ = Occ3: No_URI
 | | | | อาชีพ = Occ2: No_URI
 | | | | อาชีพ = Occ4: URI
 | | | | อาชีพ = Occ8: null
 | | | | อาชีพ = Occ5: No_URI
 | | | | อาชีพ = Occ9: No_URI
 | | | | อาชีพ = Occ7: null
 | | | | มีน้ำมูก = Y1
 | | | | ปวดหู = N

| | | | | อาชีพ = Occ6
 | | | | | ปวดกระบอกตา = N
 | | | | | หายใจมีกลิ่นเหม็น = N
 | | | | | หายใจมีเสียงหวีดๆ = N
 | | | | | ไอ = N
 | | | | | ที่อยู่ = L1: URI
 | | | | | ที่อยู่ = L2
 | | | | | อายุ = a4
 | | | | | ปวดศีรษะ = N
 | | | | | คัดจมูก = N
 | | | | | มีเสมหะ = Y: No_URI
 | | | | | มีเสมหะ = N: URI
 | | | | | คัดจมูก = Y: URI
 | | | | | ปวดศีรษะ = Y: URI
 | | | | | อายุ = a2: No_URI
 | | | | | อายุ = a3: No_URI
 | | | | | อายุ = a5: null
 | | | | | อายุ = a1: null
 | | | | | ที่อยู่ = L3: URI
 | | | | | ไอ = Y: URI
 | | | | | หายใจมีเสียงหวีดๆ = Y: No_URI
 | | | | | หายใจมีกลิ่นเหม็น = Y: No_URI
 | | | | | ปวดกระบอกตา = Y: No_URI
 | | | | | อาชีพ = Occ1
 | | | | | เพศ = F: URI
 | | | | | เพศ = M: No_URI
 | | | | | อาชีพ = Occ3
 | | | | | ที่อยู่ = L1: URI
 | | | | | ที่อยู่ = L2: null
 | | | | | ที่อยู่ = L3: No_URI
 | | | | | อาชีพ = Occ2

| | | | | | | | | | มีเสมหะ = Y
 | | | | | | | | | | ไอ = N: No_URI
 | | | | | | | | | | ไอ = Y
 | | | | | | | | | | อาเจียน = N
 | | | | | | | | | | ที่อยู่ = L1: URI
 | | | | | | | | | | ที่อยู่ = L2
 | | | | | | | | | | เพศ = F: URI
 | | | | | | | | | | เพศ = M: URI
 | | | | | | | | | | ที่อยู่ = L3
 | | | | | | | | | | ถัดจุมูก = N: No_URI
 | | | | | | | | | | ถัดจุมูก = Y: URI
 | | | | | | | | | | อาเจียน = Y: No_URI
 | | | | | | | | | | มีเสมหะ = N: No_URI
 | | | | | | | | | | อาชีพ = Occ4: No_URI
 | | | | | | | | | | อาชีพ = Occ8: URI
 | | | | | | | | | | อาชีพ = Occ5: URI
 | | | | | | | | | | อาชีพ = Occ9: null
 | | | | | | | | | | อาชีพ = Occ7: No_URI
 | | | | | | | | | | ปวดหู = Y: No_URI
 | | | | | | | | | | มีน้ำมูก = N
 | | | | | | | | | | ปวดศีรษะ = N
 | | | | | | | | | | เวียนศีรษะ = N
 | | | | | | | | | | อาชีพ = Occ6: No_URI
 | | | | | | | | | | อาชีพ = Occ1
 | | | | | | | | | | ที่อยู่ = L1: No_URI
 | | | | | | | | | | ที่อยู่ = L2: URI
 | | | | | | | | | | ที่อยู่ = L3: null
 | | | | | | | | | | อาชีพ = Occ3: No_URI
 | | | | | | | | | | อาชีพ = Occ2: No_URI
 | | | | | | | | | | อาชีพ = Occ4: No_URI
 | | | | | | | | | | อาชีพ = Occ8: No_URI

| | | | | อาชีพ = Occ5: No_URI
 | | | | | อาชีพ = Occ9: No_URI
 | | | | | อาชีพ = Occ7
 | | | | | ปวดหู = N
 | | | | | มีเสมหะ = Y: No_URI
 | | | | | มีเสมหะ = N
 | | | | | ไอ = N
 | | | | | เพศ = F: No_URI
 | | | | | เพศ = M: URI
 | | | | | ไอ = Y: URI
 | | | | | ปวดหู = Y: No_URI
 | | | | | เวียนศีรษะ = Y
 | | | | | อายุ = a4: null
 | | | | | อายุ = a2: URI
 | | | | | อายุ = a3: URI
 | | | | | อายุ = a5: No_URI
 | | | | | อายุ = a1: null
 | | | | | ปวดศีรษะ = Y
 | | | | | ปวดหู = N
 | | | | | ปวดกระบอกตา = N
 | | | | | อาชีพ = Occ6: URI
 | | | | | อาชีพ = Occ1: null
 | | | | | อาชีพ = Occ3: No_URI
 | | | | | อาชีพ = Occ2: URI
 | | | | | อาชีพ = Occ4: null
 | | | | | อาชีพ = Occ8: null
 | | | | | อาชีพ = Occ5: null
 | | | | | อาชีพ = Occ9: null
 | | | | | อาชีพ = Occ7: null
 | | | | | ปวดกระบอกตา = Y: No_URI
 | | | | | ปวดหู = Y: No_URI

| | | มีน้ำมูก = Y2: No_URI

| | | มีน้ำมูก = Y4: No_URI

| | | เหนื่อยหอบ = Y: No_URI

| | ปวดจมูก = Y: No_URI

Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 7264 99.1402 %

Incorrectly Classified Instances 55 0.7506 %

Kappa statistic 0.9849

Mean absolute error 0.008

Root mean squared error 0.0853

Relative absolute error 1.5985 %

Root relative squared error 17.0825 %

UnClassified Instances 8 0.1092 %

Total Number of Instances 7327

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.993	0.008	0.992	0.993	0.992	0.993	URI
0.992	0.007	0.993	0.992	0.993	0.993	No_URI

=== Confusion Matrix ===

a b <-- classified as

3485 26 | a = URI

29 3779 | b = No_URI

ตัวอย่างผนวกที่ 2 ผลลัพธ์ที่ได้จากโปรแกรม Weka สำหรับอัลกอริทึม C4.5 กรณีไม่มีการคัดเลือกตัวแปร กับการแบ่งชุดข้อมูลแบบ 5 Folds Cross-validation สำหรับจำแนกผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน (Acute URI)

==== Run information ====

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: URIandNoURI

Instances: 7327

Attributes: 23

โรค

เพศ

อายุ

ที่อยู่

อาชีพ

อุณหภูมิ

อาการไข้

ปวดศีรษะ

เวียนศีรษะ

เจ็บคอ

มีเสมหะ

ไอ

มีน้ำมูก

คัดจมูก

อาเจียน

เหนื่อยหอบ

หายใจมีเสียงหวีดๆ

หายใจมีกลิ่นเหม็น

แน่นหน้าอก

เจ็บหน้าอก

ปวดกระบอกตา

ปวดจมูก

ปวดหู

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

เหนือหอบ = N

| ปวดจมูก = N

| | ปวดหู = N

| | | หายใจมีกลิ่นเหม็น = N

| | | | อาเจียน = N

| | | | | ปวดกระบอกตา = N

| | | | | หายใจมีเสียงหวีดๆ = N

| | | | | มีน้ำมูก = Y3

| | | | | | อาการไอ = Y

| | | | | | | ที่อยู่ = L1

| | | | | | | | อาชีพ = Occ6

| | | | | | | | | อุณหภูมิ = temadnor: URI (2.0)

| | | | | | | | | อุณหภูมิ = temnor: No_URI (2.0)

| | | | | | | | | | อาชีพ = Occ1: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ3: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ2: No_URI (7.0)

| | | | | | | | | | อาชีพ = Occ4: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ8: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ5: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ9: No_URI (0.0)

| | | | | | | | | | อาชีพ = Occ7: No_URI (0.0)

| | | | | | | | | | ที่อยู่ = L2: URI (34.0/1.0)

| | | | | | | | | | ที่อยู่ = L3: URI (3.0)

| | | | | | | | | | อาการไอ = N

| | | | | | | | | | มีเสมหะ = Y: No_URI (79.0/1.0)

| | | | | | | | | | มีเสมหะ = N

| | | | | | | | | อาชีพ = Occ6: No_URI (5.0)
 | | | | | | | | | อาชีพ = Occ1: URI (0.0)
 | | | | | | | | | อาชีพ = Occ3: No_URI (3.0)
 | | | | | | | | | อาชีพ = Occ2: No_URI (1.0)
 | | | | | | | | | อาชีพ = Occ4: URI (23.0)
 | | | | | | | | | อาชีพ = Occ8: URI (0.0)
 | | | | | | | | | อาชีพ = Occ5: No_URI (1.0)
 | | | | | | | | | อาชีพ = Occ9: No_URI (3.0)
 | | | | | | | | | อาชีพ = Occ7: URI (0.0)
 | | | | | | | มีน้ำมูก = Y1
 | | | | | | | อายุ = a4
 | | | | | | | มีเสมหะ = Y
 | | | | | | | | | อาชีพ = Occ6
 | | | | | | | | | อาการไอ = Y: URI (115.0)
 | | | | | | | | | อาการไอ = N
 | | | | | | | | | ที่อยู่ = L1: URI (13.0)
 | | | | | | | | | ที่อยู่ = L2
 | | | | | | | | | เจ็บคอ = Y
 | | | | | | | | | ปวดศีรษะ = N: No_URI (12.0/1.0)
 | | | | | | | | | ปวดศีรษะ = Y: URI (10.0)
 | | | | | | | | | เจ็บคอ = N: URI (18.0/1.0)
 | | | | | | | | | ที่อยู่ = L3: URI (6.0)
 | | | | | | | | | อาชีพ = Occ1: URI (0.0)
 | | | | | | | | | อาชีพ = Occ3: URI (19.0/1.0)
 | | | | | | | | | อาชีพ = Occ2: URI (0.0)
 | | | | | | | | | อาชีพ = Occ4: URI (0.0)
 | | | | | | | | | อาชีพ = Occ8: URI (18.0)
 | | | | | | | | | อาชีพ = Occ5: URI (29.0)
 | | | | | | | | | อาชีพ = Occ9: URI (0.0)
 | | | | | | | | | อาชีพ = Occ7: No_URI (8.0)
 | | | | | | | | | มีเสมหะ = N: URI (736.0/2.0)

| | | | | | | | อายุ = a2
 | | | | | | | | มีเสมหะ = Y: URI (104.0/3.0)
 | | | | | | | | มีเสมหะ = N
 | | | | | | | | อุณภูมิ = temadnor
 | | | | | | | | คัดจมูก = N: No_URI (60.0)
 | | | | | | | | คัดจมูก = Y
 | | | | | | | | อาการไอ = Y: URI (9.0)
 | | | | | | | | อาการไอ = N: No_URI (4.0)
 | | | | | | | | อุณภูมิ = temnor
 | | | | | | | | เจ็บคอ = Y: URI (14.0)
 | | | | | | | | เจ็บคอ = N: No_URI (3.0)
 | | | | | | | | อายุ = a3
 | | | | | | | | คัดจมูก = N
 | | | | | | | | อาการไอ = Y: URI (433.0)
 | | | | | | | | อาการไอ = N
 | | | | | | | | เจ็บคอ = Y: URI (133.0)
 | | | | | | | | เจ็บคอ = N
 | | | | | | | | อุณภูมิ = temadnor: No_URI (7.0)
 | | | | | | | | อุณภูมิ = temnor: URI (33.0/1.0)
 | | | | | | | | คัดจมูก = Y
 | | | | | | | | ไอ = N
 | | | | | | | | เพศ = F: URI (2.0)
 | | | | | | | | เพศ = M: No_URI (39.0/1.0)
 | | | | | | | | ไอ = Y: URI (29.0)
 | | | | | | | | อายุ = a5: URI (54.0)
 | | | | | | | | อายุ = a1: No_URI (32.0)
 | | | | | | | | มีน้ำมูก = N
 | | | | | | | | เจ็บคอ = Y: URI (758.0/6.0)
 | | | | | | | | เจ็บคอ = N
 | | | | | | | | ปวดศีรษะ = N
 | | | | | | | | อาชีพ = Occ6

| | | | | | | | | | ปวดศีรษะ = Y: URI (385.0/3.0)

| | | | | | | | | | มีน้ำมูก = Y2

| | | | | | | | | | เจ็บคอ = Y: URI (72.0/1.0)

| | | | | | | | | | เจ็บคอ = N: No_URI (189.0)

| | | | | | | | | | มีน้ำมูก = Y4

| | | | | | | | | | เจ็บคอ = Y: URI (140.0)

| | | | | | | | | | เจ็บคอ = N: No_URI (6.0)

| | | | | | | | | | หายใจมีเสียงหวีดๆ = Y: No_URI (81.0)

| | | | | | | | | | ปวดกระบอกตา = Y: No_URI (135.0)

| | | | | | | | | | อาเจียน = Y

| | | | | | | | | | เวียนศีรษะ = N

| | | | | | | | | | เจ็บคอ = Y

| | | | | | | | | | อายุ = a4: URI (7.0)

| | | | | | | | | | อายุ = a2: No_URI (12.0/1.0)

| | | | | | | | | | อายุ = a3: URI (6.0)

| | | | | | | | | | อายุ = a5: URI (0.0)

| | | | | | | | | | อายุ = a1: URI (0.0)

| | | | | | | | | | เจ็บคอ = N: No_URI (430.0/3.0)

| | | | | | | | | | เวียนศีรษะ = Y

| | | | | | | | | | เพศ = F: URI (31.0)

| | | | | | | | | | เพศ = M: No_URI (4.0/1.0)

| | | | | | | | | | หายใจมีกลิ่นเหม็น = Y: No_URI (191.0)

| | | | | | | | | | ปวดหู = Y: No_URI (396.0)

| | | | | | | | | | ปวดจมูก = Y: No_URI (821.0)

| | | | | | | | | | เหนื่อยหอบ = Y: No_URI (1153.0)

Number of Leaves : 94

Size of the tree : 145

Time taken to build model: 0.28 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7267	99.1811 %
Incorrectly Classified Instances	60	0.8189 %
Kappa statistic	0.9836	
Mean absolute error	0.0118	
Root mean squared error	0.0867	
Relative absolute error	2.3667 %	
Root relative squared error	17.3639 %	
Total Number of Instances	7327	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.993	0.009	0.99	0.993	0.991	0.995	URI
0.991	0.007	0.993	0.991	0.992	0.995	No_URI

=== Confusion Matrix ===

```

a  b  <-- classified as
3488 25 | a = URI
35 3779 | b = No_URI

```


ตัวอย่างผนวกที่ 3 ผลลัพธ์ที่ได้จากโปรแกรม Weka สำหรับอัลกอริทึม C4.5 กรณีที่มีการเลือกตัวแปรแบบ Genetic Search จากข้อมูล Use full training set และทำการสร้างตัวจำแนกผู้ป่วยโรค การติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน จากข้อมูลที่แบ่งแบบ 5 Folds Cross-validation

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: URIandNoURI-Weka.filters.unsupervised.attribute.Remove-R2-6,8,11-14,17-21

Instances: 7327

Attributes: 8

โรค

อาการไข้

เวียนศีรษะ

เจ็บคอ

อาเจียน

เหนื่อยหอบ

ปวดจมูก

ปวดหู

Test mode: 5-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

เหนื่อยหอบ = N

| ปวดจมูก = N

| | ปวดหู = N

| | | อาเจียน = N

| | | | เจ็บคอ = Y: URI (2090.0/109.0)

| | | | เจ็บคอ = N

| | | | | อาการไข้ = Y: URI (1341.0/202.0)

| | | | | อาการไข้ = N

| | | | | | เวียนศีรษะ = N: No_URI (962.0/284.0)

| | | | | เวียนศีรษะ = Y: URI (67.0/7.0)

| | | อาเจียน = Y

| | | | เวียนศีรษะ = N

| | | | | เจ็บคอ = Y

| | | | | อาการไอ = Y: No_URI (16.0/7.0)

| | | | | อาการไอ = N: URI (9.0/2.0)

| | | | | เจ็บคอ = N: No_URI (437.0/3.0)

| | | | เวียนศีรษะ = Y: URI (35.0/3.0)

| | ปวดหู = Y: No_URI (396.0)

| | ปวดจมูก = Y: No_URI (821.0)

เหนื่อยหอบ = Y: No_URI (1153.0)

Number of Leaves : 11

Size of the tree : 21

Time taken to build model: 0.39 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6702	91.4699 %
Incorrectly Classified Instances	625	8.5301 %
Kappa statistic	0.8292	
Mean absolute error	0.1354	
Root mean squared error	0.2611	
Relative absolute error	27.1189 %	
Root relative squared error	52.2652 %	
Total Number of Instances	7327	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.915	0.086	0.907	0.915	0.911	0.955	URI
0.914	0.085	0.921	0.914	0.918	0.955	No_URI

=== Confusion Matrix ===

a b <-- classified as

3216 297 | a = URI

328 3486 | b = No_URI

ตัวอย่างผนวกที่ 4 ผลลัพธ์ที่ได้จากโปรแกรม SPSS สำหรับอัลกอริทึม CART ที่มีการคัดเลือกตัวแปรแบบ Genetic Search จาก Use full training set และสร้างตัวจำแนก ผู้ป่วยโรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลัน จากข้อมูลที่แบ่งแบบ 5 Folds Cross-validation

Model Summary

<i>Specifications</i>	<i>Growing Method</i>	CRT
	<i>Dependent Variable</i>	Diseases
	<i>Independent Variables</i>	Fever, Dizzines, Sore throat, Vomiting, Dyspnea, Nose pain, Ear pain
	<i>Validation</i>	Cross Validation
	<i>Maximum Tree Depth</i>	5
	<i>Minimum Cases in Parent Node</i>	100
	<i>Minimum Cases in Child Node</i>	50
<i>Results</i>	<i>Independent Variables Included</i>	Sore throat, Dyspnea, Ear pain, Nose pain, Fever, Dizzines, Vomiting
	<i>Number of Nodes</i>	19
	<i>Number of Terminal Nodes</i>	10
	<i>Depth</i>	5

Risk

<i>Method</i>	<i>Estimate</i>	<i>Std. Error</i>
<i>Resubstitution</i>	.098	.003
<i>Cross-Validation</i>	.098	.003

Growing Method: CRT

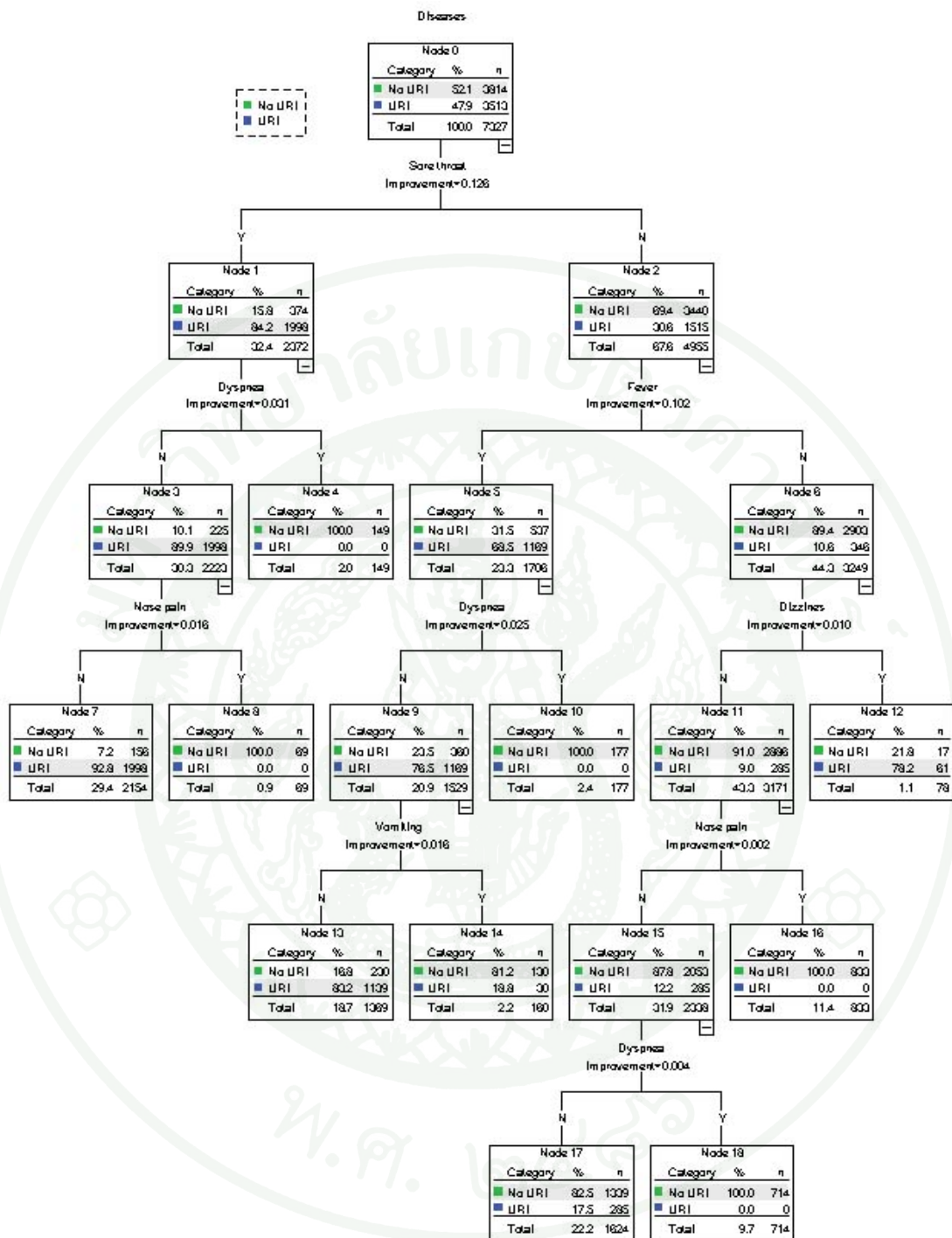
Dependent Variable: Diseases

Classification

<i>Observed</i>	<i>Predicted</i>		
	<i>No URI</i>	<i>URI</i>	<i>Percent Correct</i>
<i>No URI</i>	3411	403	89.4%
<i>URI</i>	315	3198	91.0%
<i>Overall Percentage</i>	50.9%	49.1%	90.2%

Growing Method: CRT

Dependent Variable: Diseases



ภาพผนวกที่ 1 ต้นไม้การจำแนก สำหรับอัลกอริทึม CART กรณีที่มีการเลือกตัวแปร ด้วยวิธี Genetic Search จาก Use full training set และสร้างต้นไม้การจำแนกผู้ป่วย โรคการติดเชื้อทางเดินหายใจส่วนบนแบบเฉียบพลันจากข้อมูลที่ แบ่งแบบ 5 Folds Cross-validation

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายดิษฐ์พล มั่นธรรม
วัน เดือน ปี ที่เกิด	31 พฤษภาคม 2526
สถานที่เกิด	อำเภอพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา
ประวัติการศึกษา	ระดับปริญญาตรี คณะศึกษาศาสตร์ วิชาเอกคณิตศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ตำแหน่งปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและ/หรือรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-