การประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ : กรณีศึกษาที่โรงพยาบาลพระนครศรีอยุธยา

ดิษฐพล มั่นธรรม*
ลี่ลี อิงศรีสว่าง*

บทคัดย่อ การศึกษานี้มีวัตถุประสงค์ ๑) เพื่อประยุกต์วิธีสืบค้นความรู้จากฐานข้อมูล โดยใช้เทคนิคขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ ในการจำแนกผู้ป่วย ๓ โรค คือ โรคติดเชื้อทางหายใจส่วนบนเฉียบพลัน, โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และโรคปอดอักเสบ ของโรงพยาบาลพระนครศรีอยุธยา; และ ๒) เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีต้นไม้ตัดสินใจ ๓ วิธี คือ ID3 C4.5 และ CART ในการจำแนกหรือคัดกรองผู้ป่วย ๓ โรคข้างต้น. ข้อมูลที่ศึกษาได้จากเวชระเบียนผู้ป่วยนอกโรคทางหายใจ ของโรงพยาบาลพระนครศรีอยุธยา ช่วง พ.ศ. ๒๕๔๗ - ๒๕๔๘ จำนวน ๗,๓๒๗ ราย. ตัวแปรที่นำมาพิจารณาประกอบด้วย อายุ อุณหภูมิร่างกาย เขตที่อยู่อาศัย อาชีพ อาการต่างๆ เช่น มีน้ำมูก มีไข้ คัดจมูก แน่นหน้าอก ปวดกระบอกตา ปวดศีรษะ หายใจเหนื่อยหอบ ไอ. วิธีการศึกษา ทำการสืบค้นความรู้จากฐานข้อมูลผู้ป่วยโรคระบบการหายใจด้วยขั้นตอนวิธี ID3 C4.5 และ CART โดยคัดกรองตัวแปรสำคัญต่อการจำแนกผู้ป่วยแต่ละโรค พร้อมทั้งเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีทั้งสามในการจำแนกผู้ป่วยแต่ละโรคด้วยการแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบตามหลักการทำให้ถูกต้องไขว้ (cross-validation) และการแยกค่าร้อยละ (percentage split).

จากการศึกษาวิธีการสืบค้นความรู้สำหรับจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลันพบว่าการใช้ตัวแปรที่คัดเลือกได้เพียง ๗ ตัวแปร กับอัตราส่วนข้อมูลฝึกสอนต่อข้อมูลทดสอบ ๗๐ : ๓๐. ขั้นตอนวิธี C4.5 ให้ค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้องของการจำแนกร้อยละ ๘๒.๓๒. ส่วนการสืบค้นความรู้สำหรับจำแนกผู้ป่วยโรคปอดอักเสบ พบว่าการใช้ตัวแปรที่คัดเลือกได้เพียง ๘ ตัวแปรกับอัตราส่วนข้อมูลฝึกสอนต่อข้อมูลทดสอบ ๗๐ : ๓๐. ขั้นตอนวิธี C4.5 ให้ค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้องของการจำแนกร้อยละ ๘๔.๗๐ และการสืบค้นความรู้สำหรับจำแนกผู้ป่วยโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน พบว่าการใช้ตัวแปรที่คัดเลือกได้เพียง ๗ ตัวแปร กับอัตราส่วนข้อมูลฝึกสอนต่อข้อมูลทดสอบ ๕๐:๕๐. ขั้นตอนวิธี CART ให้ค่าวัดประสิทธิภาพสูงสุด คือให้ค่าความถูกต้องของการจำแนกร้อยละ ๘๔.๖๘ ซึ่งผลลัพธ์ที่ได้จากการประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจ สามารถนำไปเป็นแนวทางสนับสนุนการคัดกรองเบื้องต้นผู้ป่วยโรคระบบการหายใจเพื่อการวินิจฉัยยืนยันต่อไป.

คำสำคัญ: *การจำแนกข้อมูล, ขั้นตอนวิธีต้นไม้ตัดสินใจ, โรคระบบการหายใจ, การวินิจฉัยโรค, การทำให้ถูกต้องไขว้*

*ภาควิชาสถิติ คณะวิทยาศาสตร์, มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพมหานคร

Abstract **An Application of Decision Tree Algorithms for Diagnosis of the Respiratory System: A Case Study of Pranakorn Sri Ayudthaya Hospital**

Dittapol Muntham*, Lily Ingsrisawang*

**Department of Statistics, Faculty of Science, Kasetsart University*

The objectives of this study involved (a) the application of methods of knowledge discovery from database using decision tree algorithms for respiratory system diagnosis to classify patients of the Pranakorn Sri Ayudthaya Hospital into three groups: acute upper respiratory tract infection, acute sinusitis, and pneumonia, and (b) the comparison of performance of the three decision tree algorithms, i.e., ID3, C4.5, and CART, for the classification or screening of 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 who attended Pranakorn Sri Ayudthaya Hospital in the period from 2003 to 2006. The variables considered were age, body temperature, residential area, occupation, and certain symptoms, e.g., rhinorrhea, fever, nasal congestion, periorbital pain, headache, wheezing and coughing. The study methods were knowledge discovery with the employment of ID3, C4.5, and CART decision tree algorithms from the hospital's medical records and determination of the effectiveness of the three algorithms. The validity of the decision tree algorithms was studied by dividing the data into two sets: training and testing data sets, which were based on the cross-validation and the percentage split methods.

The results of the knowledge discovery method found that, for the patients with acute URI with only seven selected variables and a ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective, with a classification accuracy of 92.31 per cent. For the classification of the patients with acute sinusitis with only eight selected variables and ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective, with a classification accuracy of 94.70 per cent. For the classification of the patients with pneumonia with only seven selected variables and ratio 50:50 of the training data set and the testing data set, the CART algorithm was the most effective, with a classification accuracy of 94.69 per cent. The results obtained could be used to support the diagnosis of patients with respiratory diseases.

Key words: *data classification, decision tree, respiratory diseases, diagnosis, cross-validation*

ภูมิหลังและเหตุผล

โรคระบบการหายใจเป็นปัญหาที่สำคัญของประเทศไทย. จากรายงานการเฝ้าระวังโรคประจำปีของสำนักระบาดวิทยา พบว่าโรคระบบการหายใจมีอัตราการป่วยต่อแสนประชากรเท่ากับ ๓๐๘.๒๑, ๓๑๑.๘๐ และ ๓๑๕.๓๓ ใน พ.ศ. ๒๕๔๖ - ๒๕๔๘ ตามลำดับ และมีโรคทางหายใจที่เกี่ยวข้องกับการทำงานในภาคส่วนอุตสาหกรรมเพิ่มขึ้น ได้แก่ โรคภูมิแพ้ โรคหืด โรคปอดอักเสบ เป็นต้น[๑]. จังหวัดพระนครศรีอยุธยาเป็นจังหวัดที่มีนิคมอุตสาหกรรมเป็นอันดับ ๑ ของภาคกลาง[๒] และจากรายงานสรุปจำนวนผู้ป่วยนอกของฝ่ายเวชระเบียนและสถิติประจำ พ.ศ. ๒๕๔๕ - ๒๕๔๙ พบว่า โรคทางหายใจเป็นกลุ่มที่มีจำนวนผู้ป่วยเข้ามารับการรักษามากที่สุดเป็นอันดับ ๑ ในอัตราการเข้ารับการรักษาร้อยละ ๑๔.๔๗, ๑๕.๑๓ และ ๑๕.๗๓ ช่วง พ.ศ. ๒๕๔๖ - ๒๕๔๘ ตามลำดับ โดยโรคที่เข้ามารับการรักษามากที่สุด ๓ อันดับ ได้แก่ โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน[๓], โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และโรคปอดอักเสบ ตามลำดับ ซึ่งมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง.

ปัจจุบันได้มีการเรียนรู้และนำเทคนิคเกี่ยวกับขั้นตอนวิธีต้นไม้ตัดสินใจ (decision tree algorithm) มาใช้ทางการแพทย์เพื่อจำแนกผู้ป่วยโรคต่างๆ เช่น โรคหัวใจเต้นผิดจังหวะ[๔], โรคหลอดเลือดแดง[๕], โรคเบาหวาน โรคตับอักเสบ โรคหัวใจ และโรคผิวหนัง[๖], วัณโรคและโรคเยื่อหุ้มปอดอักเสบ[๗]. สำหรับประเทศไทยก็มีการนำเทคนิคขั้นตอนวิธีต้นไม้ตัดสินใจมาใช้

ในด้านการแพทย์ เช่น พุธษดี[๘] ใน พ.ศ. ๒๕๕๑ ทำการศึกษาระบบจำแนกประเภทแบบทดสอบสุขภาพจิต, วราภรณ์[๙] ใน พ.ศ. ๒๕๕๑ วิเคราะห์ปัจจัยเสี่ยงและจัดจำแนกกลุ่มการดื้อยาในผู้ป่วยวัณโรคปอดกลับด้วยการวิเคราะห์การถดถอยโลจิสติกส์ และการวิเคราะห์การจำแนกขั้นตอนวิธีต้นไม้.

คณะผู้ศึกษานี้จึงสนใจที่จะประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ เพื่อจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน, โรคปอดอักเสบ และโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และเพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีต้นไม้ตัดสินใจ ๓ วิธี คือ ID3, C4.5 และ CART ในการจำแนกหรือคัดกรองผู้ป่วย ๓ โรคดังกล่าว.

ระเบียบวิธีศึกษา

ข้อมูลและแหล่งข้อมูล

ข้อมูลที่ใช้ศึกษาได้จากเวชระเบียนผู้ป่วยนอกของโรงพยาบาลพระนครศรีอยุธยา เฉพาะผู้ที่เข้ามารับบริการรักษาเป็นครั้งแรกในช่วง พ.ศ. ๒๕๔๗-๒๕๔๙ ด้วยโรคทางหายใจ ๓ โรค คือ โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน, โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และ โรคปอดอักเสบ ตั้งแต่เริ่มมีอาการจนกระทั่งได้รับการวินิจฉัยในช่วง ๒ สัปดาห์ มีจำนวน ๗,๓๒๗ คน. ตัวแปรตาม คือ ผลการวินิจฉัยจากแพทย์ที่ระบุว่าเป็นโรคทางหายใจ ๓ โรคที่ต้องการศึกษา. ส่วนตัวแปรอิสระที่เกี่ยวข้องกับผู้ป่วยมี ๒๒ ตัว ได้แก่ อายุ (ปี) อุณหภูมิกาย (องศาเซลเซียส) เขตที่อยู่อาศัย อาชีพ และอาการ (มีน้ำมูก มีเสมหะ มีไข้ คัดจมูก แน่นหน้าอก ปวดกระบอกตา ปวดศีรษะ ปวดจมูก ปวดหู เวียนศีรษะ เจ็บคอ เจ็บหน้าอก หายใจเหนื่อยหอบ/หายใจลำบาก หายใจมีเสียงหวีดๆ หายใจมีกลิ่นเหม็น ไอ และอาเจียน).

การวิเคราะห์ข้อมูลเบื้องต้น

ศึกษาลักษณะข้อมูลผู้ป่วยด้วยค่าสถิติ จำนวน ค่าร้อยละ ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน.

การประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจ

วิเคราะห์ตัวแปรทั้งหมดด้วยเทคนิคการจำแนกข้อมูลโดยใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ ๓ วิธี คือ ID3 C4.5 และ CART ด้วยขั้นตอนดังนี้

ก. การคัดเลือกตัวแปร

ใช้โปรแกรม Weka โดยเลือกใช้ขั้นตอนวิธีคัดเลือกตัวแปร ๓ วิธี ได้แก่

๑. วิธี Best First เป็นการคัดเลือกตัวแปรอิสระซึ่งอาจใช้: ๑.๑. การคัดเลือกเพิ่ม (forward) คือคัดเลือกตัวแปรอิสระเข้าทีละตัว; ๑.๒. การคัดเลือกลด (backward) โดยนำตัวแปรอิสระทุกตัวเข้าสมการก่อน แล้วจึงคัดเลือกออกทีละตัว; หรือ ๑.๓. ทำรายการตัวแปรแล้วเลือกตัวแปร ณ ตำแหน่งตรงกลาง และค้นหาตัวแปรถัดไปที่อยู่ทางซ้ายและทางขวา เพื่อค้นหาตัวแปรถัดไปที่จะถูกคัดเข้า และในเวลาเดียวกัน ตัวแปรนี้ก็อาจจะถูกพิจารณาคัดออกในขั้นตอนของการคัดเลือกถัดไปก็ได้[๑๐].

๒. วิธี Greedy Stepwise เป็นการคัดเลือกตัวแปรอิสระ ที่ใช้หลักการคล้ายกับวิธี Stepwise ซึ่งเป็นการคัดเลือกตัวแปรทีละขั้นตอน โดยตอนแรกจะคัดเลือกตัวแปรอิสระเข้าสมการ โดยวิธีคัดเลือกเพิ่ม. จากนั้นจึงทำการเลือกตัวแปรอิสระที่ผ่านเกณฑ์ที่กำหนดไว้เข้าสมการพร้อมทั้งพิจารณาว่ามีตัวแปรอิสระใดอยู่ในตัวแบบก่อนหน้าที่ควรจะถูกคัดออกหรือไม่ ทำเช่นนี้จนกระทั่งไม่สามารถเลือกตัวแปรอิสระเข้าตัวแบบและไม่สามารถคัดตัวแปรอิสระออกจากตัวแบบได้[๑๐].

๓. วิธี Genetic Search เป็นวิธีการคัดเลือกตัวแปรอิสระ โดยใช้ขั้นตอนวิธีพันธุศาสตร์ (genetic algorithm) อย่างง่าย คือเป็นกระบวนการค้นหาคำตอบที่มีพื้นฐานอยู่บนทฤษฎีทางพันธุศาสตร์ และการคัดเลือกตามธรรมชาติ. ระเบียบวิธีเลียนแบบพันธุศาสตร์มีกระบวนการหลัก ๓ ส่วน คือ การคัดเลือก, การไขว้สายพันธุ์ และการผ่าเหล่า ทำงานร่วมกันทำให้ระเบียบวิธีนี้เป็นการหาคำตอบที่มีประสิทธิภาพสูง[๑๑].

ใช้ทั้ง ๓ ขั้นตอนวิธีทำการคัดเลือกตัวแปรจากชุดข้อมูลฝึกสอนที่ได้จัดเตรียมไว้ ๒ แบบ คือ ๑) การใช้ข้อมูลหมดทั้งชุด (Use full training set) เป็นชุดข้อมูลฝึกสอน; และ ๒) การแบ่งข้อมูลเป็น ๑๐ ส่วน (10-fold cross-validation)

โดยใช้ข้อมูล ๙ ส่วนเป็นชุดข้อมูลฝึกสอนและส่วนที่เหลือเป็นชุดข้อมูลทดสอบ ทำการเปลี่ยนชุดข้อมูลทดสอบจนครบทุกส่วน.

ข. การสร้างตัวจำแนก

โดยใช้โปรแกรม Weka สำหรับขั้นตอนวิธี ID3 และ C4.5 และใช้โปรแกรม SPSS สำหรับขั้นตอนวิธี CART กับชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบที่มาจากการแบ่งข้อมูล ๒ แบบ.

แบบที่ ๑ ใช้การแบ่งข้อมูลแบบการแยกค่าร้อยละ (percentage split) โดยกำหนดอัตราส่วนระหว่างชุดข้อมูลฝึกสอนกับชุดข้อมูลทดสอบ ดังนี้

- แยกร้อยละ ๕๐ หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = ๕๐% : ๕๐%
- แยกร้อยละ ๖๐ หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = ๖๐% : ๔๐%
- แยกร้อยละ ๗๐ หมายถึง ชุดข้อมูลฝึกสอน : ชุดข้อมูลทดสอบ = ๗๐% : ๓๐%

แบบที่ ๒ ใช้การแบ่งข้อมูล แบบ k- fold cross-validation โดยแบ่งชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ ดังนี้

- 5 - fold cross-validation
- 10 - fold cross-validation
- 15 - fold cross-validation

ค. การทดสอบตัวจำแนก

เพื่อทดสอบความถูกต้องของตัวจำแนกที่ได้จากขั้นตอนการสร้างตัวจำแนกกับชุดข้อมูลฝึกสอน จึงนำตัวจำแนกดังกล่าวมาใช้กับชุดข้อมูลทดสอบ เพื่อให้ระบบทำนายและพิจารณาเปรียบเทียบประสิทธิภาพของตัวจำแนกที่ได้จากแต่ละขั้นตอนวิธีต้นไม้ตัดสินใจ ด้วยค่าวัดประสิทธิภาพ ๕ ค่า คือ ค่าความถูกต้อง, ค่าความไว, ค่าพยากรณ์บวก (PPV), ค่าพยากรณ์ลบ (NPV) และค่าพื้นที่ใต้ ROC curve (AUC).

ผลการศึกษา

ลักษณะทั่วไป

ผู้ป่วยนอกที่เข้ารับการรักษาครั้งแรกในกลุ่มโรคทางหายใจจำนวน ๗,๓๒๗ คน เป็นชายร้อยละ ๕๑.๔๐ และหญิงร้อยละ ๔๘.๖๐, อายุในกลุ่ม ๒๕ - ๕๙ ปี จำนวน ๓,๖๔๙ คน (ร้อยละ ๔๙.๘๐) และกลุ่มอายุ ๑๕ - ๒๔ปี จำนวน ๒,๔๘๕ คน (ร้อยละ ๓๓.๙๐), ผู้ป่วยร้อยละ ๖๑.๐๐ อาศัยอยู่นอกเขตเทศบาล ที่เหลืออาศัยอยู่ในเขตเทศบาล และอื่นๆ, ร้อยละ ๔๒.๙๐ ประกอบอาชีพรับจ้าง และร้อยละ ๒๗.๑๐ เป็นนักเรียน/นักศึกษา. ผู้ป่วยร้อยละ ๔๗.๙๐ เป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลันหลายแห่งพร้อมกัน, ร้อยละ ๒๙.๑๐ เป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และที่เหลือเป็นโรคปอดอักเสบ (ตารางที่ ๑).

การเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีต้นไม้ตัดสินใจ ๓ วิธี

ได้แก่ ID3 C4.5 และ CART ในการจำแนกหรือคัดกรองผู้ป่วย ๓ โรค คือ โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน, โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และโรคปอดอักเสบ.

ผลการคัดเลือกตัวแปรจาก ๓ ขั้นตอนวิธี ได้แก่ Best First, Greedy Stepwise และ Genetic Search ที่ได้จากการเตรียมชุดข้อมูลฝึกสอน ๒ แบบ คือ แบบใช้ชุดข้อมูลทั้งหมดและแบบการทำให้ถูกต้องไขว้ ๑๐ เท่า (10 fold cross-validation) ได้จำนวนตัวแปร จำแนกตามโรคและการแบ่งชุดข้อมูลดังตารางที่ ๒.

ผลการคัดเลือกตัวแปร พบว่าในแต่ละโรคได้จำนวนตัวแปรที่ใกล้เคียงกัน. ขั้นต่อไปจะเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีต้นไม้ตัดสินใจ โดยใช้ตัวแปรที่คัดเลือกได้จากแต่ละโรคดังนี้

โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน ใช้ ๗ ตัวแปร คือ มีไข้, ปวดศีรษะ, เวียนศีรษะ, เจ็บคอ, อาเจียน, หายใจเหนื่อยหอบ/หายใจลำบาก, ปวดจมูก และปวดหู.

โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน ใช้ ๗ ตัวแปร คือ มีไข้, เวียนศีรษะ, หายใจมีเสียงหวีด, หายใจมีกลิ่นเหม็น, ปวดกระบอกตา, ปวดจมูก และปวดหู.

สำหรับโรคปอดอักเสบ ใช้ ๘ ตัวแปร คือ เจ็บคอ, อาเจียน, หายใจเหนื่อยหอบ/หายใจลำบาก, หายใจมีเสียงหวีด, แน่นหน้าอก, เจ็บหน้าอก, ปวดจมูก และปวดหู.

ตารางที่ ๑ ข้อมูลพื้นฐานของผู้ป่วยโรคระบบการหายใจ ที่เข้ารับการรักษาครั้งแรกที่โรงพยาบาลพระนครศรีอยุธยา ช่วง พ.ศ. ๒๕๔๗-๒๕๔๙ จำนวน ๗,๓๒๗ ราย

ตัวแปร	ราย	ร้อยละ
เพศ		
ชาย	๓,๗๖๖	๕๑.๔๐
หญิง	๓,๕๖๑	๔๘.๖๐
อายุ (ปี)		
น้อยกว่า ๑ ปี	๒๒๔	๓.๑๐
๑ - ๑๔	๗๕๘	๑๐.๕๐
๑๕ - ๒๔	๒,๔๘๕	๓๓.๕๐
๒๕ - ๕๕	๓,๖๔๕	๔๕.๘๐
๖๐ ขึ้นไป	๑๗๑	๒.๓๐
ค่าเฉลี่ย +/- ค่าเบี่ยงเบนมาตรฐาน = ๒๖.๓๗±๑๒.๘๕		
ที่อยู่อาศัย		
ในเขตเทศบาล	๒,๓๘๒	๓๒.๕๐
นอกเขตเทศบาล	๔,๔๖๖	๖๑.๐๐
อื่น ๆ	๔๗๕	๖.๕๐
อาชีพ		
ในความปกครอง	๔๐๘	๕.๖๐
นักเรียน/นักศึกษา	๑,๕๔๘	๒๗.๑๐
ข้าราชการ	๔๕๐	๖.๗๐
พนักงานรัฐวิสาหกิจ	๒๕๕	๔.๐๐
พนักงานบริษัทเอกชน	๓๕๕	๔.๘๐
รับจ้าง	๓,๑๔๐	๔๒.๕๐
ธุรกิจส่วนตัว	๒๘๖	๓.๕๐
กสิกรรม	๒๓๒	๓.๒๐
อื่น ๆ	๑๓๗	๑.๕๐
โรคทางหายใจ		
โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน	๓,๕๑๓	๔๗.๕๐
โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน	๒,๑๓๒	๒๕.๑๐
โรคปอดอักเสบ	๑,๖๘๒	๒๓.๐๐

ตัวแปรที่ผ่านการคัดเลือกถูกนำมาพัฒนาตัวจำแนกผู้ป่วยรายโรคโดยใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ ID3 C4.5 และ CART กับชุดข้อมูลเรียนรู้ที่มาจากการจัดเตรียม ๖ แบบ คือ การทำให้ถูกต้องไขว้ ๕ เท่า, ๑๐ เท่า, ๑๕ เท่า, การแยกค่าร้อยละ ๕๐, ๖๐ และ ๗๐ และทำการเปรียบเทียบประสิทธิภาพของทั้ง ๓ ขั้นตอนวิธี ดังตารางที่ ๓.

จากตารางที่ ๓ พบว่า การจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ ด้วยขั้นตอนวิธี C4.5 กับข้อมูลที่มีการแบ่งแบบแยกค่าร้อยละ ๗๐ จะให้ประสิทธิภาพสูงสุด ส่วนการจำแนกผู้ป่วยโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน ด้วยขั้นตอนวิธี CART กับข้อมูลที่มีการแบ่งแบบการแยกค่าร้อยละ ๕๐ จะให้ประสิทธิภาพสูงสุด.

การประยุกต์การใช้ขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ

ผลจากการประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจ ทำให้ได้แผนภาพต้นไม้ตัดสินใจ สำหรับจำแนกผู้ป่วยทั้ง ๓ โรค ดังนี้

๑. โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน.

ด้วยขั้นตอนวิธี C4.5.

สรุปเป็นกฎการจำแนกข้อมูลผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลันที่สำคัญได้ดังนี้

๑. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีโอกาสเป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๒๐๙๐.๐/๑๐๙.๐).

๒. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ แต่มีไข้ มีโอกาสเป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๑๓๔๑.๐/๒๐๒.๐).

๓. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน ไม่เจ็บคอ ไม่มีไข้ แต่มีอาการเวียนศีรษะ มีโอกาสเป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๖๗.๐/๗.๐).

๔. ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู มีเวียนศีรษะ มีโอกาสเป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๓๕.๐/๓.๐).

ตัวอย่าง ผู้ที่มีอาการเหนื่อยหอบจะไม่มีโอกาสเป็นโรค

ตารางที่ ๒ จำนวนตัวแปรที่คัดเลือกได้จากขั้นตอนวิธี Best First, Greedy Stepwise และ Genetic Search จำแนกตามโรคและการแบ่งชุดข้อมูล

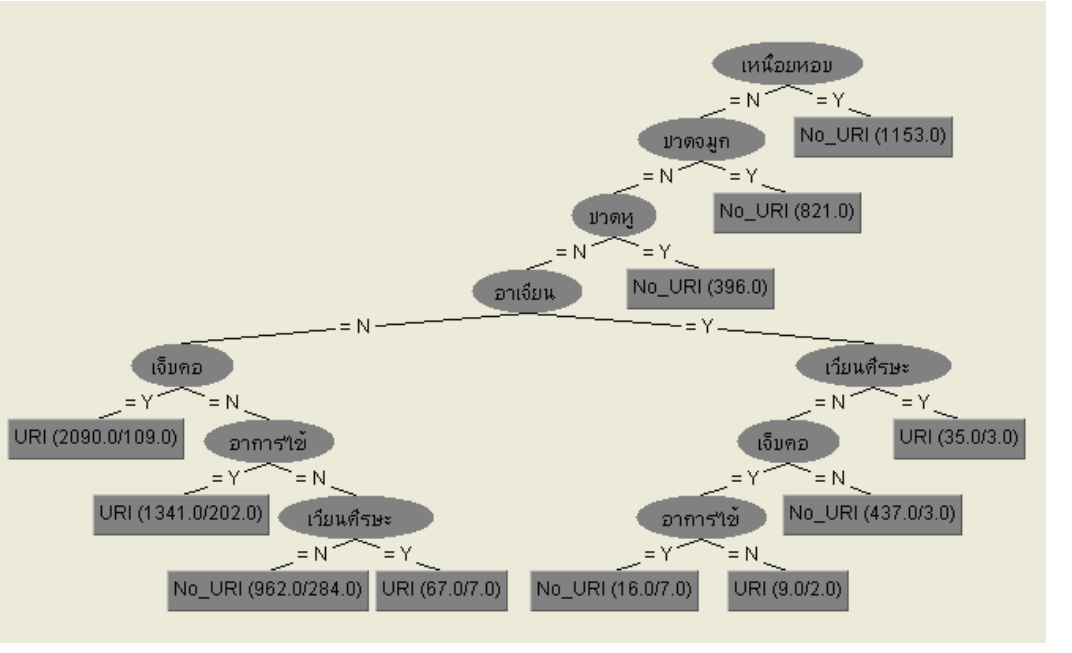
โรคระบบการหายใจ	Best First		Greedy Stepwise		Genetic Search	
	Use full training set	Cross-validation	Use full training set	Cross-validation	Use full training set	Cross-validation
โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน	๙	๙	๙	๙	๑๐	๗
โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน	๘	๗	๘	๗	๘	๗
โรคปอดอักเสบ	๙	๘	๙	๘	๙	๘

ตารางที่ ๓ เปรียบเทียบขั้นตอนวิธี ID3, C4.5 และ CART ในการจำแนกผู้ป่วยกลุ่มโรคระบบการหายใจ ๓ โรค ตามการแบ่งชุดข้อมูลแบบต่างๆ

(ค่าร้อยละ)

กลุ่มโรคระบบการหายใจ	ความถูกต้อง	ความไว	พยากรณ์บวก	พยากรณ์ลบ	AUC
โรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน					
ID3# แยกร้อยละ ๗๐	๙๒.๓๑	๙๐.๖๐	๙๒.๘๕	๙๑.๘๖	๙๖.๓๐
C4.5# แยกร้อยละ ๗๐	๙๒.๓๒	๙๒.๔๐	๙๕.๖๓	๙๐.๑๘	๙๖.๒๐
CART# แยกร้อยละ ๗๐	๙๐.๖๗	๘๙.๖๘	๙๐.๒๙	๙๐.๙๙	๙๖.๒๐
โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน					
ID3§ ถูกต้องไขว้	๙๔.๓๖	๑๐๐	๘๐.๖๓	๑๐๐	๙๖.๑๐
C4.5§ ถูกต้องไขว้	๙๔.๓๖	๑๐๐	๘๐.๖๓	๑๐๐	๘๙.๗๐
CART§ แยกร้อยละ ๕๐	๙๔.๖๙	๑๐๐	๘๑.๕๗	๑๐๐	๙๔.๗๐
โรคปอดอักเสบ					
ID3¶ แยกร้อยละ ๗๐	๙๔.๖๘	๙๓.๒๓	๘๓.๒๔	๙๘.๑๖	๙๖.๕๐
C4.5¶ แยกร้อยละ ๗๐	๙๔.๗๐	๙๓.๔๒	๘๓.๒๔	๙๘.๑๖	๙๒.๒๐
CART¶ ถูกต้องไขว้	๙๓.๖๙	๙๒.๐๗	๗๙.๓๗	๙๗.๙๖	๙๕.๙๐

หมายเหตุ
- **ID3# = ขั้นตอนวิธี ID3 แบ่งชุดข้อมูลแบบแยกร้อยละ ๗๐**
- **C4.5# = ขั้นตอนวิธี C4.5 แบ่งชุดข้อมูลแบบแยกร้อยละ ๗๐**
- **CART# = ขั้นตอนวิธี CART แบ่งชุดข้อมูลแบบแยกร้อยละ ๗๐**
- **ID3§ = ขั้นตอนวิธี ID3 แบ่งชุดข้อมูลแบบการทำให้ถูกต้องไขว้ ๕, ๑๐ และ ๑๕ เท่า. การทำให้ถูกต้องไขว้ซึ่งได้ค่าวัดประสิทธิภาพเท่ากัน**
- **C4.5§ = ขั้นตอนวิธี C4.5 แบ่งชุดข้อมูลแบบการทำให้ถูกต้องไขว้ ๕, ๑๐ และ ๑๕ เท่า การทำให้ถูกต้องไขว้ ซึ่งได้ค่าวัดประสิทธิภาพเท่ากัน**
- **CART§ = ขั้นตอนวิธี CART แบ่งชุดข้อมูลแบบแยกร้อยละ ๕๐**
- **ID3¶ = ขั้นตอนวิธี ID3 แบ่งชุดข้อมูลแบบแยกร้อยละ ๗๐**
- **C4.5¶ = ขั้นตอนวิธี C4.5 แบ่งชุดข้อมูลแบบแยกร้อยละ ๗๐**
- **CART¶ = ขั้นตอนวิธี CART แบ่งชุดข้อมูลแบบการทำให้ถูกต้องไขว้ ๕, ๑๐ และ ๑๕ เท่า การทำให้ถูกต้องไขว้ ซึ่งได้ค่าวัดประสิทธิภาพเท่ากัน**



รูปที่ ๑ ต้นไม้ตัดสินใจในการจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน ด้วยขั้นตอนวิธี C4.5

ติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๑๑๕๓.๐) หมายถึง ผู้ที่มีอาการเหนื่อยหอบ มีทั้งหมด ๑,๑๕๓ คน และทั้ง ๑,๑๕๓ คน ไม่เป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน.

หรืออีกตัวอย่างหนึ่ง ผู้ที่ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีโอกาสเป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน (๒๐๙๐.๐/๑๐๙.๐) หมายถึง มีผู้ที่ไม่มีอาการหายใจเหนื่อยหอบ ไม่ปวดจมูก ไม่ปวดหู ไม่อาเจียน แต่มีอาการเจ็บคอ มีทั้งหมด ๒,๐๙๐ คน แต่เป็นผู้ที่เป็นโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลันทั้งสิ้น ๑๐๙ คน เป็นต้น.

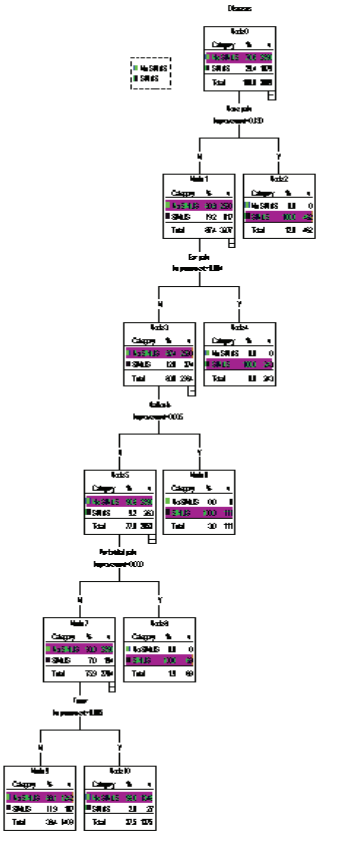
๓.๒ โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน.

สรุปเป็นกฎการจำแนกผู้ป่วยโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน ที่สำคัญได้ดังนี้

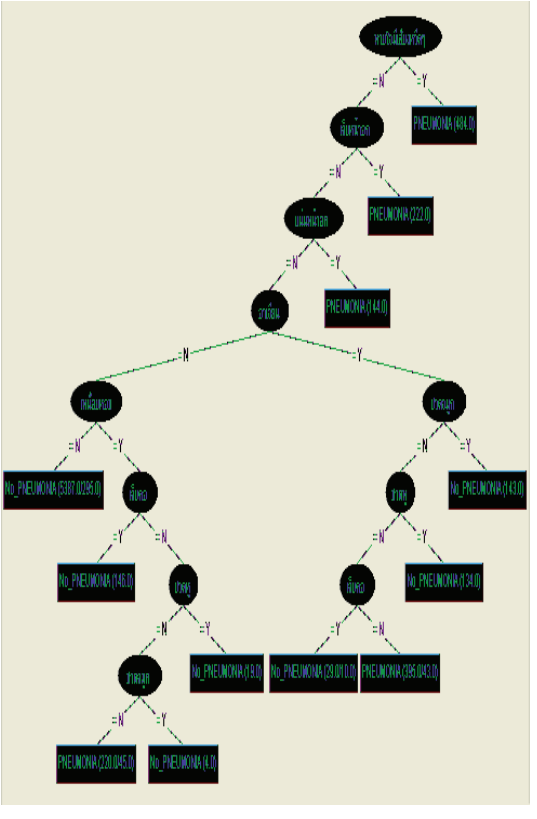
๑. มีอาการปวดจมูก มีโอกาสเป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน (๔๐๒.๐).

๒. ไม่มีอาการปวดจมูก แต่มีอาการปวดหู มีโอกาสเป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน (๒๔๐.๐).

๓. ไม่มีอาการปวดจมูก ไม่ปวดหู แต่มีอาการหายใจมีกลิ่นเหม็น มีโอกาสเป็นโรคโพรงอากาศข้างจมูก



รูปที่ ๒ ต้นไม้การจำแนกสำหรับผู้ป่วยโรคไซนัสอักเสบเฉียบพลัน ด้วยขั้นตอนวิธี CART



รูปที่ ๓ ต้นไม้ตัดสินใจสำหรับการจำแนกผู้ป่วยโรคปอดอักเสบด้วยขั้นตอนวิธี C4.5

อักเสบเฉียบพลัน (๑๑๑.๐).

๔. ไม่มีอาการปวดจมูก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น แต่มีอาการปวดกระบอกตามีโอกาสเป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน (๘๙.๐).

๕. ไม่มีอาการปวดจมูก ไม่ปวดหู ไม่หายใจมีกลิ่นเหม็น ไม่ปวดกระบอกตา ไม่มีอาการปวดกระบอกตา แต่มีไข้มีโอกาสเป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน (๑๓๗๕.๐/๒๗.๐).

๓.๓ โรคปอดอักเสบ.

สรุปเป็นกฎการจำแนกผู้ป่วยโรคปอดอักเสบที่สำคัญได้ดังนี้

๑. มีอาการหายใจมีเสียงหวีด มีโอกาสเป็นโรคปอดอักเสบ (๔๘๔.๐).

๒. ไม่มีอาการหายใจมีเสียงหวีด แต่มีอาการเจ็บหน้าอก มีโอกาสเป็นโรคปอดอักเสบ (๒๒๒.๐).

๓. ไม่มีอาการหายใจมีเสียงหวีด ไม่เจ็บหน้าอก แต่มีอาการแน่นหน้าอก มีโอกาสเป็นโรคปอดอักเสบ (๑๔๔.๐).

๔. ไม่มีอาการหายใจมีเสียงหวีด ไม่เจ็บหน้าอก ไม่แน่นหน้าอก ไม่ปวดจมูก ไม่ปวดหู ไม่เจ็บคอ แต่มีอาการอาเจียน มีโอกาสเป็นโรคปอดอักเสบ (๓๙๕.๐/๔๓.๐).

วิจารณ์

เมื่อพิจารณาประสิทธิภาพของขั้นตอนต้นไม้ตัดสินใจที่ดีที่สุด ที่ใช้ในการจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน, โรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน และโรคปอดอักเสบ จะเห็นว่าขั้นตอนวิธี C4.5 ที่ใช้กับการแบ่งข้อมูลชุดฝึกสอนและชุดทดสอบในอัตรา ๗๐ : ๓๐ สำหรับจำแนกผู้ป่วยโรคติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ การใช้เพียง ๗ ลักษณะอาการก็สามารถจำแนกผู้ป่วยทั้ง ๒ โรค ได้ประสิทธิภาพสูงถึงร้อยละ

๙๒ และ ๙๕ ตามลำดับ ซึ่งเป็นการประหยัดคำถามในการซักประวัติผู้ป่วยสำหรับการคัดกรองผู้ป่วยเบื้องต้น. ในทำนองเดียวกัน การจำแนกผู้ป่วยโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลัน ใช้เพียง ๘ ลักษณะอาการ ก็สามารถใช้ขั้นตอนวิธี CART ในการช่วยจำแนกผู้ป่วยได้ถูกต้องถึงร้อยละ ๙๕ เนื่องจากกลุ่มโรคที่ผู้วิจัยศึกษาทั้ง ๓ โรค มีตัวแปรที่มีลักษณะเฉพาะของแต่ละโรคอย่างชัดเจน เช่น ตัวแปรอาการหายใจมีกลิ่นเหม็น จะพบเฉพาะในผู้ป่วยที่เป็นโรคโพรงอากาศข้างจมูกอักเสบเฉียบพลันเท่านั้นจึงทำให้ได้ประสิทธิภาพของการจำแนกมีค่าสูง.

จากผลการศึกษาที่ได้นี้ ควรให้โรงพยาบาลพระนครศรีอยุธยาได้นำผลลัพธ์ กฎ หรือเงื่อนไขการตัดสินใจที่ได้จากขั้นตอนวิธี C4.5 และ CART ไปใช้ควบคู่กับระบบการตรวจคัดกรองปัจจุบัน เพื่อเป็นการตรวจสอบความถูกต้องและประสิทธิภาพของขั้นตอนวิธี สำหรับโรงพยาบาลอื่นที่จะนำวิธีการนี้ไปใช้ ต้องจัดเตรียมข้อมูลให้ตรงตามข้อกำหนดและเงื่อนไขของแต่ละขั้นตอนวิธี และโรคที่สนใจศึกษาควรเป็นกลุ่มโรคที่มีความซุกสูง เพราะวิธีการต้นไม้ตัดสินใจเป็นเทคนิคการเรียนรู้ที่เหมาะสมกับการศึกษาข้อมูลขนาดใหญ่ ดังนั้นจึงไม่เหมาะสมที่จะศึกษากับกลุ่มโรคที่มีความซุกของโรคต่ำ.

กิตติกรรมประกาศ

ขอขอบคุณ ผู้อำนวยการโรงพยาบาลพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา ได้อนุเคราะห์ให้ข้อมูลในการศึกษาครั้งนี้.

เอกสารอ้างอิง

๑. สำนักระบาดวิทยา. สรุปรายงานการเฝ้าระวังโรคประจำปี พ.ศ ๒๕๔๗ - ๒๕๔๘. พิมพ์ครั้งที่ ๑. กรุงเทพมหานคร: โรงพิมพ์องค์การรับส่งสินค้าและพัสดุภัณฑ์; ๒๕๔๘.

๒. การนิคมอุตสาหกรรมแห่งประเทศไทย. รายงานนิคมอุตสาหกรรมในประเทศไทย [online]. [๒๗ มกราคม ๒๕๕๐] แหล่งข้อมูล: http://www.ieat.go.th/index_thtest.php?lang=en&lang=en&CLmIEAT2=12

๓. สำนักนโยบายและแผนงานสาธารณสุข สำนักงานปลัดกระทรวงสาธารณสุข. บัญชีจำแนกโรคระหว่างประเทศ ฉบับแก้ไข ครั้งที่ ๑๐ ไทย-อังกฤษ. นนทบุรี: สำนักงานปลัดกระทรวงสาธารณสุข; ๒๕๔๑

๔. Soman T. and Bobbie O. Patrick. Classification of arrhythmia using machine learning techniques [online]. 2005 [cited 2007 Jan 27]; Available from: URL: http://cse.spsu.edu/pbobbie/SharedFile/ECGDiagnosis_ICOSSE_2005_VFinal.pdf,

๕. Carlos O. Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine 2006;10:334-43.

๖. Aftarczuk K. Evaluation of selected data mining algorithms implemented in medical decision support systems (Master of Engineering). Department of Software Engineering: Swedish University; 2007.

๗. Jose PM, Aleman C, Bielsa S, Sarrapio J, de Sevilla TF, Esquerda A. A decision tree for differentiating tuberculous from malignant pleural effusions. Respir Med 2008; 102: p.1159-64.

๘. พุธษดี ศิริแสงตระกูล. ระบบจำแนกประเภทแบบทดสอบสำหรับผู้ทดสอบสุขภาพจิตด้วยเทคนิค Decision Tree (วิทยานิพนธ์ปริญญาวิทยาศาสตร์มหาบัณฑิต). ภาควิชาวิทยาการคอมพิวเตอร์, บัณฑิตวิทยาลัย, มหาวิทยาลัยขอนแก่น ขอนแก่น; ๒๕๕๑.

๙. วราภรณ์ พิมา. การวิเคราะห์ปัจจัยเสี่ยงและการจัดจำแนกกลุ่มของการดื้อยาในผู้ป่วยวัณโรคปอดที่เกิดโรคกลับ (วิทยานิพนธ์ปริญญาวิทยาศาสตร์มหาบัณฑิต). ภาควิชาสถิติ, บัณฑิตวิทยาลัย, มหาวิทยาลัยเชียงใหม่ เชียงใหม่; ๒๕๕๑.

๑๐. Witten HI, Eibe F. Data mining practical machine learning tools and tecniques. 2nd Ed. Los Angelis: Morgan Kaufmann; 2005.

๑๑. Melanie M. An introduction to genetic algorithms. The MIT Press: Massachusetts; 1998. (221)

๑๒. Ross QJ. Induction of decision trees. Machine learning. New York: McGraw-Hill; 1986. p. 81-106.

๑๓. Ross QJ. C4.5: programs for machine learning: Morgan Kaufmann; 1992.

๑๔. Breiman L, Stone JC, Olshen RA, Friedman J. Classification and regression trees. London: Chapman & Hall; 1984. (368).