การจำแนกผู้ที่มีความเสี่ยงต่อโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ กรณีศึกษา: โรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม

Classification of people at risk for breast cancer using decision tree algorithm, A case study of Suddhavej Hospital, Mahasarakham University

ชัยยันต์ สุขหมั่น[1*] และ สุภาวดี วิชิตชาญ[2]
Chaiyarn Sukmun[1*] and Supawadee Wichitchan[2]

Received: 16 August 2023 ; **Revised**: 19 September 2023 ; **Accepted**: 16 October 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกประเภทโรคมะเร็งเต้านม (breast cancer) และศึกษาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม ผู้วิจัยได้ใช้ข้อมูลเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านมจากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 จากการทำความสะอาดข้อมูลเหลือข้อมูลทั้งหมด 1,524 ระเบียน ซึ่งมีข้อมูลผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม จำนวน 1,343 ระเบียน และข้อมูลผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม จำนวน 181 ระเบียน จากผลการศึกษาพบว่าต้นไม้ตัดสินใจ C4.5, C5.0 และ Random forest ให้ค่าความถูกต้อง (accuracy) ค่อนข้างสูง แต่ค่าเกณฑ์ในการทำนาย AUC (area under ROC curve) ค่อนข้างต่ำ เนื่องจากการทำนายโมเดลไม่สามารถแยกกลุ่ม (class) ได้ดีพอ ซึ่งพบว่าข้อมูลที่ใช้ในการจำแนกคลาสมีจำนวนของคลาสมากน้อยไม่เท่ากัน (class imbalance) เพื่อแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้ใช้เทคนิคการสุ่มเพิ่ม (oversampling) เพื่อเพิ่มจำนวนตัวอย่างในคลาสที่น้อยเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน และวิธีสุ่มลด (undersampling) ลดตัวอย่างในคลาสที่มีจำนวนมากลงเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน พบว่าต้นไม้ตัดสินใจ C4.5 และ C5.0 ให้ผลลัพธ์ไม่ต่างจากเดิมและผลลัพธ์ที่ได้ไม่ต่างกันมากนัก ส่วน Random forest ให้ค่า AUC และค่าความระลึก (recall) ที่ดีขึ้นเมื่อเปรียบเทียบกับต้นไม้ตัดสินใจ C4.5 และ C5.0 ซึ่งสูงกว่าประมาณ 15-20%

คำสำคัญ: มะเร็งเต้านม, ต้นไม้ตัดสินใจ, ข้อมูลไม่สมดุล

Abstract

This research focused on evaluating the effectiveness of the Decision Tree Algorithm in classifying breast cancer, as well as investigating the associated risk factors. The study employed medical record data from breast mass patients at Mahasarakham University's Faculty of Medicine, spanning 2010 to 2022. The dataset, post-cleansing, comprised 1,524 records, with 1,343 representing low-risk breast cancer patients and 181 representing high-risk cases. The study indicates that the Decision Tree Algorithms, specifically C4.5, C5.0, and Random Forest, had substantial classification accuracy. However, their area under the ROC curve (AUC) values were relatively low due to insufficient class separation, which stems from class imbalance. This issue was addressed by employing oversampling to augment the minority class instances and undersampling to reduce the majority class instances. The outcomes revealed that both C4.5 and C5.0 Decision Trees yielded comparable results, while Random Forest demonstrated a superior AUC and recall, approximately 15-20% higher than C4.5 and C5.0.

Keywords: Breast cancer, decision tree, class-imbalance

---

[1] สาขาวิทยาการจัดการสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม ประเทศไทย
[2] อาจารย์, คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม ประเทศไทย
[1] Department of Statistical Management Science, Faculty of Science, Mahasarakham University, Maha Sarakham, Thailand
[2] Lecturer, Faculty of Science, Mahasarakham University, Maha Sarakham, Thailand
* Corresponding author E-mail: 63010257003@msu.ac.th

บทนำ

มะเร็ง คือ กลุ่มของโรคที่เกิดเนื่องจากเซลล์ของร่างกายมีความผิดปกติ ที่ DNA หรือสารพันธุกรรม ส่งผลให้เซลล์มีการเจริญเติบโตมีการแบ่งตัวเพื่อเพิ่มจำนวนเซลล์รวดเร็วและมากกว่าปกติ ดังนั้น จึงอาจทำให้เกิดก้อนเนื้อผิดปกติ และในที่สุดก็จะทำให้เกิดการตายของเซลล์ในก้อนเนื้อนั้น เนื่องจากขาดเลือดไปเลี้ยงถ้าเซลล์พวกนี้เกิดอยู่ในอวัยวะใดก็จะเรียกชื่อ “มะเร็ง” ตามอวัยวะนั้น เช่น มะเร็งปอด, มะเร็งสมอง, มะเร็งเต้านม, มะเร็งปากมดลูก, มะเร็งเม็ดเลือดขาว และมะเร็งผิวหนัง เป็นต้น (สถาบันมะเร็งแห่งชาติ, 2563) “มะเร็งเต้านม” เป็นมะเร็งที่พบมากที่สุดเป็นอันดับ 1 ของผู้หญิงไทย และเป็นสาเหตุของการเสียชีวิตอันดับต้น ๆ ในผู้หญิงทั่วโลก แนวโน้มคนไทยป่วยเป็นโรคมะเร็งสูงขึ้นทุกปีแต่ยังพบน้อยกว่าประเทศทางตะวันตกมาก โดยผู้หญิงไทยมีอัตราการพบมะเร็งประมาณ 40 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ซึ่งถ้าเทียบกับประเทศตะวันตกพบมะเร็งเต้านมได้มากกว่า 100 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ส่วนในผู้ชายก็พบมะเร็งเต้านมได้เช่นกันแต่ไม่บ่อยนัก โดยมีอุบัติการณ์ของโรคนี้น้อยกว่าผู้หญิงเกือบ 100 เท่า

สถานการณ์ของโรคมะเร็งในภาพรวมของประเทศไทย จากสถิติพบว่าโรคมะเร็งเป็นสาเหตุการเสียชีวิตอันดับ 1 คิดเป็นร้อยละ 16 ของต้นเหตุการเสียชีวิตทั้งหมดสูงกว่าอันตราการเสียชีวิตจากอุบัติเหตุ และโรคหัวใจเฉลี่ย 2 ถึง 3 เท่า หรือมีผู้เสียชีวิตจากโรคมะเร็งเฉลี่ย 8 รายต่อชั่วโมง ปัจจุบันโรคมะเร็งถือเป็นปัญหาสาธารณสุขที่สำคัญของประเทศไทยและมีแนวโน้มอัตราการเกิดโรคสูงขึ้นอย่างต่อเนื่อง จากสถิติพบว่ามีผู้ป่วยโรคมะเร็งรายใหม่ 139,206 คนต่อปี และในจำนวนนี้มีผู้เสียชีวิต 84,073 คนต่อปี สำหรับ 5 อันดับแรกของมะเร็งที่พบบ่อยที่สุด ได้แก่ 1. มะเร็งตับและท่อน้ำดี 2. มะเร็งเต้านม 3. มะเร็งปอด 4. มะเร็งลำไส้ใหญ่และทวารหนัก และ 5. มะเร็งปากมดลูก (ณัฏฐพร นันทิวัฒนา, 2563)

การทำเหมืองข้อมูล (data mining) คือ การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (knowledge discovery from very large databases: KDD) (สายชล สินสมบูรณ์ทอง, 2560) หรือที่เรียกกันว่าการทำเหมืองข้อมูล เป็นวิธีการที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์ หรือทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้นซึ่งการค้นหาความรู้และความจริงที่แฝงอยู่ในข้อมูล (knowledge discovery) เป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่มีอยู่ เพื่อให้ได้สารสนเทศที่มีประโยชน์ (useful Information) ที่เรายังไม่ทราบ (unknown data) โดยเป็นสารสนเทศที่มีเหตุผล (valid information) และสามารถนำไปใช้ได้ (actionable) ซึ่งเป็นสิ่งสำคัญที่จะช่วยการตัดสินใจในการทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ ซึ่งมีนักวิจัยหลายคนได้นำกระบวนการทำเมืองข้อมูลมาประยุกต์ใช้ในการพยากรณ์ในด้านต่าง ๆ อาทิ อุกฤษฏ์ ศรีสุข และจารี ทองคำ (2564) ทำการการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูล สำหรับพยากรณ์การเกิดโรค ซึ่งรวบรวมมาจากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล โดยนำเอาเทคนิคการเรียนรู้ของเครื่อง (machine learning) มาใช้กับการทำเหมืองข้อมูล 5 เทคนิค ได้แก่ Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest, Deep Learning มาทำการสร้างแบบจำลอง เพื่อการพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ จากการทดลองพบว่า เทคนิค Decision Tree C4.5 เป็นเทคนิคที่ดีในการสร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์ และโรคมะเร็งเต้านม โดยให้ค่าความถูกต้อง 99.86% และ 75.52% ตามลำดับ และเทคนิค Deep Learning เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคเบาหวาน โดยให้ค่าความถูกต้อง 77.47% รวมถึงงานวิจัยของ Nemade & Fegade (2023) ใช้เทคนิคการเรียนรู้ของเครื่องในการทำนายมะเร็งเต้านม ได้แก่ Naïve Bayes, Logistic Regression, ซับพอร์ต เวกเตอร์ แมชชีน (support vector machine), วิธีการเพื่อนบ้านใกล้ที่สุด (k-nearest neighbors), ต้นไม้ตัดสินใจ (decision tree), Random forest, อดาบูสท์ (adaboost) และเอ็กซ์จีบูสท์ (XGBoost) ถูกนำมาใช้กับชุดข้อมูลมะเร็งเต้านม พบว่าต้นไม้ตัดสินใจ (decision Tree) ให้มีค่าความถูกต้องสูงสุดถึง 97%

ดังนั้น งานวิจัยนี้ผู้วิจัยมีแนวคิดในการนำเอาเทคนิคการทำเหมืองข้อมูลต้นไม้ต้นสินใจ (decision tree) มาใช้ในการจำแนกการเป็นโรคมะเร็งเต้านม ซึ่งเป็นโรคที่ยังไม่ทราบสาเหตุที่แน่ชัด และเป็นกันมากทั่วโลกรวมถึงในประเทศไทย งานวิจัยนี้จะเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกการเป็นโรคมะเร็งเต้านม โดยใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และ Random forest เพื่อหาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม โดยการแบ่งกลุ่มข้อมูลเป็นชุดการเรียนรู้ (training data) และชุดข้อมูลทดสอบ (testing data) และวัดประสิทธิภาพโมเดลของแบบจำลองในแง่ของค่าความถูกต้อง (accuracy), เกณฑ์ในการทำนาย AUC (area under ROC curve) และค่าความระลึก (recall) ผลการวิจัยจะสามารถนำไปใช้สำหรับวางแผนการรักษาหรือให้คำแนะนำผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และเมื่อหาแบบจำลอง (model) ที่มีความน่าเชื่อถือได้ก็สามารถที่จะนำข้อมูลผู้ที่เข้ารับการตรวจมะเร็งเต้านมมาเข้าในแบบจำลองเพื่อจะดูว่าผู้ที่เข้ารับการตรวจมีโอกาสที่จะเป็นโรคมะเร็งเต้านมหรือไม่ และถ้าผู้ป่วยอยู่ในกลุ่มที่เป็นมะเร็งเต้านมทางแพทย์จะได้ทำการตรวจอย่างละเอียดมากยิ่งขึ้น

เพื่อที่จะได้ทำการรักษาได้ทันท่วงที อีกทั้งยังประชาสัมพันธ์ถึงสาธารณะชนเพื่อให้ความรู้เกี่ยวกับโรคมะเร็งเต้านม และบุคคลที่มีปัจจัยเสี่ยงในการเป็นโรคมะเร็งเต้านมให้เข้ารับการตรวจโดยแพทย์ผู้เชี่ยวชาญได้อย่างทันการ

วิธีดำเนินการวิจัย

การวิจัยเรื่องการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านม มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกประเภทโรคมะเร็งเต้านม (breast cancer) และศึกษาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม (breast cancer)

1. วิธีการเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลทำโดยการสังเคราะห์ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 มีตัวแปรอิสระ (independent variable) ทั้งหมด 10 ตัวแปร โดยเป็นตัวแปรเชิงคุณภาพทั้งหมดดัง Table 1 พบว่าข้อมูลทั้งหมดมีค่าที่ขาดหายไป (missing value) 1.60 % จึงต้องทำการลบข้อมูลที่ขาดหายไปออกจากข้อมูลทั้งหมด ซึ่งจะทำให้เหลือข้อมูลทั้งหมด 1,524 ระเบียน โดยผู้วิจัยได้ทำการแปลงข้อมูลอายุและดัชนีมวลกายจากตัวแปรเชิงปริมาณเป็นตัวแปรเชิงคุณภาพ เนื่องจากลดความซับซ้อนของอัลกอริทึมต้นไม้ตัดสินใจทำให้โมเดลง่ายต่อการทำนาย และลดโอกาสเกิดการเรียนรู้มากเกินไป (overfitting) รวมทั้งช่วยในการดูแนวโน้มของข้อมูลได้ง่ายขึ้น

ตัวแปรตาม (dependent variable) คือ ผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม

ผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (88.12%) หมายถึง ผู้ที่เข้ารับการตรวจแมมโมแกรมโดยมีค่า BIRADs Score คิดจากลักษณะรูปภาพที่เห็นซึ่งอยู่ในระดับคะแนน 0-3 (negative class)

ผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (11.88%) หมายถึง ผู้ที่เข้ารับการตรวจแมมโมแกรมโดยมีค่า BIRADs Score คิดจากลักษณะรูปภาพที่เห็นซึ่งอยู่ในระดับคะแนน 4-6 (positive class)

Table 1 Independent Variable

ตัวแปร	กลุ่มตัวแปร (สัดส่วนข้อมูล)
1. เพศ (sex)	ชาย (0.79%), หญิง (99.21%)
2. อายุ (age)	อายุ 20-32 ปี (50.20%), อายุ 33-45 ปี (25.39%), อายุ 46-58 ปี (21.00%), อายุ 59-71 ปี (1.97%), อายุ 72-84 ปี (1.25%), อายุ 85-97 ปี (0.20%)
3. ดัชนีมวลกาย (body mass index: BMI)	<18.5 (3.54%), 18.5-22.9 (39.44%), 23.0-24.9 (20.28%), 25.0-29.9 (28.41%), ≥30 (8.33%)
4. สูบบุหรี่ (smoking)	สูบ (0.46%), ไม่สูบ (91.21%), ไม่ทราบ (8.33%)
5. ดื่มสุรา (binge)	ดื่ม (1.05%), ไม่ดื่ม (90.68%), ไม่ทราบ (8.27%)
6. อาการที่นำมาพบแพทย์ (chief complaint)	มีอาการ (30.97%), ไม่มีอาการ (10.37%), มาตามนัด (58.66%)
7. ก้อนหรือถุงน้ำ (mass or cyst)	Yes (48.82%), No (51.18%)
8. ความสมมาตรของเต้านมสองข้าง (asymmetries)	Yes (72.57%), No (27.43%)
9. แคลเซียม (calcification)	Yes (50.72%), No (49.28%)
10. โครงสร้างเต้านม (architectural distortion)	Yes (9.78%), No (90.22%)

2. วิธีวิเคราะห์ข้อมูล

2.1 อัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูล

1. อัลกอริทึมต้นไม้ตัดสินใจ C4.5

วิธีการวิเคราะห์ต้นไม้ตัดสินใจแบบ C4.5 เป็นวิธีที่ใช้ในการสร้างและประเมินต้นไม้การตัดสินใจ (decision tree) ที่พัฒนาโดย Ross Quinlan ในปี 1986 ซึ่งเป็นวิธีที่มีการใช้ค่า Information Gain ในการเลือก attribute ในการแยกกลุ่มข้อมูล ซึ่งเป็นค่าที่บ่งบอกถึงความสำคัญของ attribute ในการลดความไม่แน่นอนของข้อมูล (entropy) ในกลุ่มย่อย ๆ ที่สร้างขึ้น

2. อัลกอริทึมต้นไม้ตัดสินใจ C5.0

วิธีการวิเคราะห์ต้นไม้ตัดสินใจแบบ C5.0 เป็นวิธีการสร้างและประเมินต้นไม้การตัดสินใจ (decision tree) ที่พัฒนาโดย Ross Quinlan ซึ่งเป็นตัวอัพเกรดของ C4.5 ซึ่งนิยม

ใช้งานอย่างแพร่หลายในการแก้ปัญหาที่มีข้อมูลหลายมิติและต้องการทำนายหรือตัดสินใจเกี่ยวกับหลายกลุ่มข้อมูล (multi-class decision-making) อย่างไรก็ตาม C5.0 มีประสิทธิภาพในการทำงานและให้ผลลัพธ์ที่ดีกว่า C4.5 ในบางกรณีเนื่องจากใช้เทคนิคการเลือกแบบผสมผสาน (ensemble selection) และคำนวณค่าย่อย (subset) ของกฎที่เป็นไปได้ให้มากขึ้น

3. วิธี Random forest

วิธี Random forest เป็นเทคนิคในการสร้างแบบจำลองทำนาย (predictive model) ที่มาจากเทคนิคของ ensemble learning โดยใช้หลาย ๆ ต้นไม้ตัดสินใจ (decision trees) แล้วรวมผลลัพธ์ของทุกต้นไม้เพื่อทำนายผลลัพธ์ที่ถูกต้องและเสถียรขึ้น โดย Random forest เป็นวิธีที่ดีในการแก้ปัญหาการเรียนรู้มากเกินไป (overfitting) และมีความแม่นยำสูงกว่าต้นไม้ตัดสินใจแบบเดียวเมื่อใช้กับข้อมูลที่ซับซ้อนและมีความหลากหลาย

2.2 การแก้ปัญหาข้อมูลไม่สมดุล (solving the imbalanced data)

1. วิธีสุ่มเกิน (oversampling) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิม ในการวิจัยครั้งนี้จะใช้วิธีการสุ่มแบบเป็นระบบ 30% และ 35% โดยผลการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก

2. วิธีสุ่มลด (undersampling) เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย ในการวิจัยครั้งนี้จะใช้วิธีการสุ่มแบบเป็นระบบโดยทำการสุ่มเพิ่ม 50% และทำการสุ่มลด 20% ซึ่งผลการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย

ในงานนี้ผู้วิจัยได้ทำการแบ่งชุดข้อมูลออกเป็นสองส่วน โดยใช้ 75% ของข้อมูลเพื่อเป็นชุดข้อมูลการฝึก (training set) และ 25% ของข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ (test set)

3. เกณฑ์ในการวัดประสิทธิภาพความแม่นยำ

3.1 เกณฑ์การวัดด้วยค่าความถูกต้อง

ในการทดสอบประสิทธิภาพความแม่นยำ โดยใช้เกณฑ์ในการวัดด้วยค่าความถูกต้อง โดยคำนวณจากค่าในแนวเส้นทแยงมุมของเมทริกซ์ความสับสน (confusion matrix: CM)

Table 2 Confusion Matrix 2 x 2

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	True Positive (TP)	False Positive (FP)
Negative (0)	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

โดยที่

ค่าความถูกต้อง (accuracy) คือ ค่าอยู่ระหว่าง 0 - 1 เมื่อค่าเข้าใกล้ 1 นั่นคือตัวแบบสามารถจำแนกประเภทได้ดีมาก

TP คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (true positive)

FN คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (false negative)

TN คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (true negative)

FP คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (false positive)

3.2 เกณฑ์ในการทำนาย Area Under the Curve (AUC) เป็นตัววัดประสิทธิภาพของโมเดลทำนาย (predictive model) ในงานการจำแนกประเภท (classification) AUC จะวัดพื้นที่ใต้เส้น Curve ที่เกิดจากการพล็อต ROC (receiver operating characteristic) ซึ่งเป็นกราฟที่แสดงความสัมพันธ์ระหว่าง sensitivity และ specificity ของโมเดล ซึ่ง ROC Curve เป็นกราฟที่แสดงความสามารถของโมเดลในการแยกแยะ (discriminate) ระหว่างคลาสบวก (positive class) และคลาสลบ (negative class)

โดยที่

ค่าความไว (sensitivity) คือ ค่าของความถูกต้องในการพยากรณ์ของคลาสที่เกิดโรคต่อจำนวนทั้งหมดในกลุ่มของคลาสที่พยากรณ์ว่าเกิดโรค หรือเรียกอีกอย่างว่า True Positive Rate (TPR)

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

ค่าความจำเพาะ (specificity) คือ ค่าความถูกต้องในการพยากรณ์ของคลาสที่ไม่เกิดโรคต่อจำนวนทั้งหมดที่ไม่เกิดโรคจริง

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

False Positive Rate (FPR) หรือค่า 1-Specificity หมายถึงอัตราส่วนของ False Positives ต่อทั้งหมดที่เป็น Actual Negatives

$$FPR = \frac{FP}{FP + TN} \quad (4)$$

สามารถสรุปได้ดังเกณฑ์ต่อไปนี้

$0.50 \le AUC < 0.70$ คือ ตัวแบบมีประสิทธิภาพต่ำ

$0.70 \le AUC < 0.80$ คือ เกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่

$0.80 \le AUC < 0.90$ คือ ตัวแบบทำงานได้ดี

$AUC > 0.90$ คือ ตัวแบบทำงานได้ดีมาก

3.3 ค่าความระลึก (recall) คือความน่าจะเป็นที่โมเดลสามารถตรวจจับผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านมจากจำนวน ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านมทั้งหมดในข้อมูล

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (5)$$

ผลการวิจัย

ข้อมูลที่ใช้ในการวิจัยคือ ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 เมื่อแบ่งข้อมูลออกเป็นชุดฝึก (training set) และชุดทดสอบ (test set) โมเดลจะถูกฝึกอย่างต่อเนื่องในชุดฝึกและนำไปทดสอบบนชุดทดสอบเพื่อวัดประสิทธิภาพ แต่ถ้าความแตกต่างในชุดฝึกและชุดทดสอบมีความแตกต่างกันมาก ๆ อาจทำให้โมเดลมีประสิทธิภาพในการทำนายที่แตกต่างกัน ทำให้ค่าการพยากรณ์มีผลที่แตกต่างกันเนื่องจากความแตกต่างในข้อมูลที่ใช้ในการทดสอบและวัดประสิทธิภาพ ในงานวิจัยนี้ได้ทำการศึกษาอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยการแบ่งข้อมูลออกเป็นสัดส่วน 75% สำหรับการฝึกและ 25% สำหรับการทดสอบ

Table 3 The accuracy, area under the curve (AUC), and recall values of the models using the decision tree algorithms C4.5, C5.0, and the Random Forest method, with data split into 75% for training and 25% for testing.

Implement	Accuracy	AUC	Recall
C4.5	0.8770	0.5463	0.0208
C5.0	0.8976	0.5000	0.0000
Random forest	0.8635	0.5271	0.0208
Oversampling 30% + C4.5	0.7203	0.6468	0.1719
Oversampling 30% + C5.0	0.6970	0.5391	0.0151
Oversampling 30% + Random forest	0.7394	0.7100	0.3721
Oversampling 35% + C4.5	0.6963	0.7200	0.3452
Oversampling 35% + C5.0	0.6639	0.5620	0.0807
Oversampling 35% + Random forest	**0.7267**	**0.7648**	**0.4593**
Combining Random Oversampling and Undersampling + C4.5	0.6540	0.7026	0.3629
Combining Random Oversampling and Undersampling + C5.0	0.6138	0.5538	0.0145
Combining Random Oversampling and Undersampling + Random forest	0.7142	0.7612	0.4571

จาก Table 3 แสดงค่าความถูกต้อง (accuracy), ค่าเกณฑ์ในการทำนาย AUC (area under ROC curve) และค่าความระลึก (recall) ของแบบจำลองที่ฝึกฝนด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และ Random forest โดยการแบ่งข้อมูลออกเป็นสัดส่วน 75% สำหรับการฝึกและ 25% สำหรับการทดสอบ พบว่า อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ให้ค่าความถูกต้องค่อนข้างสูง แต่ค่าเกณฑ์ในการทำนาย AUC และค่าความระลึก (recall) ค่อนข้างต่ำ เนื่องจากการทำนายโมเดลไม่สามารถแยกกลุ่ม (class) ได้ดีพอ โมเดลอาจทำนายคลาสเดียวทั้งหมดหรือทำนายผิดพลาดในการแยกแยะกลุ่มของข้อมูลที่ซับซ้อน ผู้วิจัยทำการสุ่มข้อมูลเพิ่ม (oversampling) 30% และ 35% จะเห็นว่าค่า AUC และค่าความระลึก (recall) มากกว่าตอนที่ยังไม่ได้ทำการสุ่มเพิ่ม จากนั้นทำการสุ่มข้อมูลเพิ่มและลดข้อมูล (combining random oversampling and undersampling) พบว่าโมเดลที่ได้มีค่าความถูกต้อง, ค่า AUC และค่าความระลึก (recall) อยู่ในเกณฑ์ที่พึงพอใจและยอมรับได้ โดยวิธีที่ให้ผลลัพธ์ดีที่สุดคือวิธี Random forest ร่วมกับการสุ่มข้อมูลเพิ่ม (oversampling) 35% จากนั้นผู้วิจัยทำการหาจำนวนต้นไม้, ความลึกของต้นไม้ และ k-fold cross-validation เพื่อประเมินประสิทธิภาพของโมเดลในแต่ละค่าพารามิเตอร์ ซึ่งจำนวนต้นไม้ที่กำหนดคือ 50, 75, 100, 150, 200, 300, 400, 500 และ 1000 ความลึกของต้นไม้ที่กำหนดคือ 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 และ 19 ส่วนสุดท้าย k-fold cross-validation ที่กำหนดคือ 3, 5, 7, 9 และ 10 ซึ่งช่วยให้การประเมินมีความเสถียรและถูกต้องมากยิ่งขึ้น โดยลดผลกระทบจากการแบ่งข้อมูลแบบเฉพาะเจาะจง (specific) ที่อาจเกิดขึ้นในการแบ่งแยกแบบเดิม (train-test split) ที่ใช้เฉพาะชุดทดสอบ (test set) และชุดฝึก (train set) แบบเดียวเท่านั้น ผลลัพธ์แสดงให้เห็นว่าค่าที่ดีที่สุดคือ จำนวนต้นไม้เท่ากับ 200 ต้น, ความลึกของต้นไม้เท่ากับ 14 และ k-fold cross-validation เท่ากับ 7 (k=7) ซึ่งการปรับค่าพารามิเตอร์เหล่านี้ช่วยในการควบคุมความซับซ้อนของโมเดลโดยเฉพาะความลึกของต้นไม้ การเลือกค่าความลึกที่เหมาะสมสามารถป้องกันโมเดลจากการเรียนรู้ข้อมูลเกินไป (overfitting) หรือการไม่เรียนรู้เพียงพอ (underfitting) ซึ่งจะช่วยให้โมเดลมีประสิทธิภาพมากที่สุดในการทำนายข้อมูลที่ไม่เคยเห็นมาก่อน รวมทั้งการทดลองค่าพารามิเตอร์ต่าง ๆ ในขั้นตอนการฝึกและประเมินโมเดลสามารถช่วยประหยัดเวลาและทรัพยากรในกระบวนการพัฒนาโมเดลโดยไม่ต้องสร้างและทดลองทุก ๆ ค่าพารามิเตอร์ที่เป็นไปได้ หลังจากค้นพบค่าพารามิเตอร์ที่ดีที่สุดทำการฝึกและทดสอบโมเดล Random forest ด้วยการใช้ค่าพารามิเตอร์เหล่านี้ และประเมินประสิทธิภาพของโมเดลโดยใช้ข้อมูลที่ไม่เคยเห็นมาก่อน เพื่อให้แน่ใจว่าโมเดลทำงานได้ดีที่สุด พบว่าให้ค่าความถูกต้องในการทำนายคลาสต่าง ๆ บนชุดข้อมูลทดสอบที่ไม่เคยเห็นมาก่อนเท่ากับ 72.27% ดัง Figure 1 และค่า AUC อยู่ที่ 0.76 ดัง Figure 2 และมีค่าความระลึก (recall) เท่ากับ 0.4756 ซึ่งจากการใช้จำนวนต้นไม้ทั้งหมด 200 ต้นในการสร้างแบบจำลอง Random forest ทำให้ได้ปัจจัยที่ส่งผลในการจำแนกโรคมะเร็งเต้านม 5 อันดับแรก ได้แก่ ความสมมาตรของเต้านมสองข้าง, มีอาการที่นำมาพบแพทย์, ก้อนหรือถุงน้ำ, แคลเซียม และโครงสร้างเต้านม ตามลำดับ

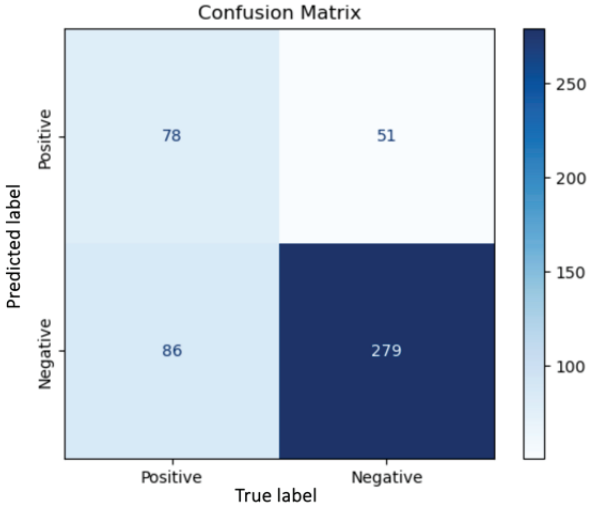

Figure 1 The Confusion Matrix of oversampling data at a 35% ratio using the Random Forest method with 7-fold cross-validation.

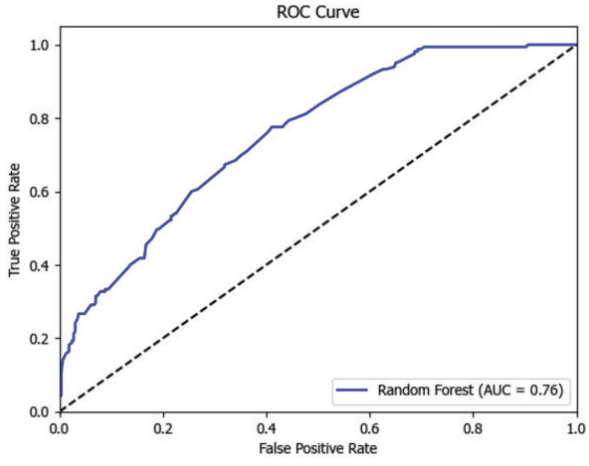

Figure 2 ROC curve for oversampling data at a 35% ratio using the Random Forest method with 7-fold cross-validation.

สรุปผลการวิจัย

การวิจัยการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมโดยการสังเคราะห์ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านมจากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่าง

ปี พ.ศ. 2553 ถึง พ.ศ. 2565 ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยพบว่าข้อมูลที่ใช้ในการจำแนกคลาสมีจำนวนของคลาสต่างกันมากน้อยไม่เท่ากัน (class imbalance) ซึ่งอาจทำให้โมเดลที่ได้สร้างขึ้นมีความสามารถในการทำนายคลาสที่มีจำนวนตัวอย่างมาก มากกว่าคลาสที่มีจำนวนตัวอย่างน้อย ๆ นำมาซึ่งผลลัพธ์ที่ไม่เสถียรและไม่แม่นยำในคลาสที่มีจำนวนตัวอย่างน้อย โดยปัญหานี้เกิดขึ้นได้ทั่วไปโดยเฉพาะข้อมูลทางการแพทย์ เนื่องจากการตรวจจับโรคหรือความเสี่ยงของโรคที่มีอัตราการเกิดต่อประชากรต่ำ ทำให้คลาสของผู้ป่วยหรือผู้ที่มีความเสี่ยงสูงมีจำนวนน้อยเมื่อเปรียบเทียบกับคลาสของผู้ที่ไม่มีโรคหรือมีความเสี่ยงต่ำ เพื่อแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้ใช้เทคนิคการสุ่มเพิ่ม (oversampling) เพื่อเพิ่มจำนวนตัวอย่างในคลาสที่น้อยเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน และวิธีสุ่มลด (undersampling) ลดตัวอย่างในคลาสที่มีจำนวนมากลงเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน รวมทั้งการแบ่งข้อมูล (split data) แบบต่าง ๆ ทำให้ค่าการพยากรณ์มีผลต่างกันเนื่องจากความแตกต่างในชุดข้อมูลที่ถูกใช้ในการสร้างและทดสอบโมเดล และความแตกต่างในแบบจำลองที่ใช้ในการวิเคราะห์ข้อมูล จากการแก้ปัญหาจะเห็นว่า C4.5 และ C5.0 ให้ผลลัพธ์ที่ไม่ต่างจากเดิมและผลลัพธ์ที่ได้ไม่ต่างกันมากนัก ส่วนวิธี Random forest ให้ค่า AUC ที่ดีขึ้นเมื่อเปรียบเทียบกับ C4.5 และ C5.0 ซึ่งสูงกว่าประมาณ 15-20% รวมทั้งค่าความระลึก (recall) ที่เพิ่มมากขึ้น อาจเกิดขึ้นเนื่องจากคุณสมบัติของวิธี Random forest ใช้วิธีการเรียนรู้แบบรวมกลุ่ม (ensemble) ของต้นไม้ตัดสินใจโดยการสุ่มข้อมูลและสุ่มคุณลักษณะ (feature) ที่ใช้ในการสร้างแต่ละต้นไม้ วิธีการเรียนรู้แบบรวมกลุ่มช่วยลดความเสี่ยงในการเรียนรู้โมเดลจากข้อมูลที่ไม่สมดุล (class imbalance) และช่วยลดการเรียนรู้มากเกินไป (overfitting) โดยที่อัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 อาจมีแนวโน้มที่จะเกิดขึ้นได้ อีกทั้งวิธี Random forest สร้างต้นไม้หลายต้นและรวมผลลัพธ์จากทุกต้นในการตัดสินใจ (voting) ซึ่งช่วยลดความผิดพลาดและเพิ่มความแม่นยำของโมเดล ในทางตรงกันข้ามอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีเพียงต้นไม้ต้นเดียวซึ่งมีความหลากหลายที่น้อยกว่า จากผลลัพธ์ที่ได้วิธี Random forest ร่วมกับการสุ่มข้อมูลเพิ่ม (oversampling) 35% เป็นวิธีที่ดีที่สุดสำหรับชุดข้อมูลที่ทำการศึกษา จากนั้นผู้วิจัย ทำการหาจำนวนต้นไม้ที่เหมาะสมและความลึกของต้นไม้พร้อมกับการทำ k-fold cross validation เพื่อดูค่าที่เปลี่ยนไปในแต่ละรอบ ผลการทดสอบแสดงให้เห็นว่าจำนวนต้นไม้ที่ดีที่สุดคือ 200 ต้น ความลึกของต้นไม้ คือ 14 และ k-fold ที่ดีที่สุดคือ 7 (k=7) สำหรับการแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้เป็นอีกหนึ่งแง่ที่สำคัญที่ช่วยให้โมเดลทำนายได้แม่นยำและเสถียร ซึ่งเป็นประโยชน์ในการวิเคราะห์โรคและการตรวจสอบความเสี่ยงในโดเมนทางการแพทย์ ซึ่งเป็นงานที่ความถูกต้องและน่าเชื่อถือมีความสำคัญเป็นอย่างยิ่ง การจัดการกับปัญหาความไม่สมดุลข้อมูลและการแบ่งข้อมูลออกเป็นชุดฝึกและชุดทดสอบเป็นปัจจัยสำคัญในการทำนาย การสุ่มข้อมูลเพิ่มและลดข้อมูลอาจช่วยปรับปรุงความแม่นยำของโมเดล แต่อาจมีผลให้ข้อมูลเสียหายและเพิ่มเวลาในการประมวลผล ควรพิจารณาความสมดุลของข้อมูลและการแบ่งข้อมูลให้อยู่ในเกณฑ์ที่เหมาะสมในแต่ละกรณีการวิเคราะห์ข้อมูลนั้น ๆ ซึ่งเป็นสิ่งสำคัญในการสร้างโมเดลที่มีประสิทธิภาพในการจำแนกข้อมูลทางการแพทย์ ฉะนั้นการเลือกอัลกอริทึมที่เหมาะสมจะขึ้นอยู่กับความต้องการของงานและลักษณะของข้อมูล และอาจจะต้องพิจารณาเพิ่มเติมเกี่ยวกับความเหมาะสมของข้อมูลและอัลกอริทึมในงานที่มีลักษณะแตกต่างกัน

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณโรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม ที่ได้อนุญาตให้ใช้ข้อมูลสำหรับนำมาศึกษาในงานวิจัย และขอขอบคุณคณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม สำหรับสถานที่ในการทำวิจัย

เอกสารอ้างอิง

ณัฏฐพร นันทิวัฒนา. (2563). *มะเร็งเต้านม มะเร็งอันดับ 1 ของผู้หญิง*. โรงพยาบาล ศิครินทร์. https://www.sikarin.com/doctor-articles/โรคมะเร็งเต้านม-มะเร็ง

สถาบันมะเร็งแห่งชาติ. (2561). *ทะเบียนมะเร็งระดับโรงพยาบาล พ.ศ. 2559*. พรทรัพย์การพิมพ์.

สายชล สินสมบูรณ์ทอง. (2560). *การทำเหมืองข้อมูล เล่ม 1 การค้นหาความรู้จากข้อมูล*. จามจุรีโปรดักส์.

อุกฤษฏ์ ศรีสุข และจารี ทองคำ. (2564). การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับอุบัติการณ์ของผู้ป่วย. *วารสาร วิทยาศาสตร์ และ เทคโนโลยี มหาวิทยาลัยมหาสารคาม*, *40*(2), 157-163.

Nemade, V., & Fegade, V. (2023). Machine learning techniques for breast cancer prediction. *Procedia Computer Science*, *218*, 1314-1320.