การจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ กรณีศึกษา: โรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม

วิทยานิพนธ์

ของ

ชัยยันต์ สุขหมั่น

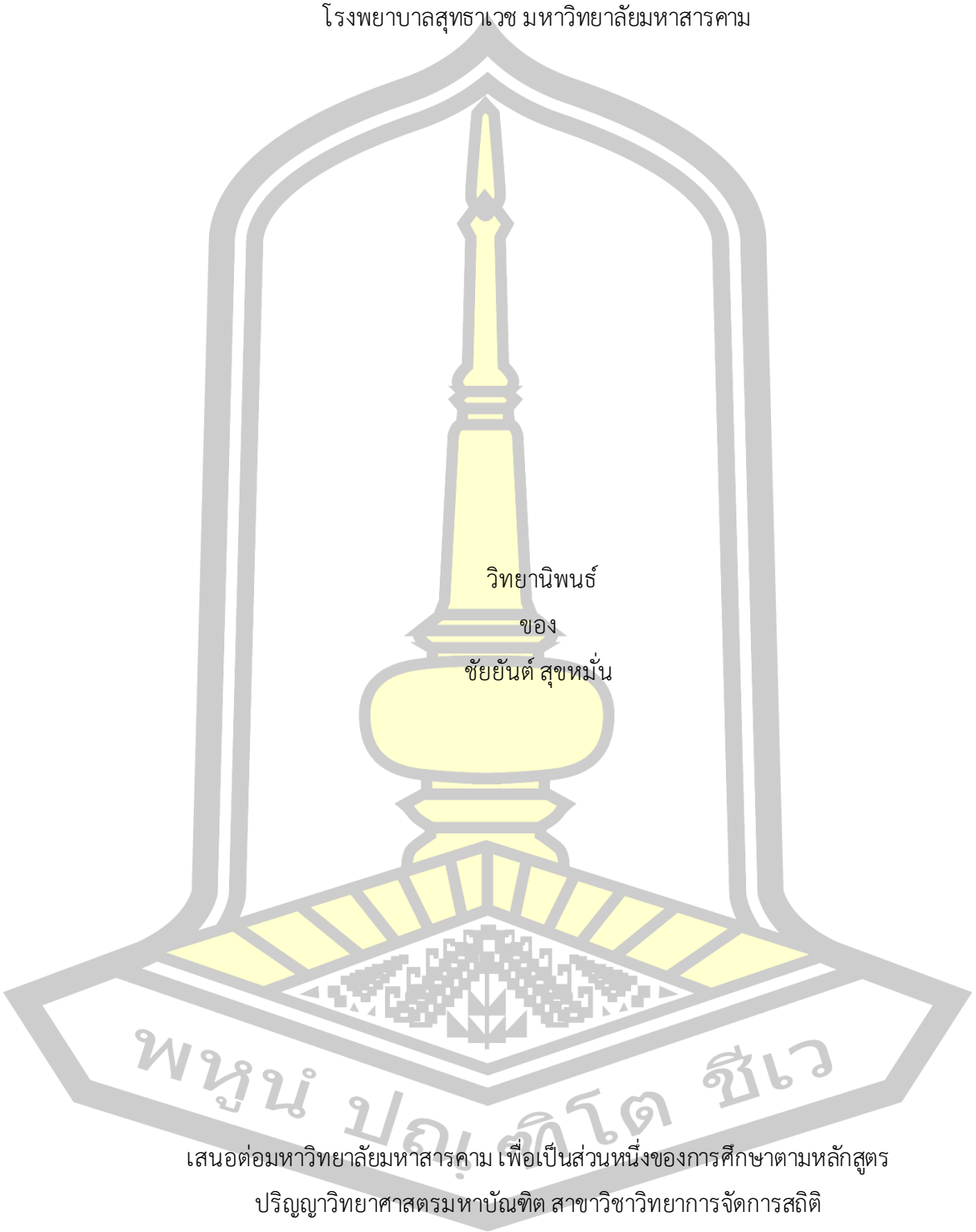
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ

ธันวาคม 2566

พหูน ปณฺ ฑิโต ชีเว

การจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ กรณีศึกษา: โรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม



วิทยานิพนธ์

ของ

ชัยยันต์ สุขหมั่น

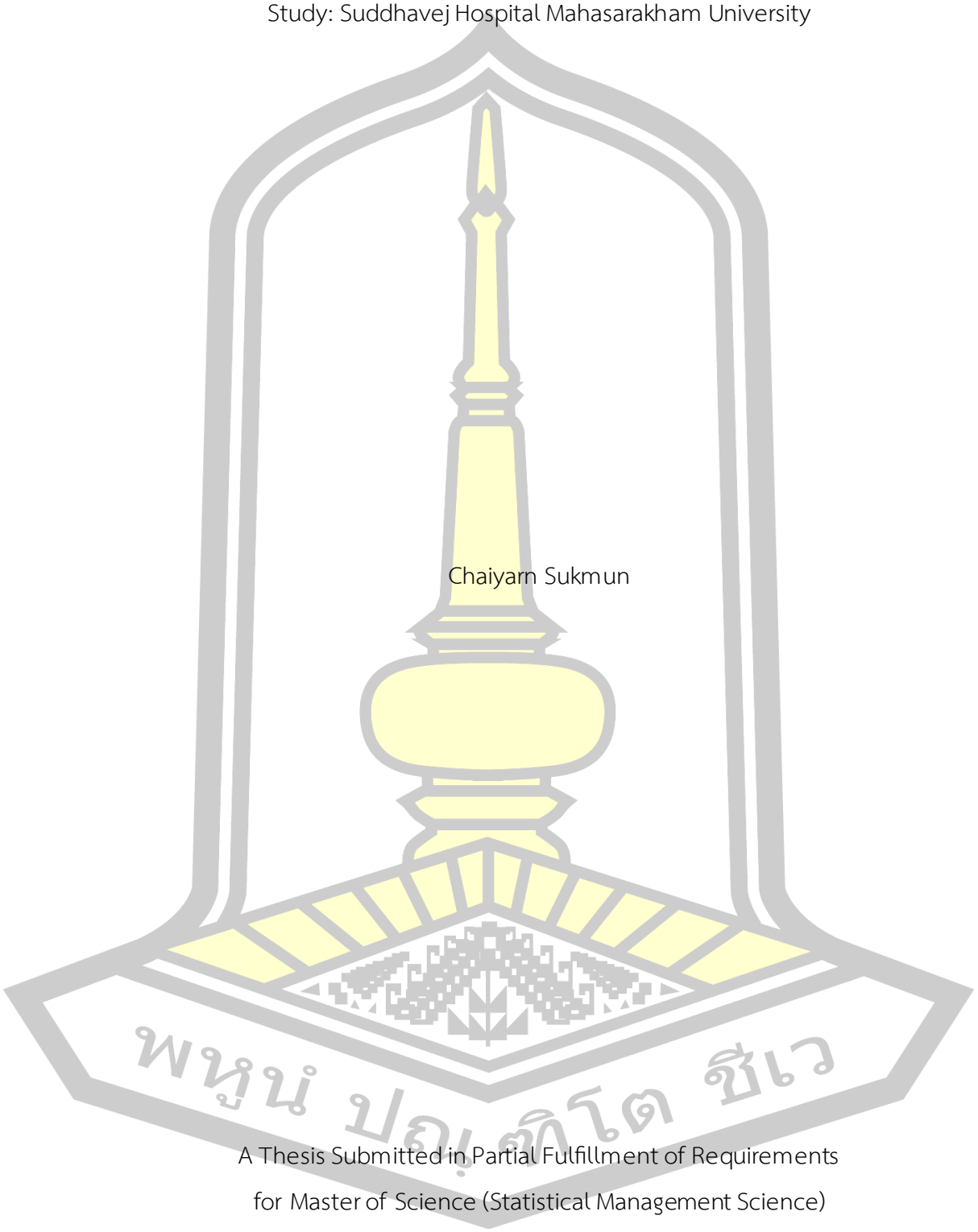
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ

ธันวาคม 2566

Classification of People at Risk for Breast Cancer Using Decision Tree Algorithm, Case Study: Suddhavej Hospital Mahasarakham University

Chaiyarn Sukmun



A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Statistical Management Science)

December 2023

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายชัยยันต์ สุขหมั่น แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. ธนพงศ์ อินทระ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อ. ดร. สุภาวดี วิชิตชาญ)

.....กรรมการ
(ผศ. ดร. สุจิตตา สุระภี)

.....กรรมการ
(ผศ. ดร. มนชยา เจียงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

.....
(ศ. ดร. ไพโรจน์ ประมวล)
คณบดีคณะวิทยาศาสตร์

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ กรณีศึกษา: โรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม

ผู้วิจัย ชัยยันต์ สุขหมั่น

อาจารย์ที่ปรึกษา อาจารย์ ดร. สุภาวดี วิชิตชาญ

ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการจัดการสถิติ

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2566

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกประเภท (classification) โรคมะเร็งเต้านม (breast cancer) และศึกษาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม ผู้วิจัยได้ใช้ข้อมูลเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 จากการทำความสะอาดข้อมูลเหลือข้อมูลทั้งหมด 1,524 ระเบียน ซึ่งมีข้อมูลผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม จำนวน 1,343 ระเบียน และข้อมูลผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านมจำนวน 181 ระเบียน จากผลการศึกษาพบว่า อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ให้ค่าความถูกต้อง (accuracy) ค่อนข้างสูง แต่ค่าเกณฑ์ในการทำนาย AUC (area under ROC curve) ค่อนข้างต่ำ เนื่องจากการทำนายโมเดลไม่สามารถแยกกลุ่ม (class) ได้ดีพอ ซึ่งพบว่าข้อมูลที่ใช้ในการจำแนกคลาสมีจำนวนของคลาสมากน้อยไม่เท่ากัน (class imbalance) เพื่อแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้ใช้วิธีการสุ่มเกิน (oversampling) เพื่อเพิ่มจำนวนตัวอย่างในคลาสที่น้อยเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน และวิธีการสุ่มลด (undersampling) ลดตัวอย่างในคลาสที่มีจำนวนมากลงเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน รวมทั้งการแบ่งข้อมูล (split data) แบบต่าง ๆ พบว่าอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ให้ผลลัพธ์ไม่ต่างจากเดิมและผลลัพธ์ที่ได้ไม่ต่างกันมากนัก ส่วนวิธี Random forest ให้ค่า AUC ที่ดีขึ้นเมื่อเปรียบเทียบกับอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ซึ่งปัจจัยที่ส่งผลในการจำแนกโรคมะเร็งเต้านม 5 อันดับแรก ได้แก่ ดัชนีมวลกาย, กลุ่มอายุ, อาการที่นำมาพบแพทย์, ความสมมาตรของเต้านมสองข้าง และก้อนหรือถุงน้ำ ตามลำดับ

คำสำคัญ : มะเร็งเต้านม, ต้นไม้ตัดสินใจ, ข้อมูลไม่สมดุล

TITLE Classification of People at Risk for Breast Cancer Using Decision Tree Algorithm, Case Study: Suddhavej Hospital Mahasarakham University

AUTHOR Chaiyarn Sukmun

ADVISORS Supawadee Wichitchan , Ph.D.

DEGREE Master of Science **MAJOR** Statistical Management Science

UNIVERSITY Mahasarakham University **YEAR** 2023

ABSTRACT

This research focuses on evaluating the effectiveness of the Decision Tree Algorithm in classifying risk level of breast cancer, as well as investigating the associated risk factors. The study employs medical record data from breast mass patients at Faculty of Medicine, Mahasarakham University, spanning 2010 to 2022. The dataset, post-cleansing, comprises 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, exhibit substantial classification accuracy. However, their area under the ROC curve (AUC) values are relatively low due to insufficient class separation, which stems from class imbalance. The research tackles this issue by employing oversampling to augment the minority class instances and undersampling to reduce the majority class instances. Various data-splitting techniques were also explored. The outcomes reveal that both C4.5 and C5.0 Decision Trees yield comparable results, while Random Forest demonstrates a superior AUC, higher than C4.5 and C5.0. The top 5 factors affecting the classification of breast cancer are body mass index, age group, chief complaint, symmetry and mass or cyst, respectively.

Keyword : Breast cancer, Decision Tree, Class-Imbalance

พหูน ปณฺ ฑิโต ชีเว

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี ด้วยความกรุณาในการให้คำปรึกษาแนะนำเป็นอย่างดียิ่งจากอาจารย์ ดร.สุภาวดี วิชิตชาญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ชี้แนวทางในการศึกษาวิทยานิพนธ์ ให้คำแนะนำในการค้นคว้าวิเคราะห์ข้อมูล ตลอดจนแก้ไข ปรับปรุงข้อบกพร่องต่าง ๆ ทำให้วิทยานิพนธ์สำเร็จสมบูรณ์ด้วยดี รวมทั้ง ผศ. ดร.ธนพงศ์ อินทระ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. สุจิตตา สุระภี และ ผู้ช่วยศาสตราจารย์ ดร.มนชยา เจียงประดิษฐ์ กรรมการสอบวิทยานิพนธ์ที่กรุณาให้คำแนะนำและให้ข้อเสนอแนะ ตลอดจนแก้ไขปรับปรุงข้อบกพร่องต่าง ๆ อันเป็นประโยชน์ต่อวิทยานิพนธ์ ทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น ผู้วิจัยขอกราบขอบพระคุณอย่างสูงยิ่งไว้ ณ โอกาสนี้ด้วย

ทั้งนี้ผู้วิจัยขอขอบคุณโรงพยาบาลสุทธาเวช คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ที่ได้อนุเคราะห์ข้อมูลของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม ซึ่งเป็นประโยชน์อย่างมากต่อการทำวิทยานิพนธ์ในครั้งนี้

ผู้วิจัยขอกราบขอบพระคุณบุพการี ซึ่งเป็นผู้ให้ชีวิต ความรัก ความอบอุ่น และบูรพคณาจารย์ ซึ่งเป็นผู้ให้ความรู้ มีส่วนวางรากฐานทางการศึกษา รวมถึงพี่น้องสาขาวิทยาการจัดการสถิติ และผู้มีส่วนเกี่ยวข้องทุกท่าน ซึ่งเป็นผู้ให้แรงบันดาลใจในการทำงาน คอยเตือนสติ และเป็นกำลังใจให้แก่ผู้ทำวิทยานิพนธ์จนบรรลุผลสำเร็จไปได้ด้วยดี

ชัยยันต์ สุขหมั่น

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1 บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

มะเร็ง คือ กลุ่มของโรคที่เกิดเนื่องจากเซลล์ของร่างกายมีความผิดปกติ ที่ DNA หรือสารพันธุกรรม ส่งผลให้เซลล์มีการเจริญเติบโต มีการแบ่งตัวเพื่อเพิ่มจำนวนเซลล์รวดเร็วและมากกว่าปกติ ดังนั้นจึงอาจทำให้เกิดก้อนเนื้อผิดปกติ และในที่สุดก็จะทำให้เกิดการตายของเซลล์ในก้อนเนื้อนั้น เนื่องจากขาดเลือดไปเลี้ยง ถ้าเซลล์พวกนี้เกิดอยู่ในอวัยวะใดก็จะเรียกชื่อ “มะเร็ง” ตามอวัยวะนั้น เช่น มะเร็งปอด, มะเร็งสมอง, มะเร็งเต้านม, มะเร็งปากมดลูก, มะเร็งเม็ดเลือดขาว, มะเร็งต่อมน้ำเหลือง และมะเร็งผิวหนัง เป็นต้น (สถาบันมะเร็งแห่งชาติ กรมการแพทย์ กระทรวงสาธารณสุข, 2563) “มะเร็งเต้านม” เป็นมะเร็งที่พบมากที่สุดเป็นอันดับ 1 ของผู้หญิงไทย และเป็นสาเหตุของการเสียชีวิตอันดับต้น ๆ ในผู้หญิงทั่วโลก แนวโน้มคนไทยป่วยเป็นโรคมะเร็งสูงขึ้นทุกปีแต่ยังพบน้อยกว่าประเทศทางตะวันตกมาก โดยหญิงไทยมีอัตราการพบมะเร็งประมาณ 40 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ซึ่งถ้าเทียบกับประเทศตะวันตกพบมะเร็งเต้านมได้มากกว่า 100 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ส่วนในผู้ชายก็พบมะเร็งเต้านมได้เช่นกันแต่ไม่บ่อยนัก โดยมีอุบัติการณ์ของโรคนี้น้อยกว่าผู้หญิงเกือบ 100 เท่า (ณัฏฐพร นันทิวัฒนา, 2563)

สถานการณ์ของโรคมะเร็งในภาพรวมของประเทศไทย จากสถิติพบว่า โรคมะเร็งเป็นสาเหตุการเสียชีวิตอันดับ 1 คิดเป็นร้อยละ 16 ของต้นเหตุการเสียชีวิตทั้งหมด สูงกว่าอันตราการเสียชีวิตจากอุบัติเหตุ 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 (สายชล สินสมบูรณ์ทอง, 2560) หรือที่เรียกกันว่าการทำเหมืองข้อมูล เป็นวิธีการที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์หรือทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้นซึ่งการค้นหาความรู้และความจริงที่แฝงอยู่ในข้อมูล (knowledge discovery) เป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่มีอยู่เพื่อให้ได้สารสนเทศที่มีประโยชน์ (useful Information) ที่เรายังไม่ทราบ (unknown data) โดยเป็นสารสนเทศที่มีเหตุผล (valid information) และสามารถนำไปใช้ได้ (actionable) ซึ่งเป็นสิ่งสำคัญที่จะช่วยการตัดสินใจในการทำธุรกิจการทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในการค้นหา

ความรู้จากฐานข้อมูลขนาดใหญ่ ซึ่งมีนักวิจัยหลายคนได้นำกระบวนการทำเมืองข้อมูลมาประยุกต์ใช้ในการพยากรณ์ในด้านต่าง ๆ อาทิ Mosayebi, A. et al. (2020) ได้ทำการศึกษาในหัวข้อเรื่อง “การสร้างแบบจำลองและการเปรียบเทียบอัลกอริทึมการทำเหมืองข้อมูลเพื่อทำนายการกลับเป็นซ้ำของมะเร็งเต้านม” โดยได้ทำการเปรียบเทียบ 7 อัลกอริทึม ผลปรากฏว่าอัลกอริทึมต้นไม้ตัดสินใจ (decision tree: C5.0) มีความแม่นยำมากที่สุด งานวิจัยอื่นที่ทำโดย Rajeswari, S. และ Suthendran, K. (2019) ได้นำการทำเหมืองข้อมูลมาประยุกต์ใช้ในด้านการเกษตรในหัวข้อเรื่อง “โมเดลการจำแนกต้นไม้ตัดสินใจขั้นสูง (ADT) สำหรับการวิเคราะห์ข้อมูลทางการเกษตรบน cloud” ซึ่งทำการศึกษาระดับความอุดมสมบูรณ์ของดินที่เหมาะสมกับการปลูกพืชแต่ละชนิดในประเทศอินเดียผ่านแอพพลิเคชัน ซึ่งพบว่า C5.0 ADT พร้อมกับการคัดเลือกคุณลักษณะ (feature selection) ให้ความแม่นยำที่สูงกว่าอัลกอริทึม C4.5 และ C5.0 อีกทั้งยังมีนักวิจัยหลายคนได้นำเทคนิคการทำเหมืองข้อมูลไปใช้รวมถึงชัดชัย แก้วตา และอัจฉรา มหาวีรวัฒน์ (2553) กล่าวถึงการวินิจฉัยคดีด้วยเทคนิคต้นไม้ตัดสินใจ เพื่อจำแนกความผิดออกเป็นมาตราต่าง ๆ ที่เหมาะสมกับคดีความนั้น และเปรียบเทียบความถูกต้องของการจำแนกข้อมูล โดยจากผลการทดลองพบว่าต้นไม้ตัดสินใจที่สร้างจาก C4.5 สามารถจำแนกความผิดได้ถูกต้องมากกว่า ID3 เป็นต้น จะเห็นว่าการทำเหมืองข้อมูลโดยเฉพาะอัลกอริทึมต้นไม้ตัดสินใจ (decision tree) เป็นวิธีที่ศึกษากันอย่างแพร่หลายและเป็นที่นิยมในปัจจุบัน

ดังนั้น งานวิจัยนี้ผู้วิจัยมีแนวคิดในการนำเอาเทคนิคการทำเหมืองข้อมูลต้นไม้ต้นสินใจ (decision tree) มาใช้ในการจำแนกการเป็นโรคมะเร็งเต้านม ซึ่งเป็นโรคที่ยังไม่ทราบสาเหตุที่แน่ชัดและเป็นกันมากทั่วโลกรวมถึงในประเทศไทย งานวิจัยนี้จะเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกการเป็นโรคมะเร็งเต้านม โดยใช้วิธี C4.5, C5.0 และ Random forest เพื่อหาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม โดยการแบ่งกลุ่มข้อมูลเป็นชุดการเรียนรู้ (training data) และชุดข้อมูลทดสอบ (testing data) และวัดประสิทธิภาพโมเดลของแบบจำลองในแง่ของค่าความถูกต้อง (accuracy) และเกณฑ์ในการทำนาย Area Under Curve (AUC) ผลการวิจัยจะสามารถเป็นข้อมูลให้แพทย์ตัดสินใจในการยืนยันผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม และเมื่อหาแบบจำลอง (model) ที่มีความน่าเชื่อถือได้ก็สามารถที่จะนำข้อมูลผู้ที่เข้ารับการตรวจมะเร็งเต้านมมาเข้าในแบบจำลองเพื่อจะดูว่าผู้ที่เข้ารับการตรวจมีความเสี่ยงในการเป็นโรคมะเร็งเต้านมหรือไม่ อีกทั้งยังประชาสัมพันธ์ถึงสาธารณะชนเพื่อให้ความรู้เกี่ยวกับโรคมะเร็งเต้านม และบุคคลที่มีปัจจัยเสี่ยงในการเป็นโรคมะเร็งเต้านมให้เข้ารับการตรวจโดยแพทย์ผู้เชี่ยวชาญได้อย่างทันการ

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาและเลือกตัวแบบ (model) ที่เหมาะสมในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ

1.2.2 เพื่อวิเคราะห์ปัจจัยที่มีผลต่อการจำแนกผู้ที่มีความเสี่ยงต่อการเป็นโรคมะเร็งเต้านม

1.3 สมมติฐานการวิจัย

1.3.1 วิธี Random forest เป็นวิธีการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมได้มีประสิทธิภาพสูงกว่าอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm)

1.3.2 ดัชนีมวลกายและอายุเป็นปัจจัยสำคัญที่มีผลต่อการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม

1.4 ขอบเขตการวิจัย

1.4.1 ตัวแปรที่ใช้ในการวิจัย คือ เพศ (sex), อายุ (age), ดัชนีมวลกาย (body mass index: BMI), สูบบุหรี่ (smoking), ดื่มสุรา (binge), อาการที่นำมาพบแพทย์ (chief complaint), ก้อนหรือถุงน้ำ (mass or cyst), ความสมมาตรของเต้านมสองข้าง (asymmetries), หินปูน (calcification), โครงสร้างเต้านม (architectural distortion) และความเสี่ยง (risk)

1.4.2 กลุ่มตัวอย่างที่ใช้ในการศึกษา คือ ข้อมูลเวชระเบียนผู้ที่เข้ารับการตรวจมะเร็งเต้านมตั้งแต่ปี 2553 ถึง ปี 2565 โรงพยาบาลสุทธาเวช มหาวิทยาลัยมหาสารคาม จำนวน 1845 ระเบียน โดยข้อมูลได้จากการซักประวัติผู้เข้ารับการตรวจมะเร็งเต้านม พร้อมทั้งผลจากแมมโมแกรม (mammogram) โดยมีค่า BIRADs Score ว่าผู้ที่เข้ารับการตรวจอยู่ในกลุ่มผู้ที่มีกลุ่มเสี่ยงต่ำหรือกลุ่มผู้ที่มีความเสี่ยงสูง ซึ่งกลุ่มที่มีความเสี่ยงต่ำจะมีค่า BIRADs Score อยู่ระหว่าง 0-3 ส่วนกลุ่มที่มีความเสี่ยงสูงจะมีค่า BIRADs Score อยู่ระหว่าง 4-6

1.4.3 ขนาดของข้อมูลฝึก (training data) ที่ใช้ในงานวิจัยนี้คือ ข้อมูลฝึกขนาด 70%, 75%, และ 80% ของข้อมูลทั้งหมด แบ่งข้อมูลโดยใช้วิธี Split Test และวิธี K-fold cross validation

1.4.4 วิธีการจำแนกประเภทในชุดข้อมูลที่ใช้ในงานวิจัยนี้ คือ อัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) และ วิธี Random forest

1.4.5 โปรแกรมที่ใช้ในการวิเคราะห์ข้อมูล คือ โปรแกรมภาษา R และโปรแกรมภาษา Python

1.5 นิยามศัพท์เฉพาะ

1.5.1 ต้นไม้ตัดสินใจ (decision tree) หมายถึง เครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงตั้งแต่ต้นที่มีรากและแขนงต่าง ๆ แตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้

1.5.2 คุณลักษณะ (attribute) หมายถึง คุณลักษณะภายในชุดข้อมูล ในทางสถิติคือตัวแปรอิสระ หรือ Independent Variable

1.5.3 เป้าหมาย (target) หมายถึง ข้อมูลที่เป็น Class หรือผลลัพธ์ (output) ในทางสถิติคือตัวแปรตาม หรือ Dependent Variable

1.5.4 การระบุกลุ่มข้อมูล (label) หมายถึง ตัวบ่งบอกว่าข้อมูลที่ให้ฝึกเป็นอะไร โดยจะใช้กับการเรียนรู้แบบเครื่อง (machine learning) แบบมีผู้ช่วยสอน (supervised) เช่น ต้องสอน Machine ให้จำว่า Inputs แบบนี้ แล้วจะได้ Output แบบไหน

1.5.5 ข้อมูลฝึก (training data) หมายถึง ข้อมูลสำหรับฝึกตัวแบบ โดยจะฝึกให้ผลลัพธ์ออกมาเป็นไปตามชุดข้อมูลต้นฉบับ หากข้อมูลภายในชุดข้อมูลฝึกมีค่าผิดหรือค่าที่ไม่ถูกต้อง ผลลัพธ์ที่ออกมาก็จะผิด

1.5.6 ข้อมูลทดสอบ (test data) หมายถึง ข้อมูลสำหรับการทดสอบ โดยข้อมูลที่ใช้สำหรับทดสอบไม่ควรนำไปใช้ร่วมกับชุดข้อมูลฝึก เพราะถ้าหากใช้ร่วมกันจะทำให้การฝึกถูกต้อง และแม่นยำมากไปกับข้อมูลชุดนั้น ๆ (model overfitting)

1.5.7 วิธี Split Test คือ การแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% หรือ 80%) ใช้ในการสร้างโมเดลและข้อมูล ส่วนที่สอง (30% หรือ 20%) ใช้ในการทดสอบประสิทธิภาพของโมเดล (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

1.5.8 วิธี k-fold cross-validation เป็นการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี cross-validation จะทำการแบ่งข้อมูลออกเป็นหลายส่วน เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูล หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

1.5.9 BI-RADS Score คือ การตรวจแมมโมแกรม จะแปลผลเป็นศัพท์ทางการแพทย์ที่เรียกว่า BI-RADS ซึ่งย่อมาจากคำว่า Breast Imaging – Reporting and Data System โดยมีค่าความผิดปกติ แบ่งเป็น BI-RADS 0–6 ซึ่งค่า 0-3 แปลผลว่าผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และค่า 4-6 แปลผลว่าผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (Spak, D. A. et al., 2017)

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ศึกษาการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านม ผู้วิจัยได้ศึกษาเอกสาร ทฤษฎี ตำรา และงานวิจัยที่เกี่ยวข้อง โดยนำเสนอเนื้อหาตามลำดับดังต่อไปนี้

2.1 โรคมะเร็งเต้านม

2.2 การทำเหมืองข้อมูล

2.3 อัลกอริทึมต้นไม้ตัดสินใจ

2.4 วิธี Random forest

2.5 เกณฑ์ในการวัดประสิทธิภาพความแม่นยำ

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 โรคมะเร็งเต้านม

2.1.1. ประวัติความเป็นมาของโรคมะเร็งเต้านม

มะเร็งเป็นโรคที่มีมานานแล้วแต่มีหลักฐานชัดเจนแสดงไว้ครั้งแรก คือ คำจารึกบนกระดาษ papyrus ในสมัยอียิปต์ กล่าวถึงเนื้องอกที่เต้านมว่าเป็นก้อนแข็ง เย็น และไม่มีวิธีการรักษา ในสมัยนั้นคิดว่าก้อนมะเร็งนี้เป็นอาการอย่างหนึ่งคล้ายกับอาการไข้ ความคิดนี้คงอยู่มาตลอดจนถึงสมัยของนักปราชญ์ Hippocrates (400 ปีก่อนคริสตกาล) เขาได้กล่าวถึงมะเร็งในหนังสือ Corpus Hippocraticum โดยใช้คำว่า Karkinoma, Kaninos หรือ Skirros แทน Hippocrates คิดว่ามะเร็งเต้านมมีความสัมพันธ์กับประจำเดือน การที่ไม่มีประจำเดือนทำให้มีการคั่งของสารบางอย่างที่เต้านมเกิดเป็นก้อนขึ้น และกลายเป็นมะเร็งในระยะต่อมา ซึ่งการผ่าตัดรักษาจะทำให้อาการทรุดตัวลง

ประมาณปี ค.ศ. 100 Calcus ปราชญ์ชาวโรมัน เป็นคนแรกที่กล่าวว่าโรคมะเร็งเต้านมเป็นโรคที่พบบ่อยในสตรี การรักษาทำโดยการจี้ด้วยสารหนูซึ่งจะทำให้ได้ในระยะแรกเท่านั้น อีกประมาณ 100 ปีต่อมา Galen แห่งเมือง Alexandria ได้ตั้งทฤษฎี Humoral Theory (melancholos หรือ black bile) โดยยึดหลักของ Hippocrates ทฤษฎีของ Galen เป็นที่ยึดถือเป็นระยะเวลาอีกหลายร้อยปี ในทฤษฎีมีใจความว่า black bile ซึ่งเป็นชื่อของเสียที่ตับสร้างขึ้น และถูกทำลายที่ม้ามมีปริมาณมากขึ้นในร่างกาย อาจด้วยสาเหตุตับสร้างมากขึ้นหรือม้ามไม่สามารถทำลายได้หมด จากการคั่งของ black bile ร่างกายจะพยายามขับออกทำให้มีอาการเส้นเลือดขอด ริดสีดวง หรือแม้กระทั่งการมีประจำเดือนในสตรี ถ้ากระบวนการนี้เสียก็จะมีอาการแสดงออกมาที่เต้านม การรักษาคือการเอาสาร black bile ออกจากร่างกายโดยการถ่ายเลือด การอบน้ำร้อน การรับประทานอาหารบางอย่าง Galen ได้อ้างถึงผลสำเร็จที่เขาได้รับจากการรักษาด้วยวิธีนี้

หลังจากสมัยของ Galen ความรู้หยุดนิ่งลงคือระหว่างคริสต์ศตวรรษที่ 6 ถึง 15 หรือที่รู้จักกันว่าเป็นยุคมืดหรือยุคกลาง ในช่วงนี้การศึกษาความรู้ทำได้เฉพาะในโบสถ์ พระเท่านั้นที่สามารถอ่านเขียนได้ ประชาชนทั่วไปไม่มีสิทธิ์ในการศึกษา ความรู้เรื่องโรคหยุดนิ่งอยู่กับที่

เมื่อพ้นยุคมืดเข้าสู่สมัย Renaissance ซึ่งมีนักวิทยาศาสตร์มีชื่อหลายท่าน เช่น Wilhelm Febry (ค.ศ. 1565-1634) กล่าวถึงสาเหตุว่ามะเร็งเต้านมเกิดจากการคั่งของ milk curdling ศาสตราจารย์ทางกายวิภาคศาสตร์ และเป็นศัลยแพทย์ในเมือง Padua (ค.ศ. 1523-1562) ตั้งทฤษฎีเกี่ยวกับความไม่สมดุลของเลือด และ black bile เขายังสังเกตว่ามะเร็งชนิดนี้มีการอักเสบ ได้รับการพยากรณ์โรคว่าเลวร้ายที่สุด ตอนปลายคริสต์ศตวรรษที่ 16 มีการศึกษาอย่างมากในเรื่องกายวิภาค พยาธิวิทยา มีการค้นพบระบบการไหลเวียนโลหิตและเริ่มศึกษาถึงระบบน้ำเหลือง

ความคิดเรื่อง black bile เริ่มลดน้อยลงในคริสต์ศตวรรษที่ 17-18 บุคคลที่มีความสำคัญคือ Thomus Bartholin (ค.ศ. 1616-1630) เป็นคนแรกที่อธิบายเกี่ยวกับระบบน้ำเหลือง และยังพบเซลล์มะเร็งในทางเดินน้ำเหลืองด้วย อีกท่านที่รู้จักกันดีคือ Antoni van Leeuwenhook (ค.ศ. 1632-1723) คิดประดิษฐ์กล้องจุลทรรศน์เป็นการเปิดการศึกษาทางเนื้อเยื่อ ซึ่งในศตวรรษที่ 18 ความรู้ได้เจริญอย่างมาก โดยผลงานของหลายท่าน เช่น Alfred Velpeau (ค.ศ. 1835), Hermann lebert (ค.ศ. 1851), Johannes Muller (ค.ศ. 1801-1858), Rudolf Virchow (ค.ศ. 1821-1902) มีการศึกษาความแตกต่างของเซลล์มะเร็ง พฤติกรรมของเซลล์มะเร็ง การกระจายของโรคแม้ความเชื่อของ black bile หมดไป การรักษาก็เปลี่ยนไปด้วย มีการทำผ่าตัดเต้านมมากขึ้น มีการใช้ความร้อนจี้ แต่อย่างไรก็ตามการผ่าตัดสมัยนั้นมีความยากลำบาก เพราะไม่มีการดมยาสลบ ภาวะแทรกซ้อนสูงจากการติดเชื้อหลังผ่าตัดมีการค้นคว้าที่น่าสนใจอันหนึ่งในสมัยนี้ โดยนักสถิติทางการแพทย์และศัลยกรรมแพทย์ชาวอิตาลีชื่อ Domenico Antonio Rigoni-Stern ได้รวบรวมข้อมูลของผู้ป่วยจำนวน 150,673 ราย ในระยะเวลา 80 ปี ระหว่างปี ค.ศ. 1760-1839 และสรุปว่า

1. อัตราการเกิดโรคมะเร็งเพิ่มตามอายุ
2. มะเร็งเต้านมมีความสัมพันธ์กับมะเร็งมดลูก
3. สตรีที่ยังไม่แต่งงาน มีโอกาสเป็นมะเร็งเต้านมมากกว่าสตรีที่แต่งงาน

การผ่าตัดเริ่มมีบทบาทอย่างจริงจังในการรักษาโรคมะเร็งเต้านมเมื่อศตวรรษที่ 19 ทั้งนี้เพราะมีการนำเอาการดมยาสลบมาใช้ในการผ่าตัดในปี ค.ศ. 1846 โดย William Thomus Green Morton ในเมือง Boston โดยใช้ sulphuric ether หนึ่งปีหลังจากนั้นมีการใช้กันอย่างแพร่หลาย อีกสิ่งหนึ่งที่ทำให้การผ่าตัดก้าวหน้าไปอย่างรวดเร็วคือ การค้นพบหลักการใช้ยาฆ่าเชื้อในการผ่าตัด และการรักษาแผลผ่าตัด ซึ่งจากหลักการอันนี้ทำให้อัตราการตายจากการผ่าตัดเต้านมในปี ค.ศ. 1877 ลดลงจาก 15% เป็น 5.8% การรักษาด้วยการผ่าตัดจึงเป็นการรักษาหลักของโรคมะเร็งเต้านม

เรื่อยมา ทำให้มีการคิดวิธีการผ่าตัดแบบต่าง ๆ ผลการผ่าตัดในระยะนั้นพบว่า อัตราการมีชีวิตอยู่ 10 ปีประมาณ 10% ซึ่งดีกว่าการรักษาด้วยวิธีอื่นอย่างมาก

2.1.2 โรคมะเร็งเต้านมในปัจจุบัน

มะเร็งเต้านม (breast cancer) เกิดจากความผิดปกติของเซลล์ที่อยู่ภายในท่อน้ำนมหรือต่อมน้ำนม ซึ่งเซลล์เหล่านี้มีการแบ่งตัวผิดปกติจนไม่สามารถควบคุมได้ มักแพร่กระจายไปตามทางเดินน้ำเหลืองไปสู่อวัยวะที่ใกล้เคียง เช่น ต่อมน้ำเหลืองที่รักแร้ หรือแพร่กระจายไปสู่อวัยวะที่อยู่ห่างไกล เช่น กระดูก ปอด ตับ และสมอง เช่นเดียวกับมะเร็งชนิดอื่น ๆ เมื่อเซลล์มะเร็งมีจำนวนมากขึ้นก็จะแย่งสารอาหารและปล่อยสารบางอย่างที่เป็นอันตรายและทำลายอวัยวะต่าง ๆ จนทำให้ผู้ป่วยเสียชีวิตในที่สุด (Ministry of Public Health, 2005) มะเร็งเต้านมเริ่มเกิดขึ้นที่ส่วนของเยื่อบุท่อน้ำนมส่วนปลายในเต้านมที่มีการเจริญเติบโตผิดปกติ (ทิพพารัตน์ แสนคาร และธนัช กนกเทศ, 2562) จากรายงานทะเบียนมะเร็งระดับโรงพยาบาล ฉบับที่ 32 ของสถาบันมะเร็งแห่งชาติ พบว่า ในปี 2559 มีผู้ป่วยมะเร็งเต้านมรายใหม่ 849 คนเป็นเพศหญิงจำนวน 842 คนและเพศชายจำนวน 7 คน (สถาบันมะเร็งแห่งชาติ, 2561) แสดงให้เห็นว่าโรคมะเร็งเต้านมพบมากในเพศหญิง และสามารถพบได้ในเพศชายด้วยเช่นเดียวกัน (Fentiman, I., 2009)

2.1.3. ปัจจัยที่เป็นสาเหตุของการเกิดโรคมะเร็ง

เป็นที่ยอมรับว่าปัจจัยที่เป็นต้นเหตุของการเกิดโรคมะเร็งยังไม่เป็นที่แน่ชัดชัด แต่เชื่อกันว่าเกิดได้หลายสาเหตุ (multifactorial) ร่วมกันแยกเป็นสาเหตุจากปัจจัยภายในร่างกาย และปัจจัยภายนอกร่างกาย (อุปมา เลี้ยงสว่างวงศ์, 2541)

1. ปัจจัยภายในร่างกาย (endogenous factors)

ปัจจัยภายในร่างกายที่เป็นสาเหตุของการเกิดมะเร็ง ซึ่งจะพบแตกต่างกันไปในแต่ละบุคคล นั้นได้แก่ ระบบภูมิคุ้มกัน กล่าวคือในสภาวะปกติร่างกายจะมีภูมิต้านทานสิ่งแปลกปลอมเนื้องอกถือเป็นสิ่งแปลกปลอมประเภทหนึ่ง ร่างกายจึงสามารถสร้างภูมิต้านทานต่อเนื้องอกได้ ทำให้สามารถทำลายเซลล์มะเร็งอันเป็นจุดเริ่มต้นของการเกิดมะเร็งตามมาได้ แต่ถ้าระบบภูมิคุ้มกันอ่อนแอลงหรือเกิดความบกพร่องในการทำหน้าที่ เช่น อาจได้รับสารเคมี/ยา ที่กดการทำงานของระบบภูมิคุ้มกันหรือได้รับผลกระทบจากสิ่งแวดล้อม การติดเชื้อที่ทำลายระบบภูมิคุ้มกัน เมื่อนั้นร่างกายจะสูญเสียระบบรักษาความปลอดภัยของตัวเองปล่อยให้เซลล์มะเร็งซึ่งเป็นสิ่งแปลกปลอมเจริญเติบโตอย่างอิสระจนกลายเป็นก้อนมะเร็ง นอกจากนี้ปัจจัยภายในร่างกายที่เกี่ยวข้องกับการเกิดมะเร็ง ยังหมายรวมถึงความแตกต่างของเชื้อชาติ มะเร็งหลายชนิดมีอุบัติการณ์ที่แตกต่างกันในระหว่างเชื้อชาติ เช่น มะเร็ง

ตับพบมากในคนเอเชีย มะเร็งโพรงจมูกพบมากในแถบเอเชียตะวันออกเฉียงใต้ มะเร็งกระเพาะอาหารพบมากในคนญี่ปุ่น เป็นต้น มะเร็งบางชนิดพบแตกต่างกันในเพศต่างกัน เช่น มะเร็งตับ มะเร็งปอด มะเร็งโพรงจมูก พบมากในเพศชาย ขณะเดียวกันมะเร็งในช่องปาก มะเร็งเต้านม มะเร็งผิวหนัง พบมากในเพศหญิง ความต่างในเรื่องอายุเป็นสาเหตุหนึ่งเช่นเดียวกัน กล่าวคือพบมะเร็งต่างชนิดกันในวัยที่ต่างกัน เช่น มะเร็งลูกตาพบเฉพาะในเด็ก เป็นต้น อย่างไรก็ตามปัจจัยภายในร่างกายที่เกี่ยวข้องกับการเกิดโรคมะเร็งที่นับว่าสำคัญที่สุดคือ ความต่างกันทางกรรมพันธุ์ (genetic makeup) มีมะเร็งหลายชนิดที่มีความสัมพันธ์ใกล้ชิดกับกรรมพันธุ์ เช่น มะเร็งเต้านม มะเร็งรังไข่ มะเร็งลำไส้ใหญ่ หรือมะเร็งลูกตา กล่าวคือถ้ามีบุคคลใดในครอบครัวเป็นมะเร็งอย่างหนึ่งอย่างใดดังกล่าวแล้ว บุคคลในครอบครัวเดียวกันหรือเครือญาติใกล้ชิด มีโอกาสที่จะเป็นมะเร็งชนิดนั้น ๆ ด้วยเช่นกัน

2. ปัจจัยภายนอกร่างกาย (exogenous factors)

นอกจากปัจจัยภายในร่างกายแล้วสาเหตุของการเกิดโรคมะเร็งยังมีส่วนเกี่ยวข้องกับสาเหตุจากปัจจัยภายนอกร่างกายหรือสาเหตุจากสิ่งแวดล้อมปัจจัยสิ่งแวดล้อมที่มีส่วนเกี่ยวข้องกับการเกิดโรคมะเร็งมีดังนี้

- สารกายภาพ (physical agent) มีสารกายภาพหลากหลายชนิดที่มีส่วนเสริมให้เกิดมะเร็งได้ เช่น รังสี การได้รับรังสีไม่ว่าจะเป็นรังสีเอ็กซ์ รังสีอัลตราไวโอเลต หรือรังสีในสนามแม่เหล็ก แม้ปริมาณน้อย ๆ แต่ในระยะเวลายาวนานมีส่วนทำให้เกิดโรคมะเร็งในอวัยวะเกือบทุกระบบของร่างกายได้ การระคายเคืองเฉพาะที่ เช่น การมีฟันเกหรือฟันปลอม โดยฟันเกทำให้เกิดการครูดกับเนื้อเยื่อในช่องปากขณะเคี้ยวอาหารเป็นสาเหตุของมะเร็งในช่องปาก เป็นต้น

- สารเคมี (chemical agent) สารเคมีมีอยู่ทั่วไปในบรรยากาศ น้ำ และสิ่งแวดล้อมรอบตัวเราโดยเฉพาะในปัจจุบันที่มีภาวะสิ่งแวดล้อมที่สะอาดบริสุทธิ์กำลังจะหมดสภาพลงไปทุกที จนอาจประเมินได้ว่าคนเราอยู่ท่ามกลางสิ่งแวดล้อมที่เป็นพิษ ในกรณีที่ก่อให้เกิดโรคมะเร็งสารเคมีดังกล่าวถือเป็นสารก่อมะเร็ง (Carcinogen)

- ไวรัส (Virus) ในปัจจุบันแม้ยังไม่มีการยืนยันแน่ชัดว่า ไวรัสสามารถทำให้เกิดโรคมะเร็งได้ในคน แต่มีไวรัสทั้งชนิดไวรัสดีเอ็นเอและไวรัสอาร์เอ็นเอหลายชนิด พบว่าเป็นสาเหตุที่ทำให้เกิดโรคมะเร็งในสัตว์ทดลอง รวมทั้งจากหลักฐานผลการวิจัยที่พบว่าไวรัสมีส่วนเกี่ยวข้องกับการเกิดมะเร็งบางชนิดในคน เช่น ไวรัสเอ็บสเตนบาร์ (Epstein Barr virus) พบว่ามีส่วนสัมพันธ์กับการเกิดโรคมะเร็งในโพรงจมูกและมะเร็งต่อมน้ำเหลืองเบอร์คิตท์ (Burkitt's Lymphoma) เป็นต้น ทำให้เชื่อได้ว่าไวรัสน่าจะมีส่วนเกี่ยวข้องกับการเกิดโรคมะเร็งในคน แม้จะไม่ใช่สาเหตุโดยตรงแต่น่าจะเป็นสาเหตุเสริมให้เกิดการนำไปสู่การเกิดโรคมะเร็ง

- สารพิษ (toxin) สารพิษที่คนเราได้รับมักปะปนมากับอาหารที่รับประทานเข้าไป เช่น สารอะฟลาท็อกซิน (Aflatoxon) จากเชื้อรา Aspergillus สารตัวนี้เป็นสาเหตุของการเกิดโรคมะเร็งตับ ซึ่งมีหลักฐานยืนยันมากมาย พบได้ในอาหารประเภทถั่วต่าง ๆ โดยเฉพาะถั่วลิสง

- พยาธิ (parasite) พยาธิบางชนิดเป็นบ่อเกิดของการเกิดโรคมะเร็งได้เช่นเดียวกัน เช่น พยาธิที่กระเพาะปัสสาวะ เป็นสาเหตุของโรคมะเร็งกระเพาะ เป็นต้น

- ภาวะขาดอาหารหรือทุพโภชนาการ (malnutrition) ภาวะขาดอาหารประเภทโปรตีนทำให้เกิดโรคตับแข็ง และการเกิดภาวะตับแข็งเป็นสาเหตุหนึ่งต่อการเกิดโรคมะเร็งตับ ส่วนทุพโภชนาการเกี่ยวข้องกับการเกิดโรคมะเร็ง โดยที่ร่างกายได้รับสารอาหารไม่ครบถ้วน เช่น ขาดวิตามิน ได้แก่ วิตามินเอ หรือ วิตามินซี ซึ่งพบว่ามีส่วนเสริมให้ร่างกายเสี่ยงต่อการเกิดโรคมะเร็งได้

- ตัวแปรอื่น ๆ นอกจากตัวแปรข้างต้นที่กล่าวมายังมีตัวแปรด้าน พฤติกรรมสุขภาพจิต และรูปแบบการดำรงชีวิตเป็นสาเหตุของการเกิดโรคมะเร็งได้เช่นกัน โดยมีหลักฐานพอสรุปได้ว่า การเกิดมะเร็งมีส่วนเกี่ยวข้องกับตัวแปรดังกล่าว รวมถึงตัวแปรทางด้านขนบธรรมเนียมประเพณี และภาวะทางเศรษฐกิจ สังคมของแต่ละบุคคลอีกด้วย

ในงานนี้กล่าวถึงโรคมะเร็งเต้านม ซึ่งสาเหตุของการเกิดมะเร็งเต้านม (จิณพิชญ์ชา มะมม, 2555) จากการศึกษาที่ผ่านมา พบว่า เกิดจากพันธุกรรมที่มีญาติสายตรงโดยเฉพาะมารดา พี่สาว หรือน้องสาวที่เป็นมะเร็งเต้านม, ผู้ที่มีอายุตั้งแต่ 40 ปีขึ้นไป, การใช้ฮอร์โมนหรือยาคุมกำเนิดตั้งแต่อายุยังน้อยและเป็นเวลานาน ๆ, ความอ้วนโดยเฉพาะในช่วงหลังหมดประจำเดือน, การเข้าสู่ระยะหมดประจำเดือนก่อนอายุ 45 ปี, มีบุตรคนแรกอายุมากกว่า 30 ปี, ผู้ที่มีประวัติเป็น benign breast disease บางโรค เช่น fibrocystic disease ที่มี proliferative pattern หรือมี atypical hyperplasia, คนที่ไม่มีบุตร, ประวัติการมีประจำเดือนครั้งแรกอายุน้อยกว่า 12 ปี, การสูบบุหรี่, การดื่มสุรา และหมดประจำเดือนอายุมากกว่า 55 ปี ซึ่งในปัจจุบันมีการศึกษาวิจัยเกี่ยวกับปัจจัยส่งเสริมในการเกิดโรคและการแพร่กระจายของโรคมะเร็งเต้านมมากมาย เช่น

- สตรีที่มีประวัติในครอบครัวเป็นมะเร็งรังไข่หรือมะเร็งเต้านม และตรวจพบความผิดปกติของยีนโดยเฉพาะยีน BRCA1 และยีน BRCA2 จะพบว่ามีโอกาสเป็นมะเร็งเต้านมก่อนอายุ 50 ปี

- สตรีที่เคยเป็นมะเร็งเต้านมมาแล้ว 1 ข้าง มีโอกาสเกิดมะเร็งเต้านมอีกข้างหนึ่งสูงกว่าคนทั่วไป และพบว่าสตรีที่มีประวัติเป็นมะเร็งลำไส้ จะมีโอกาสเกิดมะเร็งรังไข่และต่อเนื่องถึงโอกาสการเป็นมะเร็งเต้านมสูงกว่าอีกด้วย

- เคยได้รับรังสีบริเวณทรวงอก (thoracic irradiation) เพื่อการรักษามะเร็งชนิดอื่น ๆ เช่น Hodgkin disease หรือ Non-Hodgkin lymphoma

- สตรีที่อยู่ในระหว่างการตั้งครรภ์ แม้พบได้ไม่บ่อยครั้ง แต่ในรอบ 10 ปีที่ผ่านมาพบมากขึ้นเรื่อย ๆ (ปัทมา พรพึ่งบุญ, 2555)

- อายุที่มากขึ้น โดยทั่วไปความเสี่ยงจะเพิ่มขึ้นหลังอายุ 45 ปี แต่ก็สามารถพบได้ในคนที่อายุน้อยกว่า 45 ปีด้วยเช่นกัน (พิมพ์จุฑา วราทรัพย์, จิราวัลย์ จิตรถเวช และวิชิต หล่อจีระชุณห์กุล, 2561)

2.1.4. ระยะของมะเร็งเต้านม

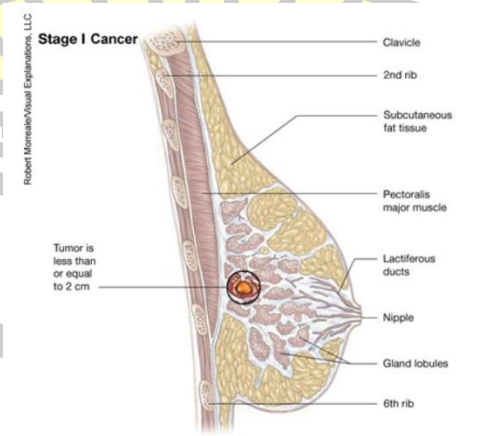
โดยทั่วไปมะเร็งเต้านมมี 4 ระยะ คือระยะที่ 1-4 แต่ในปัจจุบันนี้เรามักจะเพิ่มระยะที่ “ศูนย์” เข้าไปด้วย รวมแล้วเป็น 5 ระยะ คือ ระยะที่ 0-4 (หะสัน มูหาหมัด, 2555) เนื่องจากปัจจุบันผู้หญิงมีการตรวจเช็คมะเร็งเต้านมบ่อยมากขึ้น ทำให้พบว่ามีผู้ป่วยจำนวนหนึ่งที่เซลล์มะเร็งเต้านมเพิ่งจะก่อตัวขึ้นยังไม่ได้ลุกลามออกไปจากจุดกำเนิด ซึ่งสามารถรักษาได้ง่ายโดยการผ่าตัดเอาบริเวณนั้นออกไป มีโอกาสหายขาดสูงมากเกือบ 100 % ยกตัวอย่างในประเทศสหรัฐอเมริกา ซึ่งเป็นประเทศที่มีการรณรงค์ให้ประชาชนมีการตรวจเช็คมะเร็งเต้านมกันอย่างกว้างขวาง ตั้งแต่ทศวรรษ 80 เป็นต้นมา ทำให้ปัจจุบันนี้พบว่า ราว 1 ใน 4 หรือ 25% ของผู้ป่วยมะเร็งเต้านมจะยังคงอยู่ในระยะที่ 0 และให้ผลการรักษาที่ดีมากเรียกได้ว่าหายขาดเกือบ 100% (คณะแพทยศาสตร์ ศิริราชพยาบาล, 2555)

- มะเร็งเต้านมระยะที่ 0 หรือ DCIS (ductal carcinoma in situ)

เซลล์มะเร็งเพิ่งก่อตัวขึ้นจำกัดอยู่ภายในเนื้อเยื่อฐานรากยังไม่แบ่งตัวลุกลามสู่ภายนอกขอบเขต โอกาสอยู่รอดที่ 5 ปี ประมาณ 99%

- มะเร็งเต้านมระยะที่ 1

มะเร็งมีการลุกลามออกมานอกเนื้อเยื่อฐานราก แต่ยังไม่มีการแพร่กระจายไปสู่ต่อมน้ำเหลืองที่รักแร้ และขนาดก้อนมะเร็งไม่เกิน 2 ซม. โอกาสอยู่รอดที่ 5 ปี ประมาณ 98%

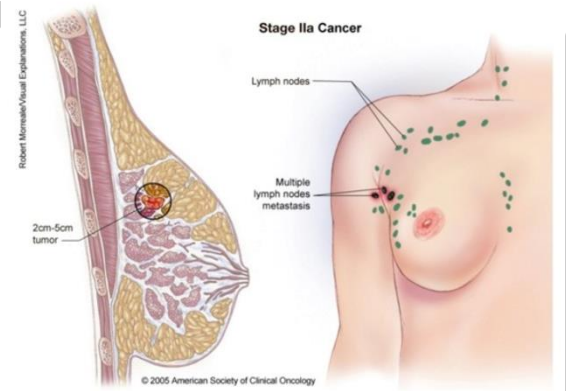


ที่มา: http://www.thaibreastcancer.com/ca-131/

รูปที่ 1 มะเร็งเต้านมระยะที่ 1

- มะเร็งเต้านมระยะที่ 2

ก้อนมะเร็งขนาดเกิน 2 ซม.แต่ไม่เกิน 5 ซม.ที่ยังไม่มีการแพร่กระจายไปสู่ต่อมน้ำเหลืองที่รักแร้ หรือมะเร็งขนาดเล็กไม่เกิน 2 ซม. แต่มีการแพร่กระจายไปสู่ต่อมน้ำเหลืองที่รักแร้แล้ว ถ้ามะเร็งยังไม่เข้าต่อมน้ำเหลือง โอกาสอยู่รอดที่ 5 ปี ประมาณ 98%

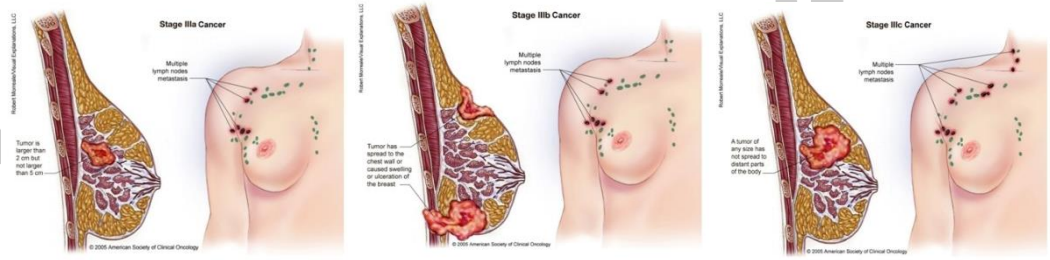


ที่มา: http://www.thaibreastcancer.com/ca-131/

รูปที่ 2 มะเร็งเต้านมระยะที่ 2

- มะเร็งเต้านมระยะที่ 3

ก้อนมะเร็งมีขนาดใหญ่กว่า 5 ซม. แพร่กระจายไปยังต่อมน้ำเหลืองที่รักแร้ข้างเดียวกันอย่างมาก จนทำให้ต่อมน้ำเหลืองเหล่านั้นมารวมติดกันเป็นก้อนใหญ่หรือติดแน่นกับอวัยวะข้างเคียงโอกาสอยู่รอดที่ 5 ปี ประมาณ 84 %



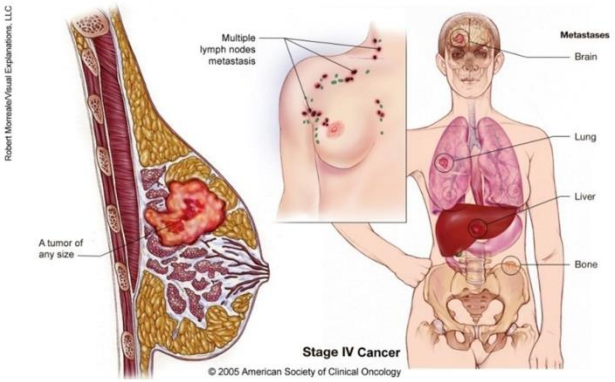
ที่มา: http://www.thaibreastcancer.com/ca-131/

(a) มะเร็งเต้านมระยะที่ IIIa (b) มะเร็งเต้านมระยะที่ IIIb (c) มะเร็งเต้านมระยะที่ IIIc

รูปที่ 3 มะเร็งเต้านมระยะที่ 3

- มะเร็งเต้านมระยะที่ 4

ก้อนมะเร็งมีขนาดโตเท่าไหร่ก็ได้ แต่พบว่ามีการแพร่กระจายไปยังส่วนอื่นของร่างกายที่อยู่ไกลออกไป เช่น กระดูก ปอด ตับ หรือสมอง เป็นต้น โอกาสอยู่รอดที่ 5 ปี ประมาณ 24 %



ที่มา: http://www.thaibreastcancer.com/ca-131/

รูปที่ 4 มะเร็งเต้านมระยะที่ 4

2.1.5. การรักษาโรคมะเร็งเต้านม

วิธีการรักษามะเร็งเต้านมที่ได้ผลดีและเป็นที่ยอมรับกันในปัจจุบันมีอยู่ 5 วิธี คือ

5.1 การรักษาโดยการผ่าตัด

เป็นวิธีการรักษาหลักสำหรับผู้ป่วยมะเร็งเต้านมระยะเริ่มแรก ซึ่งมีประโยชน์ในการควบคุมโรคและสามารถนำชิ้นเนื้อที่ได้จาการผ่าตัดไปตรวจทางพยาธิวิทยา ทำให้ทราบระยะที่แท้จริงของโรค ช่วยวางแผนการรักษาที่เหมาะสมและสามารถพยากรณ์โรคได้แม่นยำมากขึ้น ขั้นตอนการผ่าตัดรักษามะเร็งเต้านม แบ่งออกเป็น 2 ส่วน คือ การผ่าตัดที่เต้านมและการผ่าตัดต่อมน้ำเหลืองที่รักแร้ นอกจากนี้ยังมีส่วนเพิ่มเติมซึ่งไม่ใช่การรักษาโดยตรง เช่น การเสริมสร้างเต้านมใหม่เพื่อเพิ่มคุณภาพชีวิตที่ดีแก่ผู้ป่วย

5.2 การใช้ยาเคมีบำบัด (chemotherapy)

เป็นการรักษาโดยการใช้ยาที่มีฤทธิ์ทำลายหรือยับยั้งการเจริญเติบโตของเซลล์มะเร็ง โดยออกฤทธิ์ทั่วร่างกายต่างจากการผ่าตัดซึ่งให้ผลเฉพาะที่ จึงเป็นการเพิ่มประสิทธิภาพในการรักษามะเร็งเต้านมที่หลงเหลือหรือมีการหลุดรอดไปยังระบบอื่น ๆ ช่วยให้มีโอกาสหายขาดและมีชีวิตที่ยืนยาวขึ้น อย่างไรก็ตามยาเคมีบำบัดนี้ นอกจากจะออกฤทธิ์ทำลายเซลล์มะเร็งแล้ว ก็อาจมีผลต่อเซลล์ปกติที่มีการแบ่งตัวอย่างรวดเร็ว เช่น ไขกระดูก ผม และขนตามร่างกาย ระบบสืบพันธ์และเยื่อบุทางเดินอาหาร ทำให้เกิดอาการข้างเคียงจากการใช้ยาได้

5.3 การรักษามะเร็งเต้านมโดยการฉายแสง (radiation therapy)

เป็นการรักษาด้วยการใช้รังสีที่มีพลังงานสูงเพื่อหยุดยั้งการเจริญเติบโตและการแบ่งตัวของเซลล์มะเร็ง มักใช้การฉายแสงร่วมในกรณีที่ผู้ป่วยผ่าตัดแบบสงวนเต้า หรือผู้ป่วยที่มีมะเร็งลุกลามมาที่ผิวหนังหรือกล้ามเนื้อหน้าอก

5.4 การรักษามะเร็งเต้านมโดยการใช้ยาต้านฮอร์โมน (hormonal therapy)

เป็นการรักษาโดยการให้ยาในผู้ป่วยมะเร็งเต้านมชนิดที่มีตัวรับฮอร์โมนเพศ ออกฤทธิ์ทำให้เซลล์มะเร็งขาดฮอร์โมนที่เป็นตัวกระตุ้นการเจริญเติบโต ลดอัตราการกลับมาเป็นซ้ำ ซึ่งยากลุ่มนี้ส่งผลในการยับยั้งการเจริญเติบโต และทำลายเซลล์มะเร็งเต้านมชนิดที่มีตัวรับฮอร์โมน ซึ่งมีผลต่อเซลล์มะเร็งทั่วร่างกายเช่นเดียวกับการให้ยาเคมีบำบัด

5.5 การรักษามะเร็งเต้านมโดยการใช้ยาแบบพุ่งเป้า (targeted therapy)

ยาในกลุ่มนี้จัดเป็นยากลุ่มใหม่ เช่น ยาต้านเฮอร์ทู (HER2) ซึ่งมีกลไกการออกฤทธิ์แตกต่างจากยากลุ่มเดิม ๆ กล่าวคือ เซลล์มะเร็งเต้านมบางชนิดจะมีตัวรับสัญญาณเฮอร์ทูอยู่ที่ผิวเซลล์ ทำให้สามารถใช้ยาดังกล่าวเพื่อจับกับตัวรับสัญญาณเหล่านี้และให้ยาออกฤทธิ์ทำลายเซลล์มะเร็งดังกล่าวได้ ดังนั้นเซลล์อื่น ๆ ที่ไม่มีตัวรับสัญญาณก็จะไม่ได้รับผลกระทบจากยากลุ่มนี้

2.2 การทำเหมืองข้อมูล

2.2.1. เหมืองข้อมูล (data mining)

การทำเหมืองข้อมูล คือ การนำวิธีการจากการเรียนรู้ของเครื่องการรู้จำรูปแบบ (pattern recognition) สถิติ (statistics) และฐานข้อมูล (data base) เพื่อสกัดข้อมูลที่มีประโยชน์จากฐานข้อมูลขนาดใหญ่ได้ (Cabena, P. et al., 1998) หรือในอีกความหมายหนึ่งคือ การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ หรือที่เรียกกันว่าการทำเหมืองข้อมูลเป็นวิธีการที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์หรือทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้นซึ่งการค้นหาความรู้และความจริงที่แฝงอยู่ในข้อมูล (knowledge discovery) หรือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ๆ เป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่มีอยู่ซึ่งต่างจากระบบฐานข้อมูล (database system) ตรงที่การทำเหมืองข้อมูลไม่ต้องกำหนดคำสั่ง เช่น structured query language (SQL) เพื่อค้นหาข้อมูลที่ต้องการ แต่ระบบการทำเหมืองข้อมูลจะมีวิธีการซึ่งโดยปกติจะเป็นเครื่องมือในการเรียนรู้กลไกต่าง ๆ (machine learning tools) เพื่อทำหน้าที่นั่นคือแค่บอกว่าต้องการอะไร (what to be mined) แต่ไม่จำเป็นต้องระบุว่าทำอย่างไร (how to mine) ระบบฐานข้อมูลโดยทั่วไปจะบังคับให้ต้องทำทั้งสองหน้าที่นี้คือคิดก่อนว่าจะค้นหาอะไรแล้วก็ไปกำหนดคำสั่ง SQL เพื่อค้นหาข้อมูลนั้น ดังนั้นถ้าคิดไม่รอบคอบหรือคิดดีแล้วแต่แปลเป็นคำสั่งผิดก็จะได้ข้อมูลผิด ๆ หรือไม่ตรงตามความต้องการนั่นเอง (Chapman, P. et al., 2000)

2.2.2. วิวัฒนาการในการทำเหมืองข้อมูล (Revolution of data Mining)

ในปี ค.ศ. 1960 เทคโนโลยีฐานข้อมูลได้เริ่มพัฒนาจากไฟล์ประมวลผลข้อมูล (file processing) พื้นฐาน ค้นคว้าและพัฒนาระบบฐานข้อมูลมาเรื่อย ๆ

ในปี ค.ศ. 1970 ได้นำไปสู่การพัฒนาระบบการจัดเก็บข้อมูลในรูปแบบตาราง (relation database system) มีเครื่องมือจัดการตัวแบบข้อมูลและมีวิธีใช้ดัชนีและการบริหารข้อมูล นอกจากนี้ผู้ใช้ยังได้รับความสะดวกในการเข้าถึงข้อมูลโดยการใช้ภาษาในการเรียกใช้ข้อมูล (query language)

ในปี ค.ศ. 1980 เทคโนโลยีฐานข้อมูลได้เริ่มมีการปรับปรุงและพัฒนาในการหาระบบจัดการที่มีศักยภาพมากขึ้น ความก้าวหน้าเทคโนโลยีฮาร์ดแวร์ใน 30 ปีที่ผ่านมาได้นำไปสู่การจัดเก็บข้อมูลจำนวนมากที่มีความซับซ้อนได้อย่างมีประสิทธิภาพเพิ่มขึ้น

ในปี 1990 Data Warehouse and Decision Support คือการนำข้อมูลมาเก็บลงในฐานข้อมูลขนาดใหญ่ครอบคลุมการใช้งานทั้งหมดขององค์กรเพื่อช่วยสนับสนุนการตัดสินใจ

ในปี ค.ศ. 2000 ถึงปัจจุบันสามารถจัดเก็บข้อมูลได้หลายรูปแบบ แตกต่างกันทั้งระบบปฏิบัติการหรือการจัดเก็บฐานข้อมูล ซึ่งการนำข้อมูลทั้งหมดจากหลายแหล่ง หลายช่วงเวลามารวมกันและจัดเก็บไว้ในรูปแบบเดียวกันเรียกว่า คลังข้อมูล (data warehouse) เพื่อความสะดวกในการจัดการต่อไป ซึ่งเทคโนโลยีคลังข้อมูลรวมไปถึงการทำความสะอาดข้อมูล (data cleaning) การวิเคราะห์ข้อมูลในลักษณะผสมผสาน (data integration) และการประมวลผลข้อมูลเชิงวิเคราะห์แบบออนไลน์ (on-line analytical processing : OLAP) เป็นวิธีการวิเคราะห์ข้อมูลในหลาย ๆ มิติ

การเจริญเติบโตอย่างรวดเร็วของข้อมูลจำนวนมากที่สะสมไว้ในฐานข้อมูลขนาดใหญ่มาก ซึ่งเกินกว่าที่กำลังควรจะสามารถจัดการได้ เป็นผลทำให้มีความจำเป็นที่จะต้องมีเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลและหาความเป็นไปได้ของข้อมูลทั้งหมดที่เป็นประโยชน์ ซึ่งก็คือ “การทำเมืองข้อมูล”

2.2.3. ประเภทข้อมูลที่นำมาทำเหมืองข้อมูล

1. ฐานข้อมูลตาราง (relational database) เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดยตัวแบบความสัมพันธ์ที่มีอยู่จริง (entity-relationship model)

2. คลังข้อมูล (data warehouse) เป็นฐานข้อมูลที่เก็บรวบรวมข้อมูลจากหลายแหล่งหลายช่วงเวลามาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในแหล่งเดียวกัน

3. ฐานข้อมูลรายการ (transaction database) เป็นฐานข้อมูลที่แต่ละรายการแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงินจะเก็บข้อมูลในรูปชื่อลูกค้าได้รายการสินค้าที่ลูกค้านั้นซื้อ

4. ฐานข้อมูลขั้นสูง (advance database) เป็นฐานข้อมูลที่จะกินอยู่ในรูปแบบอื่น ๆ เช่น ข้อมูลเชิงวัตถุ (object-oriented) ข้อมูลไฟล์ตัวอักษร (text file) ข้อมูลมัลติมีเดียและข้อมูลในรูปของเว็บ

2.2.4. ประเภทของการทำเมืองข้อมูล

อัลกอริทึมของการประเมินข้อมูลสามารถแบ่งเป็น 2 ประเภท

1. การสร้างตัวแบบในการทำนาย (predictive modeling) หรือเรียกว่าการเรียนรู้แบบมีผู้สอน (supervised learning) คือ การนำข้อมูลในอดีตมาสร้างตัวแบบเพื่อทำการทำนายอนาคตโดยมีการใช้ข้อมูลฝึกหัด (training data) ซึ่งข้อมูลทุกตัวจะมีคุณสมบัติที่ใช้ในการทำนาย อัลกอริทึมประเภทนี้จะมุ่งเน้นในการแบ่งแยกข้อมูลออกเป็นกลุ่มตามค่าคุณสมบัติของข้อมูล ซึ่งค่าคุณสมบัติของข้อมูลมีค่าไม่ต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การจำแนกหรือการจำแนกประเภท (classification) แต่ถ้าคุณสมบัติของข้อมูลมีค่าต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การถดถอย (regression) หรือการพยากรณ์ (forecasting)

2. การสร้างตัวแบบในการพรรณนาหรือบรรยายหรืออธิบาย (descriptive modeling) หรือเรียกว่าการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) คือ การนำข้อมูลที่มีอยู่มาศึกษาเพื่อหากฎความสัมพันธ์ (association) หรือการจัดกลุ่ม (clustering) ซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย เช่น การจัดกลุ่ม, โครงข่ายโคโฮเนน และกฎความสัมพันธ์

2.2.5. เทคนิคการทำเหมืองข้อมูล

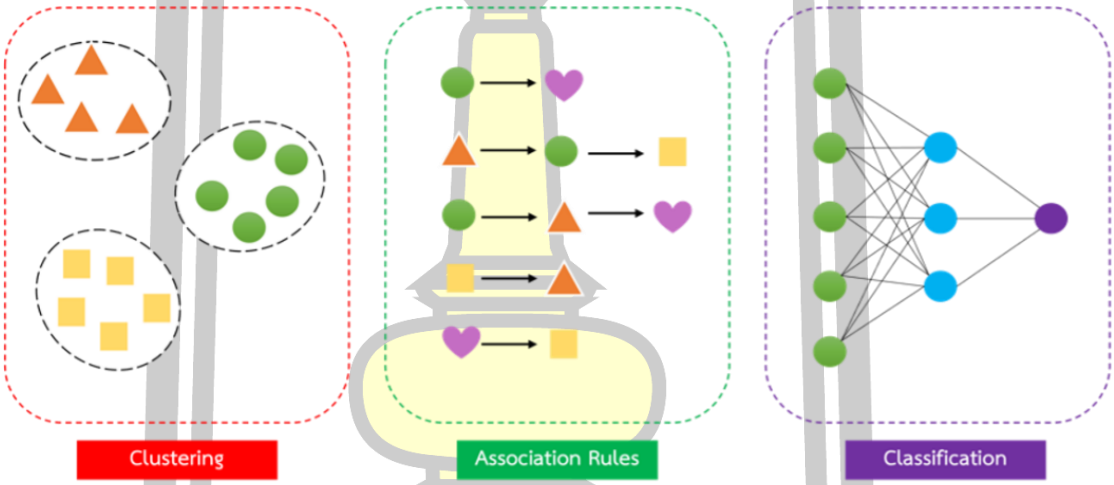
การแก้ปัญหาของงานชนิดต่าง ๆ โดยวิธีการ Data Mining ในแต่ละงานก็จะมีเทคนิคที่นำมาใช้ได้อย่างเหมาะสม โดยเทคนิค Data Mining นั้นมีหลายหลากส่วนใหญ่มาจากศาสตร์ AI (artificial intelligence) หรือศาสตร์อื่น ๆ แบ่งเป็น 3 เทคนิค ในการทำเหมืองข้อมูล ดังนี้

1) Clustering คือการจัดกลุ่มข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภท แต่จะไม่เหมือนกันโดยการแบ่งประเภทจะวิเคราะห์ข้อมูลตามต้นแบบ แต่สำหรับการแบ่งกลุ่มเป็นการวิเคราะห์โดยไม่พิจารณาจัดกลุ่มตามประเภทที่มีหรือที่รู้จัก แต่จะใช้ขั้นตอนวิธีการจัดกลุ่ม (cluster) เพื่อเป็นการค้นหาเพื่อแบ่งข้อมูลทั้งหมดลงในกลุ่มหรือกลุ่มย่อยที่เหมือนกัน ซึ่งความคล้ายคลึงกันของระเบียนที่อยู่ในกลุ่มเดียวกันจะมีมากที่สุด ส่วนความคล้ายคลึงกันของระเบียนที่อยู่ต่างกลุ่มกันจะมีน้อยที่สุด

2) Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูล โดยการค้นหาคุณลักษณะต่าง ๆ ที่ซื้อร่วมกันในโลกธุรกิจที่แพร่หลายส่วนใหญ่ คือการวิเคราะโดยการค้นหาคุณลักษณะต่าง ๆ ที่ซื้อร่วมกันในโลกธุรกิจที่แพร่หลายส่วนใหญ่คือการวิเคราะห์ความสัมพันธ์กัน (affinity analysis)

หรือการวิเคราะห์ตะกร้าตลาด (market basket analysis) งานของความสัมพันธ์เป็นการค้นหากฎความสัมพันธ์ระหว่างคุณลักษณะ 2 คุณลักษณะหรือมากกว่า 2 คุณลักษณะ กฎความสัมพันธ์อยู่ในรูป ถ้า สินค้าที่ซื้อก่อน (antecedent) แล้ว สินค้าที่ซื้อทีหลัง (consequent) ร่วมกับการวัดซับพอร์ท (support) และความเชื่อมั่น (confidence) ที่สัมพันธ์กับกฎนั้น

3) Classification เป็นการจัดแบ่งประเภทของข้อมูล โดยหาชุดต้นแบบหรือชุดของการทำงานที่อธิบายและแบ่งประเภทข้อมูล วัตถุประสงค์เพื่อให้สามารถใช้เป็นต้นแบบทำนายประเภทของวัตถุหรือข้อมูลที่ไม่มีการระบุประเภทหรือชนิดของข้อมูล ซึ่งต้นแบบสร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (training data) โดยอาจจะเป็นกลุ่มข้อมูลที่มีการระบุประเภทหรือกลุ่มเรียบร้อยแล้ว รูปแบบของต้นแบบแสดงได้หลายแบบเช่น Decision Trees หรือ Neural Networks เป็นต้น



รูปที่ 5 เทคนิคการทำเหมืองข้อมูล

2.2.6. การเตรียมข้อมูล (data preprocessing)

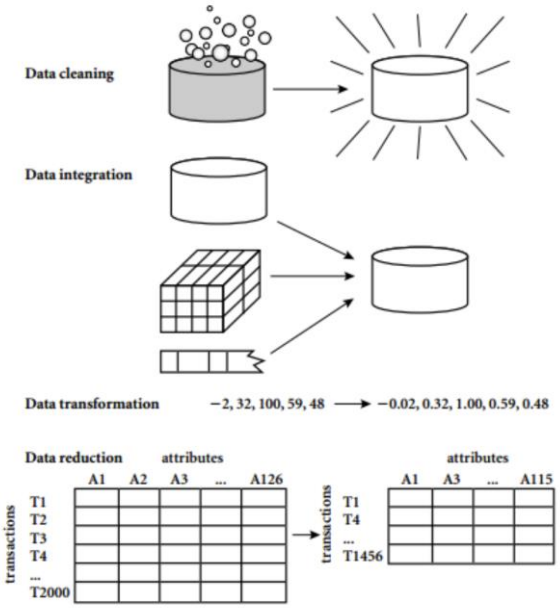
การเตรียมข้อมูล นับเป็นขั้นตอนที่สำคัญมากในกระบวนการของการทำงานด้านวิทยาการข้อมูล (Yi, M., 2019) ซึ่งถ้าการเตรียมข้อมูลทำได้ไม่ดีก็อาจจะส่งผลให้การทำงานในขั้นตอนอื่นไม่มีประสิทธิภาพตามไปด้วย โดยผลกระทบที่เกิดขึ้นร้ายแรงน้อยที่สุดอาจทำให้เสียเวลาต้องทำใหม่ หรือถ้าแย่กว่านั้นอาจส่งผลให้ผลการวิเคราะห์ หรือการตีความจากการนำข้อมูลไปใช้ผิดไปจากที่ควรจะเป็น ดังนั้นก่อนที่จะนำข้อมูลที่ได้ไปใช้งานควรมีการทำความเข้าใจและตรวจสอบคุณภาพของข้อมูล ทำข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง หรือปรับเปลี่ยนให้อยู่ในรูปแบบที่เหมาะสม เพื่อให้ข้อมูลอยู่ในลักษณะหรือรูปแบบที่สามารถนำไปใช้งานต่อในขั้นตอนอื่นอย่างมีประสิทธิภาพ (Lal, T. N. et al., 2006) ซึ่งอาจมีข้อมูลลักษณะ เช่น

- ข้อมูลไม่สมบูรณ์ (incomplete data) เช่น ค่าของคุณลักษณะขาดหาย (missing value) ขาดคุณลักษณะที่น่าสนใจหรือขาดรายละเลียดของข้อมูล

- ข้อมูลรบกวน (noisy data) เช่น ข้อมูลมีค่าผิดพลาด (error) หรือมีค่าผิดปกติ (outliers)

- ข้อมูลไม่สอดคล้อง (inconsistent data) เช่น ข้อมูลเดียวกัน แต่ตั้งชื่อต่างกัน หรือใช้ค่าแทนข้อมูลที่ต่างกัน

การตรียมข้อมูลโดยทั่วไปแบ่งการทำงานออกได้เป็น 3 ขั้นตอน ได้แก่ การทำความสะอาดข้อมูล (data cleaning) การเลือกข้อมูล (data selection) และการแปลงข้อมูล (data transformation)



ที่มา: http://www.e2matrix.com/blog/2017/10/10/data-mining/

รูปที่ 6 ตัวอย่างการเตรียมข้อมูล

2.2.7. ขั้นตอนการทำเหมืองข้อมูล

ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ ดังนี้

- Data Cleaning เป็นขั้นตอนในการกำจัดข้อมูลที่เราไม่ต้องการ หรือข้อมูลที่ไม่เป็นประโยชน์ต่อการใช้งานหรือข้อมูลที่มีความผิดพลาด

- Data Integration เป็นขั้นตอนการรวมข้อมูลทั้งหมดที่มีจากแหล่งข้อมูลต่าง ๆ มาไว้ด้วยกัน

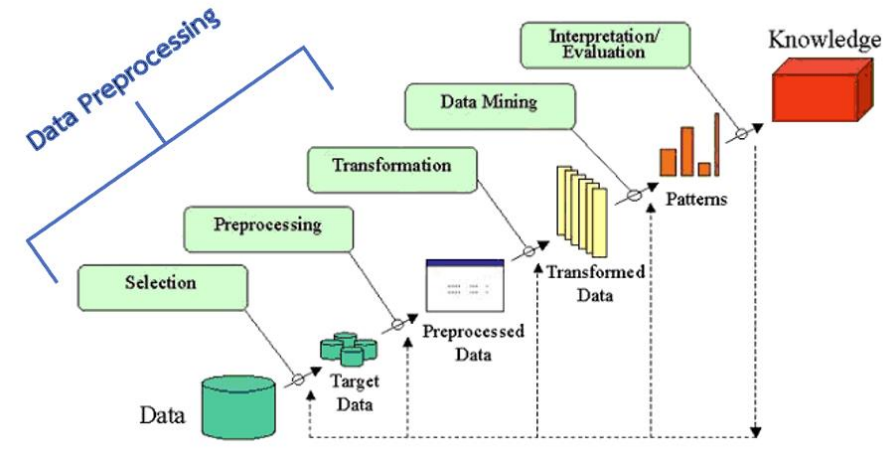
- Data Selection เป็นขั้นตอนการคัดเลือกข้อมูลที่เกี่ยวข้องกับข้อมูลที่จะใช้วิเคราะห์เลือกข้อมูลที่มีความสัมพันธ์กัน ส่งผลต่อกันและเป็นประโยชน์ต่อการทำนาย

- Data Transformation เป็นการจัดภาพแบบข้อมูลที่ได้จากขั้นตอนการคัดเลือกข้อมูลให้มีความเหมาะสมต่อการทำนาย เช่น การจัดระเบียนข้อมูลที่สัมพันธ์กันมาไว้ในระเบียนชุดเดียวกัน หรือการแปลงค่าตัวเลขให้อยู่ช่วงที่กำหนด (normalization) เป็นต้น

- Data Mining เป็นขั้นตอนที่ใช้ในการนำข้อมูลที่พร้อมแล้วมาสร้างแบบจำลอง โดยขั้นแรกจะต้องทำการเลือกเทคนิคที่เหมาะสมกับภาพชุดข้อมูลและพิจารณาปัญหา เช่น ต้องการทำนายประเภทของโรค หรือต้องการแบ่งประเภทข้อมูลของโรค หรือต้องการหาปัจจัยที่เกี่ยวข้อง เป็นต้น หลังจากได้เทคนิคที่เหมาะสมแล้วจะทำการสอน (train) ให้แบบจำลองเรียนรู้ลักษณะของข้อมูลว่าชุดข้อมูลทั้งหมดมีความสัมพันธ์กันอย่างไร และทิศทางการวิเคราะห์เป็นอย่างไร โดยการสอนให้แบบจำลองเรียนรู้นั้นจำเป็นต้องมีการกำหนดพารามิเตอร์ (parameter) หรือค่าตัวแปรต่าง ๆ ให้เหมาะสม ซึ่งในการพิจารณาค่าพารามิเตอร์นั้นขึ้นอยู่กับเทคนิคที่เลือกใช้ ประสบการณ์ในการวิเคราะห์และการลองผิดลองถูก จากนั้นจึงนำแบบจำลองที่ได้ไปทดสอบหาความผิดพลาดของแบบจำลอง โดยการนำข้อมูลจริงที่เตรียมไว้สำหรับการทดสอบ (test) มาป้อนลงในแบบจำลองแล้วดูผลของการทำนายที่ได้

- Pattern Evaluation เป็นขั้นตอนการประเมินผลของแบบจำลองที่ได้ว่า เกิดความผิดพลาดมากน้อยเพียงใด ภาพแบบการทำนายที่ได้นั้นเป็นไปตามความต้องการหรือไม่

- Knowledge Representation เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ สะดวกต่อการใช้งาน เช่น การป้อนข้อมูลเข้าและการแสดงผลลัพธ์ เป็นต้น



ที่มา: https://www.irjet.net/archives/V8/i4/IRJET-V8I4660.pdf

รูปที่ 7 กระบวนการทำงานเหมืองข้อมูล (data mining)

2.2.8. การคัดเลือกคุณลักษณะ (feature selection)

เนื่องจากกลุ่มตัวอย่างของข้อมูลการแพทย์มีแนวโน้มที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้ข้อมลูการแพทย์บางประเภทมีจำนวนคุณลักษณะมากขึ้น ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกประเภทของโรค เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวโมเดลการจำแนกประเภทโดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของข้อมูลการแพทย์ที่สูงมากได้ดี การคัดเลือกคุณลักษณะเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในตัวแบบพยากรณ์ อาจกระทำเพื่อเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว หรือเลือกกลุ่มของตัวแปรที่มี ความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณสมบัติเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลของการทำเหมืองข้อมูล เพื่อทำให้การสร้างตัวแบบพยากรณ์มีประสิทธิภาพ เพราะจะช่วยลดมิติของข้อมูลและอาจช่วยให้การเรียนรู้วิธีการพยากรณ์ดำเนินการได้เร็วขึ้นและมีประสิทธิภาพมากขึ้น (Saabith, A. L. S., Sundararajan, E., & Bakar, A. A., 2014) การคัดเลือกคุณลักษณะที่สำคัญสามารถแบ่งออกเป็น 2 วิธี ได้แก่

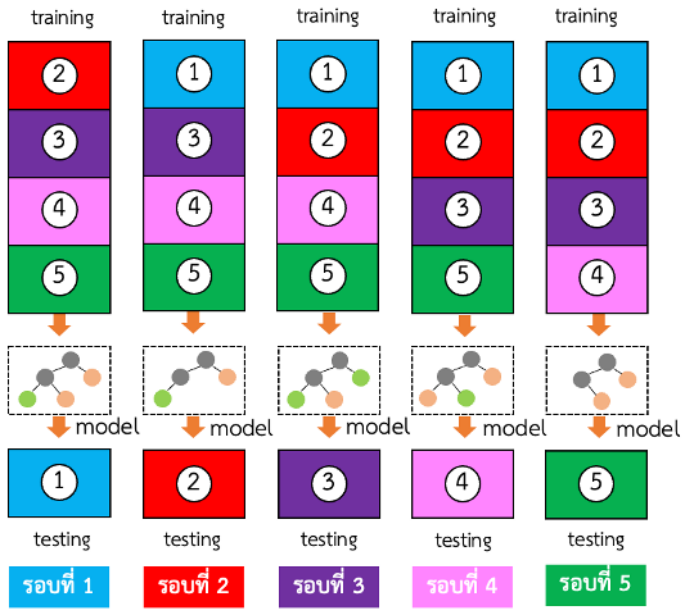
1. Filter Approach เป็นวิธีที่พยายามคัดเลือกคุณลักษณะที่สำคัญ โดยการคำนวณค่าน้ำหนัก หรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ โดยเลือกเฉพาะคุณลักษณะที่มีความสำคัญเก็บไว้ เช่น เทคนิค Correlation Based Feature Selection (CFS) เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square

2. Wrapper Approach เป็นวิธีคัดเลือกคุณลักษณะที่สำคัญ โดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มาสร้างเซตของคุณลักษณะใหม่ โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากเซตเดิม เช่น เทคนิค Forward Selection เทคนิค Backward Elimination เทคนิค Evolutionary Selection

2.2.9. k-fold cross-validation

ครอสวาลิเดชัน (cross validation) คือวิธีการในการคาดการณ์ค่าผิดพลาดของโมเดล หรือวิธีการที่นำเสนอโดยพื้นฐานของวิธีการครอสวาลิเดชัน คือกลุ่มตัวอย่าง (resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชันมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (network architecture) หรือโมเดลการคัดแยกประเภท (classification model) อาทิ ในการทำ Classify ข้อมูลโดยใช้เทคนิค Data mining จะต้องมีการแบ่งข้อมูลออกเป็นชุดสอน (training) และชุดทดสอบ (testing) แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการ Classify นั้นดีเกินจริง ดังนั้นจะมีวิธีการคิดวิธี k-fold cross-validation ขึ้นมาแก้ปัญหา

การแบ่งข้อมูลแบบ K-fold cross-validation คือการแบ่งข้อมูลออกเป็น K ชุดเท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุด จะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ ดังตัวอย่างต่อไปนี้ 5-fold cross-validation (k = 5) ชุดข้อมูลหลังจากทำการแบ่งออกเป็น 5 ชุด ข้อมูลย่อยเท่า ๆ กันโดยแต่ละกล่องคือชุดข้อมูลย่อย 1 ชุด ดังรูปที่ 8



รูปที่ 8 K-fold cross-validation (k = 5)

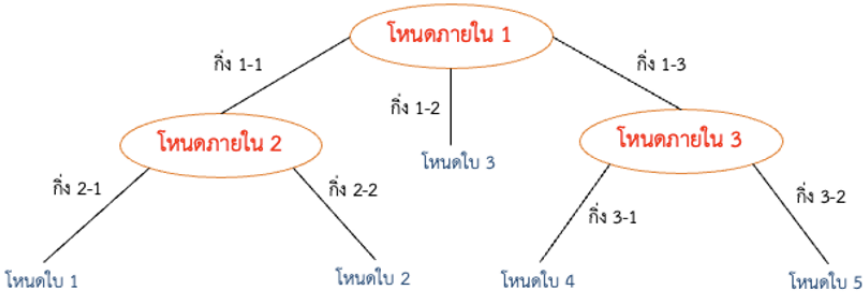
2.3 อัลกอริทึมต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจนับว่าเป็นวิธีการเรียนรู้ที่ใช้มากที่สุดแบบหนึ่งในการเรียนรู้ของเครื่อง การเรียนรู้แบบนี้เป็นการเรียนรู้โดยการจำแนก (classification) ข้อมูลออกเป็นกลุ่ม (class) ต่าง ๆ โดยใช้คุณสมบัติ (attribute) ของข้อมูลในการจำแนก ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่าคุณสมบัติใดของข้อมูลที่เป็นตัวกำหนดการจำแนก และคุณสมบัติแต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกันอย่างไร ซึ่งเป็นประโยชน์ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น ผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจจะแสดงในรูปต้นไม้ ซึ่งประกอบไปด้วย

1. โหนดภายใน (internal node) คือ คุณสมบัติต่าง ๆ ของข้อมูล ซึ่งเมื่อข้อมูลใด ๆ ตกลงมาที่โหนดจะใช้คุณสมบัตินี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้เรียกว่าโหนดราก (root node)

2. กิ่ง (branch, link) เป็นค่าคุณสมบัติของคุณสมบัติในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดภายในนั้น

3. โหนดใบ (leaf node) คือกลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์ในการแยกแยะข้อมูล ตัวอย่างของต้นไม้ตัดสินใจแสดงในรูปที่ 9



รูปที่ 9 โครงสร้างต้นไม้ตัดสินใจ

จากรูปที่ 9 แสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้

1. ต้นไม้จะเริ่มโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (training set)

2. ถ้าข้อมูลทั้งหมดอยู่ในชุดเดียวกันแล้ว ให้โหนดนั้นเป็นโหนดใบและตั้งชื่อตามกลุ่มของข้อมูลนั้น ๆ

3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ให้ทำการวัดค่าเกณ (gain) ของคุณสมบัติ (attribute) แต่ละตัวเพื่อใช้เป็นเกณฑ์ในการคัดเลือกคุณสมบัติที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดยคุณสมบัติที่มีค่าเกนมากที่สุดจะถูกเลือกเป็นตัวทดสอบหรือคุณสมบัติในการตัดสินใจและแสดงในรูปของโหนดบนต้นไม้

4. กิ่งของต้นไม้จะถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของการทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5. ทำการวนซ้ำเพื่อหาคุณสมบัติที่มีค่าเกนมากที่สุดจากข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณสมบัตินี้มาสร้างเป็นโหนดตัดสินใจตัวสินใจ คุณสมบัติที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาเป็นโหนดในลำดับต่อไปอีก

6. ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งออกไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่ง

- ถ้าข้อมูลทุกตัวในโหนดนั้นอยู่ในกลุ่มเดียวกันให้สร้างโหนดใบตามกลุ่มของข้อมูลนั้น
- ถ้าไม่เหลือคุณสมบัติใดสำหรับใช้ในการแบ่งข้อมูลแล้ว ซึ่งในกรณีนี้จะใช้กลุ่มที่มีข้อมูลสนับสนุนมากที่สุดเป็นโหนดใบ
- ถ้าไม่มีข้อมูลสนับสนุนสำหรับกิ่งนั้น ๆ แล้วให้สร้างตามกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

2.3.1. ลักษณะการเรียนรู้ของต้นไม้ตัดสินใจ

ผลการเรียนรู้แสดงอยู่ในรูปที่เข้าใจง่าย ทำให้ง่ายต่อการวิเคราะห์คุณสมบัติที่มีผลต่อการแยกแยะกลุ่มต่าง ๆ แต่ละเส้นทางจากโหนดรากถึงโหนดใบสามารถแสดงให้อยู่ในรูปกฎ IF-THEN ได้ มีความทนทานต่อข้อมูล (noisy data) เช่น คุณสมบัติที่ไม่เกี่ยวข้อง และค่าคุณสมบัติที่ผิดพลาดหรือขาดหายการเรียนรู้มีความรวดเร็วเมื่อเทียบกับอัลกอริทึมการจำแนกชนิดอื่น นำไปประยุกต์ใช้ในด้านต่าง ๆ เช่น การวิเคราะห์ความเสี่ยงของลูกหนี้, การวินิจฉัยทางการแพทย์, งานทางด้านธุรกิจ, การวิเคราะห์กลุ่มดาว และวิทยาศาสตร์ด้านอื่น ๆ

2.3.2. ค่ามาตรฐานเกน (gain criterion)

วิธีการสร้างต้นไม้ตัดสินใจแบบ ID3 จะใช้ค่ามาตรฐานเกนในการตัดสินใจเลือกคุณสมบัติที่จะใช้เป็นรากหรือโหนดในต้นไม้ โดยการคำนวณค่าเป็นของคุณสมบัติแต่ละตัว เมื่อทดลองใช้คุณสมบัตินั้นแบ่งตัวอย่างแล้วเลือกคุณสมบัติที่มีค่าเกนสูงที่สุดมาเป็นรากหรือโหนด ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ ซึ่งมีสาระสำคัญ คือ ค่าสารสนเทศของข้อมูลขึ้นอยู่กับค่าความน่าจะเป็นของข้อมูลซึ่งสามารถวัดอยู่ในรูปของบิต (bits) จากสูตร

$$\text{ค่าสารสนเทศของข้อมูล} = -\log_2(\text{ความน่าจะเป็นของข้อมูล}) \qquad (2.1)$$

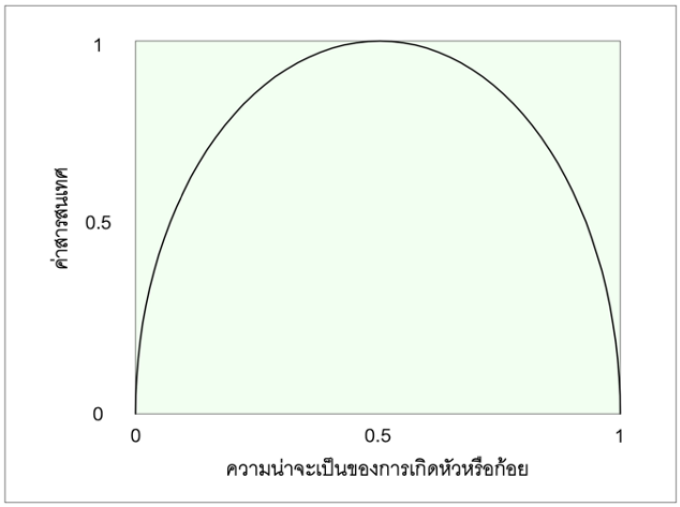
ถ้าให้ชุดของข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, \ldots, m_n\}$ และให้ความน่าจะเป็นที่จะเกิดค่า m_i มีค่าเท่ากับ $P(m_i)$ จะได้ว่าค่าสารสนเทศของ M หรือค่าเอนโทรปีของ M เขียนแทนด้วย $I(M)$ คำนวณได้จากสูตร

$$I(M) = \sum_{i}^{n} -P(M_i)\log_2 P(m_i) \qquad (2.2)$$

ตัวอย่างเช่น ในการโยนหัวโยนก้อย ชุดข้อมูล M จะประกอบด้วยค่าที่เป็นไปได้ (หัว, ก้อย) และถ้าให้ความน่าจะเป็นที่ออกหัวเท่ากับ $P(\text{หัว})$ และความน่าจะเป็นที่ออกก้อยเท่ากับ $P(\text{ก้อย})$ ดังนั้นค่าสารสนเทศของการโยนหัวโยนก้อย จะคำนวณได้จากสูตร

$$I(\text{การโยนหัวโยนก้อย}) = -P(\text{หัว})\log_2(P(\text{หัว})) - P(\text{ก้อย})\log_2(P(\text{ก้อย})) \qquad (2.3)$$

เมื่อความน่าจะเป็นของการเกิดหัวหรือก้อยมีค่าต่าง ๆ กันจะสามารถคำนวณค่าสารสนเทศของการโยนหัวโยนก้อยได้ต่าง ๆ กันดังรูปที่ 10 ซึ่งจะเห็นได้ว่าเมื่อออกหัวหมดหรือก้อยหมดค่าสารสนเทศจะเป็น 0 และค่าสารสนเทศจะค่อย ๆ เพิ่มขึ้นจนสูงที่สุดเมื่อความน่าจะเป็นของการเกิดหัวเท่ากับความน่าจะเป็นของการเกิดก้อย แสดงให้เห็นว่าค่าสารสนเทศที่น้อยจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันน้อยหรือเกือบจะเป็นพวกเดียวกัน แต่ถ้าค่าสารสนเทศสูงจะบ่งบอกว่าข้อมูลชุดนั้นมีความแตกต่างกันมากหรือประกอบด้วยตัวอย่างหลายพวกที่มีจำนวนใกล้เคียงกัน



ที่มา: https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-Classification.pdf

รูปที่ 10 ค่าสารสนเทศของการโยนหัวโยนก้อย

ในการเลือกคุณสมบัติที่จะมาเป็นโหนดรากจะอาศัยค่ามาตรฐานเกน ซึ่งคำนวณจากค่าสารสนเทศทั้งหมดของชุดข้อมูลนั้นลบด้วยค่าสารสนเทศหลังจากเลือกคุณสมบัติใดคุณสมบัติหนึ่งเป็นราก ค่าสารสนเทศหลังจากแบ่งตามคุณสมบัติที่เลือกแล้วจะคำนวณได้จากค่าผลรวมของผลคูณระหว่างค่าสารสนเทศของแต่ละโหนดกับอัตราส่วนของตัวอย่างในแต่ละกิ่งต่อตัวอย่างทั้งหมดที่โหนดนั้น ๆ หรือความน่าจะเป็นของค่าที่เป็นไปได้ของแต่ละคุณสมบัติ

ถ้าให้ข้อมูลสอนคือ T และคุณสมบัติที่เป็นโหนด คือ X และมีค่าทั้งหมดที่เป็นไปได้ n ค่า โหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น $\{t_1, t_2, \ldots, t_n\}$ ตามค่าที่เป็นไปได้ของ X ดังนั้นจึงสามารถคำนวณค่าสารสนเทศหลังจากแบ่งตามคุณสมบัติ X ดังนี้

$$I_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} I(t_i) \qquad (2.4)$$

ค่ามาตรฐานเกนของคุณสมบัติ X สามารถคำนวณได้จากการลบค่าสารสนเทศทั้งหมดที่โหนด นี้กับค่าสารสนเทศที่ได้หลังจากแบ่งด้วยคุณสมบัติ X ดังนี้

$$Gain(X) = I(T) - I_x(T) \tag{2.5}$$

2.3.3. วิธีการต้นไม้ตัดสินใจ

1. อัลกอริทึม C4.5 (C4.5 algorithm)

ในระหว่างปลายปี ค.ศ. 1970 และต้นปี ค.ศ. 1980 J. Ross Quinlan เป็นนักวิจัยทางด้านการเรียนรู้ระบบการทำงาน (machine learning) ได้พัฒนาต้นไม้ตัดสินใจ คือ ID3 (Iterative Dichotomiser) ในปี ค.ศ. 1984 มีกลุ่มของนักสถิติคือ J. Breman, J. Freidman, R. olshen and C. Stone ได้ตีพิมพ์หนังสือ Classification and Regression Trees Algorithm: CART ซึ่งอธิบายต้นไม้ตัดสินใจแบบไบนารี ต่อมาในปี ค.ศ. 1992 J. Ross Quinlan ได้เสนออัลกอริทึม C4.5 สำหรับสร้างต้นไม้ตัดสินใจ อัลกอริทึม C4.5 เป็นส่วนขยายของอัลกอริทึม ID3 ซึ่งเป็นมาตรฐานของอัลกอริทึมการเรียนรู้แบบมีผู้สอน (supervised learning) (สายชล สินสมบูรณ์ทอง, 2560)

อัลกอริทึม C4.5 ใช้แนวคิดของเกนสารสนเทศ (information gain) หรือการลดลงของเอนโทรปี (entropy reduction) เพื่อเลือกการแบ่งแยกที่เหมาะสมที่สุด สมมุติว่าเรามีตัวแปร X ตัวหนึ่ง ซึ่งมีค่าที่เป็นไปได้ k ค่า โดยมีความน่าจะเป็น $p_1, p_2, \ldots, p_k$ ตามลำดับ จำนวนบิตที่น้อยที่สุดโดยเฉลี่ยที่ใช้ส่งเรียกว่า เอนโทรปีของ X (entropy of X) นิยามได้ดังนี้

$$H(X) = -\sum_j p_j \log_2(p_j) \tag{2.6}$$

โดยที่ X แทน คุณลักษณะที่นำมาวัดค่าเอนโทรปี

p_j แทน สัดส่วนของจำนวนสมาชิกในกลุ่ม j กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

สำหรับเหตุการณ์หนึ่งที่มีความน่าจะเป็น p จำนวนสารสนเทศในบิตโดยเฉลี่ยที่ใช้ส่งคือ $-\log_2(p)$ ตัวอย่างเช่น ผลของการโยนเหรียญที่เที่ยงตรง 1 เหรียญ ด้วยความน่าจะเป็น 0.5 สามารถส่งโดยใช้ $-\log_2(0.5) = 1$ บิต ซึ่งเป็น 0 หรือ 1 ขึ้นอยู่กับผลของการโยนเหรียญ สำหรับตัวแปรที่มีผลลัพธ์หลายอย่าง จะใช้ผลรวมของน้ำหนักถ่วง โดยที่น้ำหนักเท่ากับความน่าจะเป็นของผลลัพธ์ จะได้เอ็นโทรปีก่อนการแบ่งแยกคือ

$$H(T) = -\sum_j p_j \log_2(p_j) \tag{2.7}$$

C4.5 ใช้แนวคิดของเอนโทปีดังนี้ สมมุติว่าเรามีการแบ่งแยกลูกค้า S ซึ่งแบ่งแยกชุดข้อมูลฝึกหัด T ออกเป็นเซ็ตย่อย $T_1, T_2, \ldots, T_k$ ดังนั้นความต้องการสารสนเทศเฉลี่ย (mean information requirement) หรือเอ็นโทรปีหลังการแบ่งแยกสามารถคำนวณได้ในรูปของผลรวมน้ำหนักถ่วงของเอ็นโทรปีสำหรับเซตย่อยแต่ละเซตดังนี้

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i) \tag{2.8}$$

โดยที่ P_i แทนสัดส่วนของระเบียนในเซทย่อยที่ i

ซึ่งจะนิยามผลกำไรสารสนเทศ (information gain) ได้ในรูปของ $gain(S) = H(T) - H_S(T)$ นั่นคือการเพิ่มขึ้นของสารสนเทศทำได้โดยการแบ่งข้อมูลฝึกหัด T ตามการแบ่งแยกลูกค้า S ในแต่ละโหนดการตัดสินใจ อัลกอริทึม C4.5 เลือกการแบ่งแยกที่เหมาะสมที่สุดซึ่งมีผลกำไรสารสนเทศ $gain(S)$ มากที่สุด

2. อัลกอริทึม C5.0 (C5.0 algorithm)

อัลกอริทึม C5.0 ที่นำมาใช้เป็นตัวคัดแยกนี้สามารถทำให้เกิดการสร้างต้นไม้ด้วยจำนวนกิ่งมาก ๆ ต่อหนึ่งโหนดได้ (Frey, L. et al., 2003) แสดงการสร้างต้นไม้ด้วยการใช้ Gain Ratio ซึ่งหาได้จากสมการดังต่อไปนี้

การวัดค่าความไม่บริสุทธิ์ของข้อมูล (entropy) มีสมการดังนี้

$$E(S) = -\sum_{i=1}^{n} P(S_i) \log_2 P(S_i) \tag{2.9}$$

จากสมการ (1) ค่า $i_1 - i_n$ เป็นกลุ่มข้อมูลย่อยของ Sและ $P(S_i)$คือความน่าจะเป็นของกลุ่มข้อมูล i ที่เกิดขึ้นใน Sมีการวัดค่า Information Gain เพื่อเป็นการสร้างลำดับด้วยสมการดังนี้

$$Gain(S, V) = E(S) - \sum_{v \in values(v)} \frac{|S_v|}{|S|} \times E(S_v) \tag{2.10}$$

Gain จะมีค่าอคติ (bias) ที่เป็นส่วนหนึ่งในการแบ่งแยกกลุ่มข้อมูลทำให้ในการลดค่าอคตินั้นต้องคำนวณดังสมการ

$$SplitInf\,o(S,V) = -\sum_{i=1}^{m}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|} \quad (2.11)$$

ดังนั้น เมื่อต้องการหาค่า Gain Ratio จึงเป็นการแบ่งค่า Gain ด้วยค่าการ Split Info ดังสมการ

$$GainRatio(S,V) = \frac{Gain(S,V)}{SplitInf\,o(S,V)} \quad (2.12)$$

อัลกอริทึม C5.0 พัฒนามาจากอัลกอริทึม C4.5 จึงได้เพิ่มความถูกต้องของตัวคัดแยก C5.0 ด้วยการใช้วิธี Boosting เป็นเทคนิคการสร้างการเชื่อมต่อของตัวคัดแยกหลาย ๆ ตัวให้มีความถูกต้องมากขึ้น อัลกอริทึม C5.0 ได้รวมเอาฟังก์ชันต่าง ๆ เช่น ค่าความแปรปรวนของการคัดแยกผิด (variable misclassification costs) ซึ่งจะมีการแยกค่าสูญหายในการคัดแยกของกลุ่มข้อมูล ทำให้ลดการเกิดความแปรปรวนของการคัดแยกผิดได้ในอัลกอริทึม C5.0 ยังให้ความสำคัญในการกระจายน้ำหนักบนกลุ่มข้อมูล เช่น ในงานวิจัยของ Berrar, D. et al. (2001) และ Ménard, C. et al. (2006) กำหนดค่าน้ำหนักอัลกอริทึม C5.0 ในการประมวลผลเป็นค่าเดียวกัน แต่ถ้าเกิดกรณีที่มีการคัดแยกผิดก็จะมีการเพิ่มค่าน้ำหนักให้สูงขึ้นและทำการคัดแยกข้อมูลใหม่อีกครั้ง นอกจากนี้อัลกอริทึม C5.0 ได้เพิ่มการรองรับชนิดข้อมูลใหม่ ๆ ด้วย ซึ่งทำให้เกิดข้อดี คือ อัลกอริทึม C5.0 ใช้งานง่ายขึ้น ลำดับขั้นตอนการทำงานของอัลกอริทึม C5.0 แสดงได้ดังนี้ (Govindarajan, M., 2007)

- สำหรับแต่ละคุณสมบัติหาค่า Information Gain ของแต่ละคุณสมบัติเพื่อเป็นการจัดลำดับของความสำคัญของคุณสมบัติ
- สร้างโหนดรากจากค่า Information Gain ที่มากที่สุด
- ทำการสร้างกิ่งจากโหนดรากให้เป็นโหนดลูก (children node)
- เมื่อไม่สามารถสร้างกิ่งต่อไปได้อีกจะได้โหนดสุดท้ายเป็นผลลัพธ์ หรือ โหนดใบ (leaf node)

2.4 วิธี Random forest

วิธี Random forest เป็นการสุ่มเลือกข้อมูลเพื่อนำมาทำต้นไม้ตัดสินใจ (decision tree) หลาย ๆ แบบจำลองเพื่อทำนายผล ซึ่งจำนวนผลลัพธ์ที่ซ้ำกันมากที่สุดของแต่ละแบบจำลองจะเป็นคำตอบของการทำนายผล โดยกระบวนการนี้ช่วยลดความผิดพลาดและเสถียรภาพของการทำนาย ซึ่งทำให้แต่ละต้นไม้มีความหลากหลายในการตัดสินใจและไม่เกิดการเรียนรู้มากเกินไป (overfitting) อัลกอริทึม Random forest ได้รับความนิยมมากในวงการเรียนรู้ของเครื่อง (machine learning)

และงานวิจัยที่เกี่ยวข้องเนื่องจากความสามารถในการทำนายแม่นยำและการทำงานที่นิยมในงานที่ต้องการประมวลผลข้อมูลที่มีความซับซ้อน เช่น การวิเคราะห์ภาพถ่ายทางการแพทย์ การคัดแยกข้อมูลทางการแพทย์ และการจัดแบ่งกลุ่มลูกค้า เป็นต้น ลักษณะของต้นไม้ที่อยู่ภายในป่าของเทคนิคการสุ่มป่าไม้จะถูกควบคุมด้วย 3 ปัจจัยคือ

1. ต้นไม้แต่ละต้นจะถูกสอน (train) โดยการใช้เซตย่อยจากข้อมูลตัวอย่าง

2. เมื่อต้นไม้โตขึ้นจะสามารถค้นหาโนด (node) แต่ละโนดที่อยู่ในกิ่งที่ดีที่สุดของต้นไม้โดยใช้การสุ่ม เลือกคุณสมบัติจาก N คุณสมบัติ

3. ต้นไม้แต่ละต้นจะไม่มีการตัดออก แต่จะปล่อยให้ต้นไม้โตขึ้นไปเรื่อย ๆ จนได้ผลลัพธ์ที่ดีที่สุดหลังจากการสร้างป่า แล้วทำการให้คะแนน (vote) โดยต้นไม้ภายในป่า หากต้นไม้ต้นใดได้คะแนนสูงสุด ก็จะนำเอาต้นไม้นั้นออกมาสร้างเป็นโมเดล

ข้อดีของ Random forest

1. ใช้ได้ทั้งกับปัญหา classification และ regression

2. ใช้ได้ทั้งกับข้อมูลมีโครงสร้าง (structured) และไม่มีโครงสร้าง ข้อมูลมีโครงสร้าง เช่น ข้อมูลจัดเก็บเป็นตารางหรือคอลัมน์แล้ว ข้อมูลไม่มีโครงสร้าง เช่น รูปภาพ หรือข้อความ เป็นต้น

3. ป้องกันการเกิด overfitting กับชุดข้อมูลที่นำมาสร้าง

2.5 เกณฑ์ในการวัดประสิทธิภาพความแม่นยำ

ในการวัดประสิทธิภาพความแม่นในการศึกษาครั้งนี้ใช้เกณฑ์การวัดด้วยค่าความถูกต้อง (accuracy) และเกณฑ์ในการทำนาย Area Under Curve (AUC) โดยจะใช้ค่าเฉลี่ยของค่าความถูกต้อง และค่า AUC จากอัลกอริทึมต้นไม้ตัดสินใจ มีรายละเอียดดังนี้

2.5.1. เกณฑ์การวัดด้วยค่าความแม่นยำ

ในการทดสอบประสิทธิภาพความแม่นยำโดยใช้เกณฑ์ในการวัดด้วยค่าความแม่นยำ โดยคำนวณจากค่าในแนวเส้นทแยงมุมของเมทริกซ์ความสับสน (confusion matrix) ได้ดังตารางที่ 1

ตารางที่ 1 เมทริกซ์ความสับสนแบบ 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)

สามารถคำนวณได้โดยสูตร ดังนี้

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.13)$$

โดยที่ ค่าความถูกต้อง (accuracy) คือ ค่าอยู่ระหว่าง 0 - 1 เมื่อค่าเข้าใกล้ 1 นั่นคือตัวแบบสามารถจำแนก ประเภทได้ดีมาก

TP คือ สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่าจริง และสิ่งที่เกิดขึ้นคือ จริง

FN คือ สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้นคือ จริง

FP คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้นคือ ไม่จริง

TN คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้นคือ ไม่จริง

2.5.2. 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) โดยค่า AUC มีค่าอยู่ระหว่าง 0 - 1 เมื่อค่าเข้าใกล้ 1 นั่นหมายความว่าตัวแบบในภาพรวมสามารถจำแนกประเภทได้ดีมาก

โดยที่

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (2.14)$$

ค่าความจำเพาะ (specificity) คือ ค่าความถูกต้องในการพยากรณ์ของคลาสที่ไม่เกิดโรคต่อจำนวนทั้งหมดที่ไม่เกิดโรคจริง

$$Specificity = \frac{TN}{TN+FP} \quad (2.15)$$

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

$$FPR = \frac{FP}{FP + TN} \tag{2.16}$$

สามารถสรุปได้ดังเกณฑ์ต่อไปนี้

$0.50 \leq AUC < 0.70$ คือ ตัวแบบมีประสิทธิภาพต่ำ

$0.70 \leq AUC < 0.80$ คือ เกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่

$0.80 \leq AUC < 0.90$ คือ ตัวแบบทำงานได้ดี

$AUC \geq 0.90$ คือ ตัวแบบทำงานได้ดีมาก

2.6 งานวิจัยที่เกี่ยวข้อง

2.6.1 งานวิจัยในประเทศ

อุมาพร นันทธีโร (2555) ศึกษาเกี่ยวกับอาการนำมาพบแพทย์ของผู้มาตรวจมะเร็งเต้านม โดยพบว่าอาการนำมาพบแพทย์ เช่น คลำเจอก้อน, เจ็บเต้านม มีความสัมพันธ์กับขนาดของก้อนเนื้อ (Mass) เพราะเนื่องจากการเกิดของโรคมะเร็งเต้านมของผู้ป่วยมักไม่มีอาการผิดปกติในระยะแรก จึงมีความจำเป็นและสำคัญอย่างยิ่งที่ต้องทำการตรวจค้นหามะเร็งเต้านมในระยะเริ่มต้นของผู้มาตรวจมะเร็งเต้านมทั้งในกลุ่มที่มีอาการและไม่มีอาการ โดยพบผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (BIRADS 4-6) และมีอาการนำมาพบแพทย์ คิดเป็นร้อยละ 96.3

ณัฐวุฒิ ศรีวิบูลย์ (2559) ได้นำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์โรคมะเร็ง โดยทำการเปรียบเทียบอัลกอริทึมการทำเหมืองข้อมูล ได้แก่ C4.5, k-Nearest Neighbor และ Naïve Bayes ผลการเปรียบเทียบพบว่า ต้นไม้ตัดสินใจ C4.5 มีประสิทธิภาพสูงสุดที่ 98.63%

ภรณี เหล่าอิทธิ และนภา ปริญญานิติกูล (2559) ศึกษาเกี่ยวกับมะเร็งเต้านม พบว่าผู้หญิงที่อ้วนหรือมีภาวะน้ำหนักเกินมาตรฐาน (ดัชนีมวลกายเกินมาตราฐาน) โดยเฉพาะในผู้หญิงวัยหลังหมดประจำเดือนแล้ว การสูบบุหรี่ และดื่มเครื่องดื่มที่มีแอลกอฮอล์ก็เพิ่มโอกาสความเสี่ยงของการเกิดโรคมะเร็งเต้านม รวมไปถึงอายุที่มากขึ้นทำให้มีความเสี่ยงในการเป็นมะเร็งเต้านมได้เช่นเดียวกัน

ภรัณยา ปาลวิสุทธิ์ (2559) งานวิจัยมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับการทำนายความผิดปกติของการติดยาเสพติดทางอินเทอร์เน็ต (IAD) ในวัยรุ่น ซึ่งการวิเคราะห์ข้อมูลพบปัญหาความไม่สมดุลระดับซึ่งได้รับการแก้ไขโดยใช้ SMOTE เพื่อเพิ่มข้อมูลคลาสที่น้อยกว่า โดยวิธีต้นไม้ตัดสินใจ J48, ID3, LMT, CART และ Random forest ถูกนำมาใช้ในการสร้างแบบจำลองการทำนายและประสิทธิภาพที่ถูกวัดโดยใช้ความถูกต้อง , ความไว และความจำ ผลการทดลองพบว่า Random Forest มีประสิทธิภาพสูงกว่า J48, ID3, LMT, และ CART โดยมีค่าความถูกต้อง 87.15% ค่าความไว 85.89% และค่าความจำเพาะ 87.53%

วิชญ์วิสิฐ เกสรสิทธิ์ และวิชิต หล่อจีระชุณห์กุล (2561) ได้ใช้วิธีการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน โดยได้ใช้วิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวานจำนวน 4 วิธี คือ วิธีสุ่มเกิน (oversampling), วิธีสุ่มลด (undersampling), วิธีผสมผสาน (hybrid method) และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยอัลกอริทึมต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด

อัจจิมา มณฑาพันธุ์ (2562) เปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่าง ๆ จำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection ผลการทดลองพบว่าเทคนิค Evolutionary Selection ได้ผลดีที่สุดจากจำนวน 7 เทคนิค

มุทิตา หวังคิด, อาริกา ธรรมโน และอาริต ธรรมโน (2563) ศึกษาอัลกอริทึมเหมืองข้อมูลโดยนำเสนออัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง รวมทั้งทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยภาษาไพทอน โดยมีจุดประสงค์เพื่อช่วยในการคัดกรองผู้ป่วยโรคมะเร็งเต้านมในเบื้องต้น เพื่อให้กระบวนการการวินิจฉัยเป็นไปได้อย่างรวดเร็วและสามารถรักษาได้อย่างทันเวลา

อุกฤษฏ์ ศรีสุข และจารี ทองคำ (2564) ทำการการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค ซึ่งรวบรวมมาจากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล โดยนำเอาเทคนิค Machine Learning มาใช้กับการทำเหมืองข้อมูล 5 เทคนิค ได้แก่ Decision Tree C4.5, Naïve Bayes, Neural Networks, Random forest, Deep Learning มาทำการสร้างแบบจำลองเพื่อการพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ จากการทดลองพบว่า เทคนิค Decision Tree C4.5 เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์และโรคมะเร็งเต้านม โดยให้ค่าความถูกต้อง 99.86%

และ 75.52% ตามลำดับ และเทคนิค Deep Learning เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคเบาหวาน โดยให้ค่าความถูกต้อง 77.47%

ศิวกร บันลือทรัพย์ และวราพร จิระพันธุ์ทอง (2565) ได้ทำการศึกษาอัลกอริทึมการเรียนรู้ของเครื่องสำหรับการทำนายการคัดแยกผู้ป่วย COVID-19 ได้ศึกษาข้อมูลผู้ป่วย จำนวน 1,608,923 ราย จากกรมควบคุมโรค โดยใช้อัลกอริทึมการจำแนกข้อมูลทั้งหมด 3 แบบได้แก่ Random forest, Neural network และ Naive Bayes ผลการวิจัยพบว่า Random Forest มีค่าความถูกต้องเท่ากับ 93.33% ซึ่งเป็นอัลกอริทึมที่ดีที่สุดจากทั้ง 3 แบบ ในการคัดแยกผู้ป่วยออกเป็น 4 ประเภท คือ 1. การเฝ้าระวังในกลุ่มผู้ป่วยหรือผู้มีอาการเข้าได้กับนิยามผู้สงสัยติดเชื้อไวรัสโคโรนา 2019 ที่เข้าเกณฑ์สอบสวนโรค (patient under investigation : PUI) 2. การตรวจคัดกรองในประชากรเสี่ยงตามจุดคัดกรองและด่านเข้าออกระหว่างประเทศ(screening) 3. การเฝ้าระวังในกลุ่มเป้าหมายเฉพาะหรือพื้นที่เฉพาะ (sentinel surveillance) และ 4. การเฝ้าระวังเหตุการณ์ในสถานที่เสี่ยง เก็บตัวอย่างส่งตรวจเมื่อเข้านิยาม PUI หรือเป็นกลุ่มก้อนของผู้ป่วยทางเดินหายใจ

2.6.2 งานวิจัยต่างประเทศ

Venkatesan, E. V., & Velmurugan, T. (2015) ได้ทำการวิเคราะห์ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจสำหรับการจำแนกโรคมะเร็งเต้านม โดยรวบรวมข้อมูลจากเมืองเจนไน ประเทศอินเดีย ซึ่งในงานนี้ได้ทำการเปรียบเทียบอัลกอริทึมการจำแนกประเภท 4 วิธี ได้แก่ j48, Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree) และ Best First Tree (BF Tree) ผลการทดลองแสดงให้เห็นว่า อัลกอริทึม J48 มีความแม่นยำสูงสุด 99%, อัลกอริทึม CART มีความแม่นยำ 96%, อัลกอริทึม AD Tree มีความแม่นยำ 97% และอัลกอริทึม BF Tree มีความแม่นยำ 98% จากผลการนายของอัลกอริทึมทั้ง 4 ประสิทธิภาพของ J48 ซึ่งสอดคล้องกับงานวิจัยของ Sumbaly, R., Vishnusri, N. & Jeyalatha, S. (2014)

Ray, R., et al. (2019) ทำการทดลองโดยใช้ Decision Tree, k- nearest neighbors (k-NN), Random Forest และ Gaussian Naïve Bayes ในการพยากรณ์โรคมะเร็งเต้านม จากผลการทดลอง พบว่า Decision Tree และ Random Forest มีความสามารถในการพยากรณ์สูงกว่าอีก 2 อัลกอริทึมที่นำมาศึกษา

Nwokocha, N., Kabari, L., & Agaba, F. (2019) ใช้เทคนิคการทำเหมืองข้อมูลต้นไม้ตัดสินใจ (Decision Tree) เพื่อทำนายการกลับมาเป็นซ้ำของโรคมะเร็งเต้านมโดยการใช้แอทริบิวต์ 10 แอทริบิวต์จาก UCI Repository ทำนายด้วยโปรแกรม WEKA ผลการทดลองพบว่า มีค่าความถูกต้อง 65.97%

Mosayebi, A. et al. (2020) สร้างแบบจำลองและเปรียบเทียบอัลกอริทึมการทำเหมืองข้อมูลเพื่อทำนายการกลับเป็นซ้ำของมะเร็งเต้านม โดยใช้อัลกอริทึม 7 อัลกอริทึม ได้แก่ LVQ, Bayesian, MLP, KPCA, SVM, C5.0 และ Random Forest ผลการวิเคราะห์แสดงให้เห็นว่าต้นไม้ตัดสินใจ: C5.0 ให้ค่าความถูกต้องมากที่สุดในการทำนายการกลับมาเป็นซ้ำของโรคมะเร็งเต้านม

Saad, H. & Nagarur, N. (2020) ศึกษาเกี่ยวกับการจำแนกผู้ป่วยโรคมะเร็งเต้านมในสตรีชาวลิเบียจำนวน 130 คน โดยใช้เทคนิคเหมืองข้อมูล 6 อัลกอริทึม ได้แก่ Logistic Regression, Neural Network, k-NN, Naïve bayes, Decision Tree และ SVM พร้อมทั้งการคัดเลือกคุณลักษณะ (feature selection) ที่ทำให้โมเดลมีความแม่นยำมากขึ้น พบว่าโมเดลที่ทำนายการเกิดโรคมะเร็งเต้านมได้ดีที่สุดในบรรดาทั้ง 6 อัลกอริทึมที่ศึกษา คือ Decision Tree โดยให้ค่าความแม่นยำสูงถึง 90%

บทที่ 3 วิธีดำเนินการวิจัย

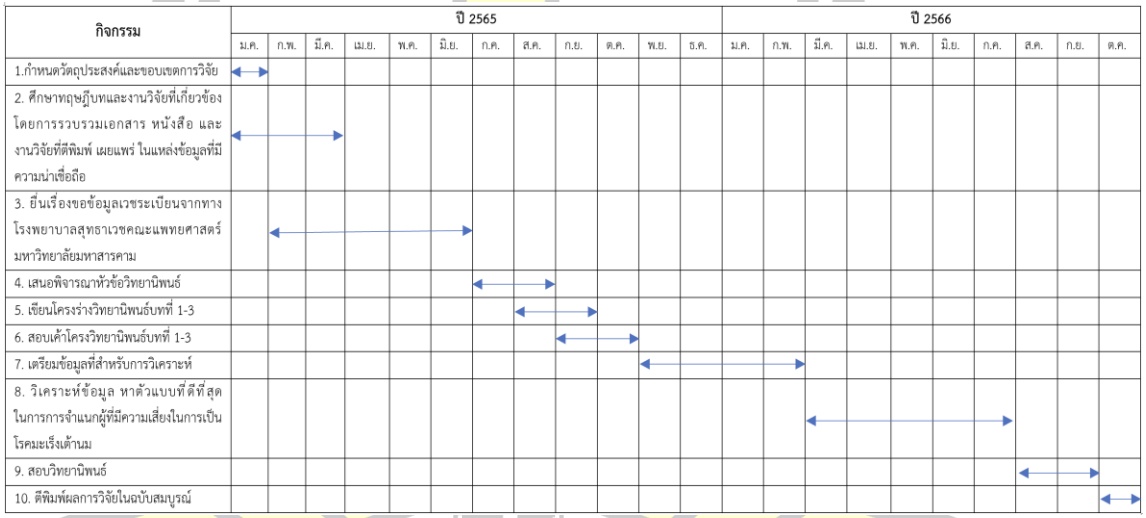
การวิจัยเรื่องการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกประเภทโรคมะเร็งเต้านม (breast cancer) และศึกษาปัจจัยเสี่ยงที่ทำให้เกิดโรคมะเร็งเต้านม (breast cancer) ด้วยวิธีการที่นำเสนอ โดยกำหนดวิธีการที่ใช้ในการศึกษาดังนี้

3.1 แผนงานปฏิบัติการวิจัย

3.2 วิธีการเก็บรวบรวมข้อมูลและการเตรียมข้อมูล

3.1 แผนงานปฏิบัติการวิจัย

ตารางที่ 2 แผนงานปฏิบัติการวิจัย (gantt chart)



3.2 วิธีการเก็บรวบรวมข้อมูลและการเตรียมข้อมูล

3.2.1 การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลทำโดยการสังเคราะห์ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 โดยมี

ตารางที่ 3 ตัวแปรอิสระ (independent variable)

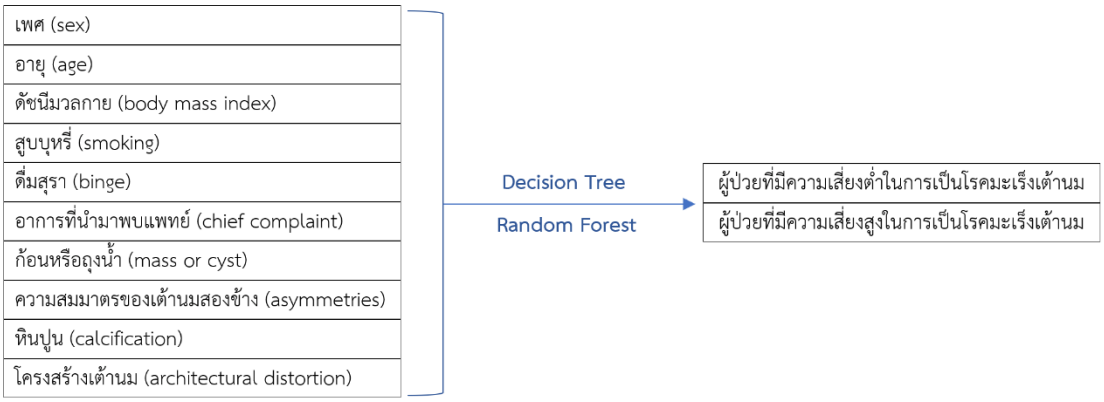
ตัวแปร	กลุ่มตัวแปร (สัดส่วนข้อมูล)
1. เพศ (sex)	ชาย (0.79%), หญิง (99.21%)
2. อายุ (age)	อายุ 20-32 ปี (50.20%), อายุ 33-45 ปี (25.39%), อายุ 46-58 ปี (21.00%), อายุ 59-71 ปี (1.97%), อายุ 72-84 ปี (1.25%), อายุ 85-97 ปี (0.20%)
3. ดัชนีมวลกาย (body mass index: BMI)	<18.5 (3.54%), 18.5-22.9 (39.44%), 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)	มี (48.82%), ไม่มี (51.18%)
8. ความสมมาตรของเต้านมสองข้าง (asymmetries)	สมมาตร (72.57%), ไม่สมมาตร (27.43%)
9. หินปูน (calcification)	มี (50.72%), ไม่มี (49.28%)
10. โครงสร้างเต้านม (architectural distortion)	ผิดปกติ (9.78%), ไม่ผิดปกติ (90.22%)

ตัวแปรตาม (dependent variable) คือ

ผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม

ผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (88.12%) หมายถึง ผู้ที่เข้ารับการตรวจแมมโมแกรม (mammogram) โดยมีค่า BIRADs Score คิดจากลักษณะรูปภาพที่เห็นซึ่งอยู่ในระดับคะแนน 0-3

ผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (11.88%) หมายถึง ผู้ที่เข้ารับการตรวจแมมโมแกรม (mammogram) โดยมีค่า BIRADs Score คิดจากลักษณะรูปภาพที่เห็นซึ่งอยู่ในระดับคะแนน 4-6



รูปที่ 11 กรอบแนวคิดของการวิจัย

3.2.2 การเตรียมข้อมูล

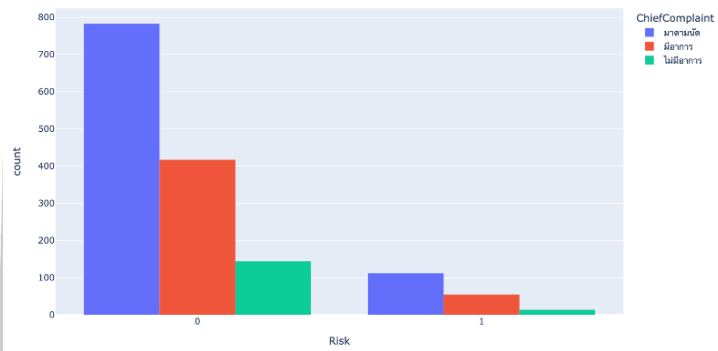
การเตรียมข้อมูล (data preparation) เป็นขั้นตอนสำคัญในกระบวนการทำเหมืองข้อมูล (data mining) ซึ่งเป็นกระบวนการค้นคว้าและวิเคราะห์ข้อมูลเพื่อค้นหาความรู้และแบ่งแยกข้อมูลที่ซ่อนอยู่ภายในข้อมูลขนาดใหญ่ การเตรียมข้อมูลมุ่งเน้นการเตรียมข้อมูลให้พร้อมใช้ในขั้นตอนการวิเคราะห์ข้อมูลและการประมวลผลข้อมูลต่าง ๆ ซึ่งข้อมูลที่นำมาทำการศึกษาเป็นข้อมูลจากโรงพยาบาลสุทธาเวช คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม โดยเป็นข้อมูลดิบ (raw data) จำนวน 1,845 ระเบียน ตัวอย่างดังรูปที่ 12 และผู้วิจัยได้ทำการเตรียมข้อมูลดังนี้

Call number	sex	age	bmi	pregnancy	smoking_type_id	drinking_type_id	symptom	xreport
650001984	2	43	29.136	N	1	1	แพทย์นัด F/U MMG	Findings ; Rt.breast Study reveals heterogeneous fibroglandular tissue without mass or calcification. No structural distortion noted. Surrounding fat and fibrous planes are unremarkable. Neither nipple, skin thickening nor retraction noted. Sonogram of Rt.breast shows unremarkable breast tissue without mass or calcification. No structural distortion noted. Surrounding fat and fibrous planes are unremarkable. Benign appearance Rt.axillary lymph nodes noted. Lt.breast Study reveals heterogeneous fibroglandular tissue without mass or calcification. No structural distortion noted. Surrounding fat and fibrous planes are unremarkable. Neither nipple, skin thickening nor retraction noted. Sonogram of Lt.breast shows unremarkable breast tissue without mass or cyst. Surrounding fat and fibrous planes are unremarkable. Benign appearance Lt.axillary lymph nodes noted. ====== [Conclusion] ====== Rt.breast : Unremarkable study. Lt.breast : Unremarkable study. BIRADS : 2 Annual check up.
650002030	2	51	27.239	N	1	1	มาทำMammography with US ตามนัด	CLINICAL INDICATION: A 50-year-old postmenopausal woman, is requested for screening. TECHNIQUE: Standard CC and MLO views. High frequency linear probe ultrasound. PREVIOUS STUDY: 1/06/2018. FINDINGS: MAMMOGRAMS: - Breast composition: The breasts are heterogeneously dense, which may obscure small masses. - Mass: Not detectable. - Calcification: Not detectable. - Architectural distortion: Not detectable. - Asymmetries: Not detectable. - Intramammary node: Not detectable. - Skin and Nipples: Normal. - Other features: Not detectable. US BREASTS: - Tissue composition: Heterogeneous background echo-fibroglandular. - Mass or cyst: Not detectable. - Ductal ectasia: Not detectable. - Axillary node: No bilateral axillary lymph node enlargement is detected. - Others feature: Not detectable. ====== [Conclusion] ======

รูปที่ 12 ตัวอย่างข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม

1. การเก็บข้อมูล (data collection) ขั้นตอนแรกคือการรวบรวมข้อมูลซึ่งผู้วิจัยได้ทำการขอข้อมูลจากแผนกเวชระเบียน โรงพยาบาลสุทธาเวช คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม

2. การสำรวจข้อมูล (explore data) เป็นขั้นตอนที่สำคัญในการเตรียมข้อมูลและเข้าใจลักษณะของข้อมูลที่จะนำมาใช้ในการทดสอบและประเมินอัลกอริทึมต่าง ๆ ของต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกโรคมะเร็งเต้านม โดยผู้วิจัยได้ทำการศึกษาข้อมูลเพื่อเข้าใจโครงสร้างของข้อมูล เช่น ดูความสัมพันธ์ของอาการที่นำมาพบแพทย์กับความเสี่ยงในการจำแนกโรคมะเร็งเต้านม ดังรูปที่ 13 อีกทั้งผู้วิจัยทำการสำรวจข้อมูลอายุ (age) ของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม พบว่าอายุต่ำสุดคือ 20 ปี อายุสูงสุดคือ 88 ปี และอายุเฉลี่ยอยู่ที่ 52 ปี ข้อมูลดัชนีมวลกาย (BMI) ของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม พบว่าดัชนีมวลกายต่ำสุดคือ 13.468 ดัชนีมวลกายสูงสุดคือ 37.333 และดัชนีมวลกายเฉลี่ยอยู่ที่ 24.136 เป็นต้น



รูปที่ 13 ความสัมพันธ์ของอาการที่นำมาพบแพทย์กับความเสี่ยงในการจำแนกโรคมะเร็งเต้านม

3. การทำความสะอาดข้อมูล (data cleaning) การตรวจสอบและแก้ไขข้อมูลที่หายไป ข้อมูลที่ซ้ำกัน, ข้อมูลที่ไม่สมบูรณ์ และข้อมูลที่ผิดพลาด โดยข้อมูลที่นำมาทำการศึกษา พบว่ามีข้อมูลสูญหาย (missing data) หรือข้อมูลที่ไม่สมบูรณ์ (incomplete data) เพื่อให้ข้อมูลมีคุณภาพและถูกต้อง ผู้วิจัยได้ทำการตรวจสอบข้อมูลโดยมีจำนวนข้อมูลสูญหายและข้อมูลที่ไม่สมบูรณ์ 1.6% จึงทำการจัดการกับข้อมูลสูญหายโดยการลบข้อมูลที่มีส่วนของข้อมูลสูญหายออก (listwise deletion) ทำให้ข้อมูลเหลืออยู่ทั้งหมด 1,524 ระเบียน ซึ่งการทำความสะอาดข้อมูลเป็นขั้นตอนสำคัญในกระบวนการวิเคราะห์ข้อมูลและการเตรียมข้อมูลเพื่อใช้ในการทำโมเดลหรือการประมวลผลข้อมูลอื่น ๆ เพื่อให้ผลลัพธ์ที่ได้มีคุณภาพและน่าเชื่อถือสูงขึ้น

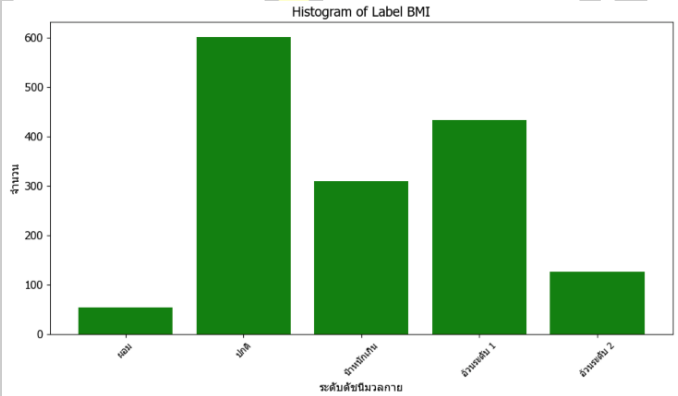
4. การแปลงข้อมูล (data transformation) การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการทำเหมืองข้อมูล ในส่วนนี้ผู้วิจัยได้ทำการแปลงข้อมูลอายุเป็นกลุ่มอายุ โดยความถี่ของข้อมูลช่วงอายุดังรูปที่ 14 และข้อมูลดัชนีมวลกายเป็นระดับการวัดผลโดยใช้ดัชนีมวลกาย โดยความถี่

ของข้อมูลระดับการวัดผลดังรูปที่ 15 ซึ่งการแบ่งกลุ่มข้อมูลอาจช่วยในการทำให้เห็นความสัมพันธ์หรือรูปแบบที่ซ่อนอยู่ในข้อมูลได้ชัดเจนขึ้น ช่วยลดความซับซ้อนของข้อมูล และทำให้การวิเคราะห์ข้อมูลง่ายขึ้น อีกทั้งสามารถช่วยในการสรุปแนวโน้มหรือรูปแบบของข้อมูลได้ดีขึ้น โดยเปลี่ยนข้อมูลเชิงปริมาณให้อยู่ในรูปเชิงคุณภาพ สามารถดำเนินการแปลงข้อมูลดัชนีมวลกายให้เป็นระดับการวัดผลโดยใช้ดัชนีมวลกาย (body mass index-BMI) ดังนี้

BMI น้อยกว่า 18.5 ถือว่าน้ำหนักต่ำกว่าเกณฑ์ BMI ระหว่าง 18.5-22.9 ถือว่าปกติ

BMI ระหว่าง 23.0-24.9 ถือว่าน้ำหนักเกิน BMI ระหว่าง 25.0-29.9 ถือว่าอ้วนระดับ 1

BMI มากกว่า 30 ถือว่าอ้วนระดับ 2

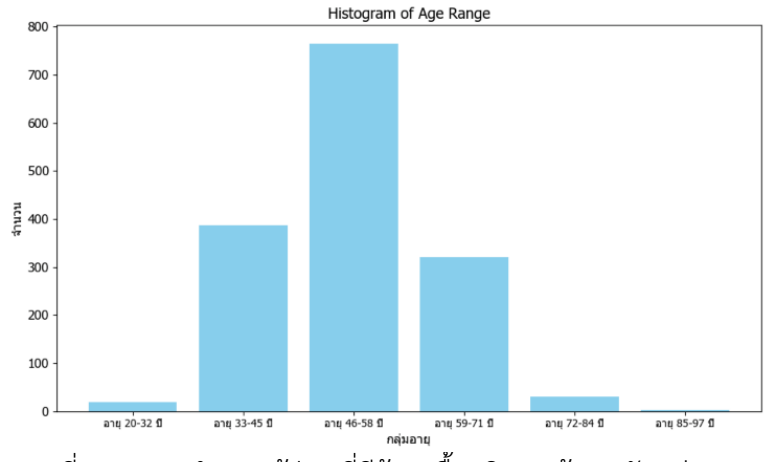


รูปที่ 14 กราฟจำนวนผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านมกับระดับดัชนีมวลกาย

นอกจากนี้ยังแปลงข้อมูลอายุให้อยู่ในกลุ่มอายุ ที่แบ่งเป็น 6 กลุ่มอายุ ได้แก่

ช่วงอายุ 20-32 ปี ช่วงอายุ 33-45 ช่วงอายุ 46-58 ปี

ช่วงอายุ 59-71 ปี ช่วงอายุ 72-84 ปี ช่วงอายุ 85-97 ปี



รูปที่ 15 กราฟจำนวนผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านมกับกลุ่มอายุ

5. การเตรียมคุณสมบัติ (feature engineering) การสร้างคุณสมบัติใหม่หรือการแปลงคุณสมบัติเพื่อเพิ่มประสิทธิภาพในการทำเหมืองข้อมูล ให้ข้อมูลสามารถเข้ากับรูปแบบหรือระบบที่ใช้งานได้ ในส่วนนี้ผู้วิจัยได้ทำการเข้ารหัสข้อมูล (one-hot-encoding) เป็นวิธีการแปลงข้อมูลที่เป็นแบบประเภท (categorical data) เป็นข้อมูลแบบตัวเลขที่เหมาะสมในการป้อนเข้าสู่ระบบประมวลผลหรือระบบเรียนรู้ของเครื่อง การเข้ารหัสนี้จะสร้างคอลัมน์ใหม่สำหรับแต่ละหมวดหมู่ของข้อมูล และใส่ค่า 1 ในคอลัมน์ที่ตรงกับหมวดหมู่ของข้อมูลนั้น และค่า 0 ในคอลัมน์อื่น ๆ ดังรูปที่ 16

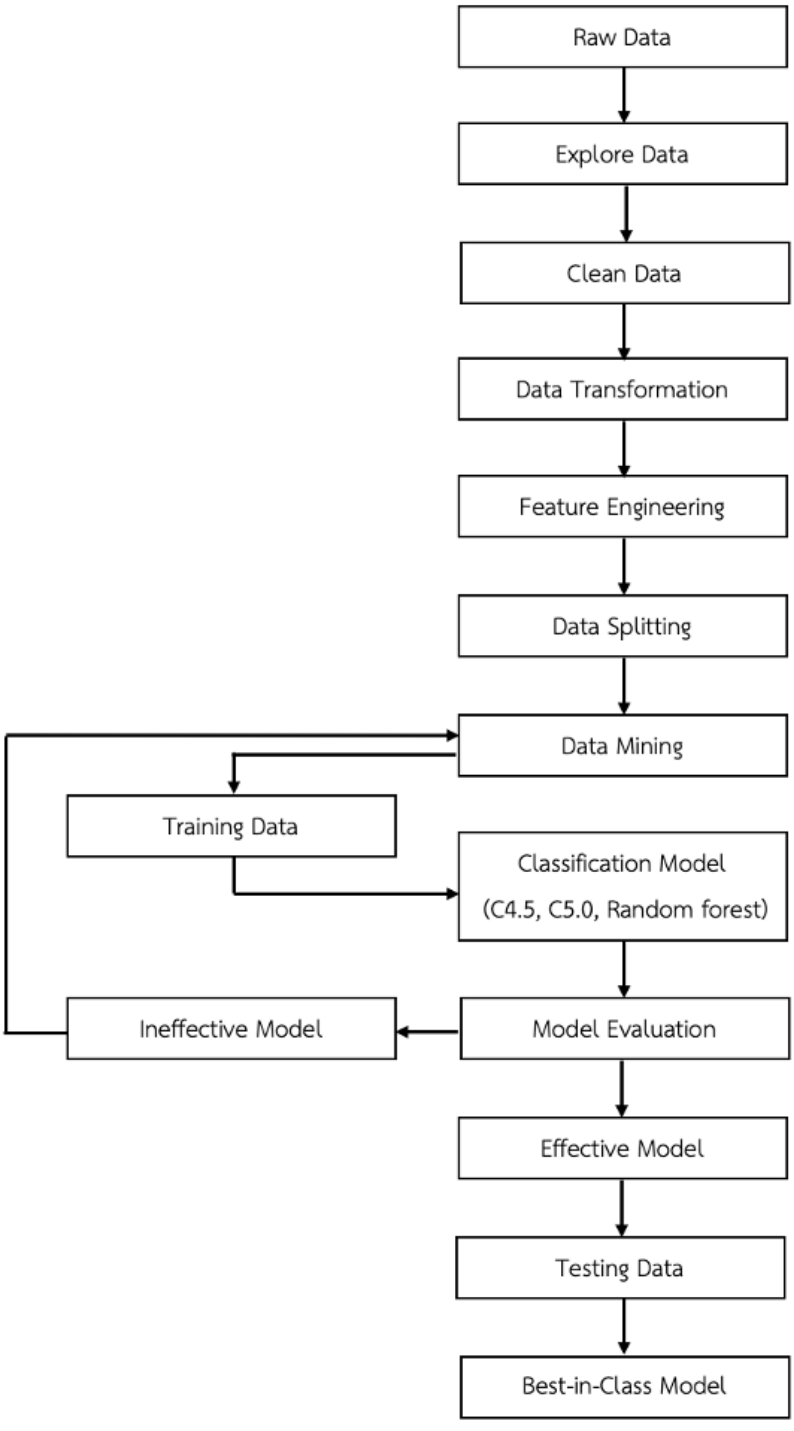
Gender	Age	Age_range	BMI	Label_BMI	Smoking	Binge	ChiefComplaint	MassORCyst	Asymmetries	Calcification	Architectural	Risk
หญิง	70	อายุ 59-71 ปี	13.896	ผอม	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	0	1	0	1	0
หญิง	27	อายุ 20-32 ปี	23.674	น้ำหนักเกิน	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มีอาการ	1	1	0	0	0
หญิง	63	อายุ 59-71 ปี	15.571	ผอม	สูบบุหรี่	ไม่ดื่มสุรา	มีอาการ	1	1	1	0	0
หญิง	29	อายุ 20-32 ปี	23.733	น้ำหนักเกิน	ไม่สูบบุหรี่	ไม่ดื่มสุรา	ไม่มีอาการ	1	1	0	0	0
หญิง	47	อายุ 46-58 ปี	16.016	ผอม	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	1	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
หญิง	48	อายุ 46-58 ปี	36.885	อ้วนระดับ 2	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	1	1	1	0	0
หญิง	88	อายุ 85-97 ปี	13.468	ผอม	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	1	0	0	0	0
หญิง	55	อายุ 46-58 ปี	37.036	อ้วนระดับ 2	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	0	1	1	0	0
หญิง	54	อายุ 46-58 ปี	37.223	อ้วนระดับ 2	ไม่สูบบุหรี่	ไม่ดื่มสุรา	ไม่มีอาการ	0	1	1	0	0
หญิง	88	อายุ 85-97 ปี	20.613	ปกติ	ไม่สูบบุหรี่	ไม่ดื่มสุรา	มาตามนัด	1	1	1	1	1

รูปที่ 16 ตัวอย่างข้อมูลการเข้ารหัสข้อมูล (one-hot-encoding)

6. การแยกชุดข้อมูล (data splitting) คือกระบวนการที่แบ่งชุดข้อมูลใหญ่ออกเป็นหลายชุดย่อย โดยการแยกชุดข้อมูลมีความสำคัญอย่างยิ่งเพราะช่วยให้สามารถประเมินประสิทธิภาพของโมเดล และป้องกันปัญหาที่เรียกว่าการเรียนรู้มากเกินไป (overfitting) ชุดข้อมูลที่แยกออกมามักจะประกอบด้วย ชุดข้อมูลสำหรับการฝึกหัด (training set) เป็นชุดข้อมูลส่วนใหญ่ที่ใช้สำหรับการฝึกหัดโมเดล และชุดข้อมูลสำหรับการทดสอบ (test set) เป็นชุดข้อมูลที่ใช้เพื่อประเมินประสิทธิภาพของโมเดลหลังจากการฝึกฝนเสร็จสิ้น เพื่อให้เห็นภาพรวมของประสิทธิภาพของโมเดลกับข้อมูลที่ไม่เคยเห็นมาก่อน โดยในงานวิจัยนี้ผู้วิจัยได้ทำการแบ่งข้อมูลเป็นสัดส่วน 70:30, 75:25 และ 80:20

การเตรียมข้อมูลเป็นขั้นตอนสำคัญเพื่อให้กระบวนการทำเหมืองข้อมูลสามารถสกัดความรู้และข้อมูลที่มีค่าจากชุดข้อมูลใหญ่ได้อย่างมีประสิทธิภาพและถูกต้อง โดยการเตรียมข้อมูลจะช่วยปรับปรุงคุณภาพข้อมูลและเพิ่มประสิทธิภาพของการวิเคราะห์และโมเดลที่ได้

สามารถสรุปแผนผังกระบวนการทำเหมืองข้อมูลได้ดังรูปที่ 17



รูปที่ 17 กระบวนการการทำงานวิจัย

บทที่ 4 ผลการวิจัย

ในบทนี้จะนำเสนอผลการวิจัยของการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest จะนำเสนอเป็น 6 ส่วน โดยมีรายละเอียดดังนี้

4.1 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest

4.2 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 30%

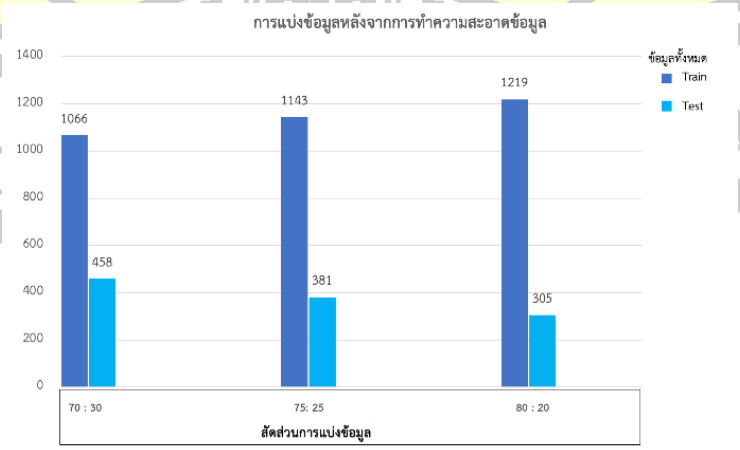
4.3 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 35%

4.4 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) และวิธีสุ่มลด (undersampling)

4.5 การหาแบบจำลองที่ดีที่สุดในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึม Random forest โดยวิธีสุ่มเกิน (oversampling) 35%

4.6 ปัจจัยที่ส่งผลในการจำแนกโรคมะเร็งเต้านม

4.1 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



รูปที่ 18 จำนวนข้อมูลฝึกหัดและข้อมูลทดสอบที่ใช้ในการวิเคราะห์ข้อมูล

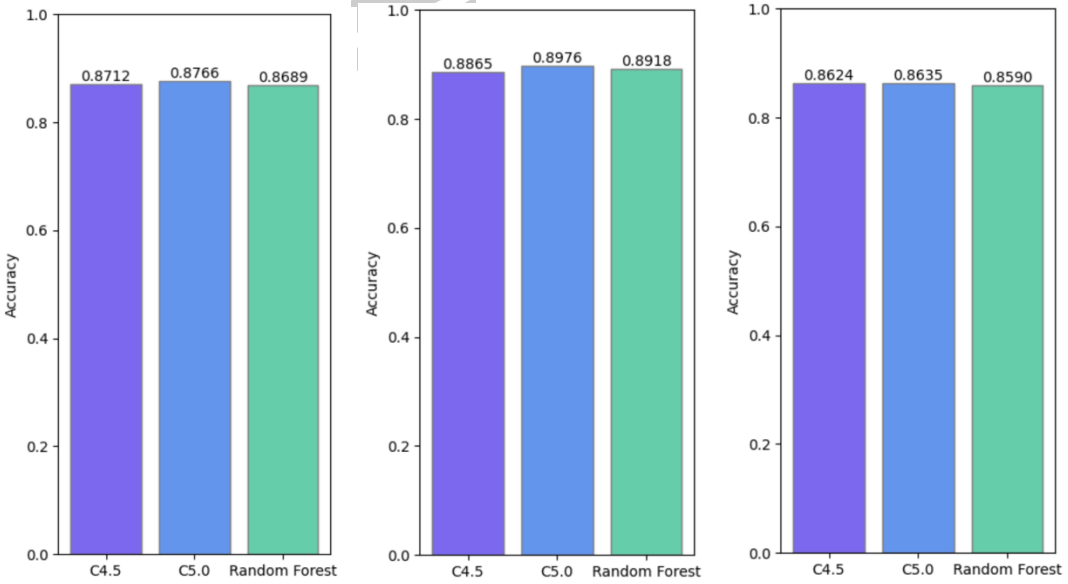
ในส่วนนี้เป็นการหาแบบจำลองโดยข้อมูลที่ทำความสะอาดเรียบร้อยแล้ว จำนวน 1,524 ระเบียน และแบ่งสัดส่วนข้อมูล 70:30, 75:25 และ 80:20 ดังรูปที่ 18 โดยแสดงผลลัพธ์ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูล (split data) แบบต่าง ๆ ดังตารางที่ 4

ตารางที่ 4 แสดงค่าความถูกต้อง, ค่า AUC และ Confusion Metrix ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5. C5.0 และวิธี Random forest และการแบ่งข้อมูลแบบต่าง ๆ

อัลกอริทึม	สัดส่วนการแบ่งข้อมูล	Confusion Metric		Accuracy	AUC
C4.5	70 : 30	0(TP)	0(FP)	87.12	0.50
		59(FN)	399(TN)		
	75 : 25	0	1	88.65	0.50
		51	406		
	80 : 20	1	5	86.24	0.54
		58	394		
C5.0	70 : 30	1	0	87.66	0.55
		47	333		
	75 : 25	0	0	89.76	0.50
		39	342		
	80 : 20	2	6	86.35	0.53
		46	327		
Random forest	70 : 30	0	0	86.89	**0.61**
		40	265		
	75 : 25	0	0	**89.18**	0.50
		33	272		
	80 : 20	2	5	85.90	0.48
		38	260		

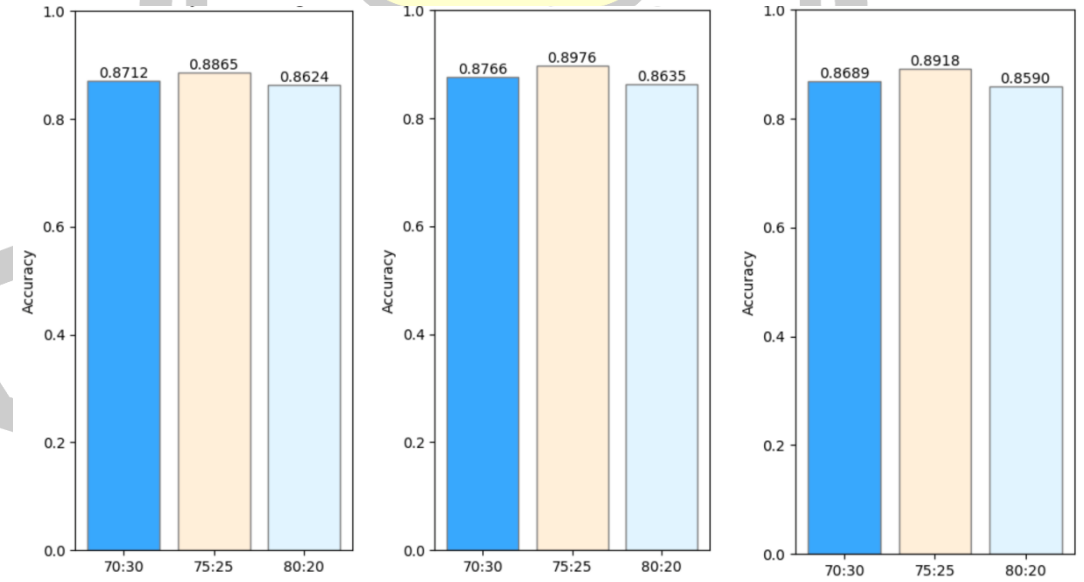
หมายเหตุ TP คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (true positive)
FN คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (false negative)
TN คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (true negative)
FP คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (false positive)

ในส่วนนี้จะแสดงค่าความถูกต้องและค่า AUC ของแต่ละวิธีของการวัดประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ดังรูปที่ 19-22



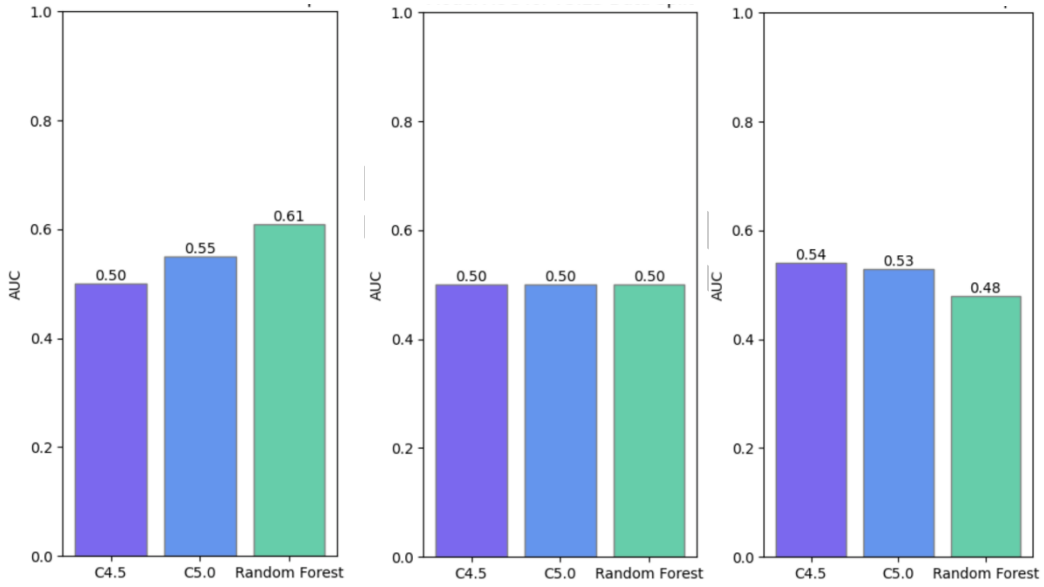
19(ก) การแบ่งข้อมูล 70:30 19(ข) การแบ่งข้อมูล 75:25 19(ค) การแบ่งข้อมูล 80:20

รูปที่ 19 ค่าความถูกต้องของการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



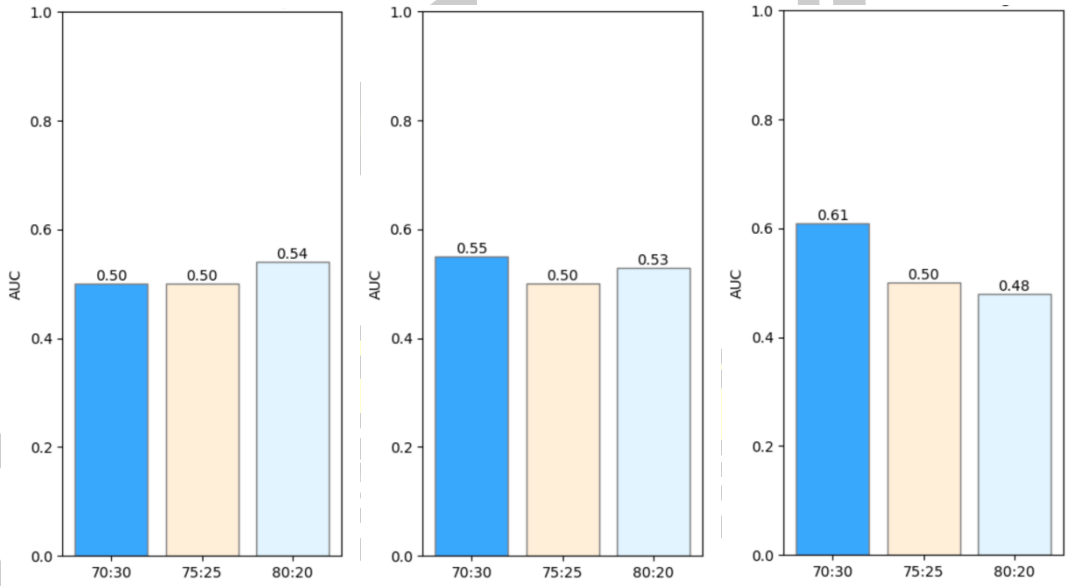
20(ก) ต้นไม้ตัดสินใจ C4.5 20(ข) ต้นไม้ตัดสินใจ C5.0 20(ค) Random forest

รูปที่ 20 ค่าความถูกต้องของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ



21(ก) การแบ่งข้อมูล 70:30 21(ข) การแบ่งข้อมูล 75:25 21(ค) การแบ่งข้อมูล 80:20

รูปที่ 21 ค่า AUC ของการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



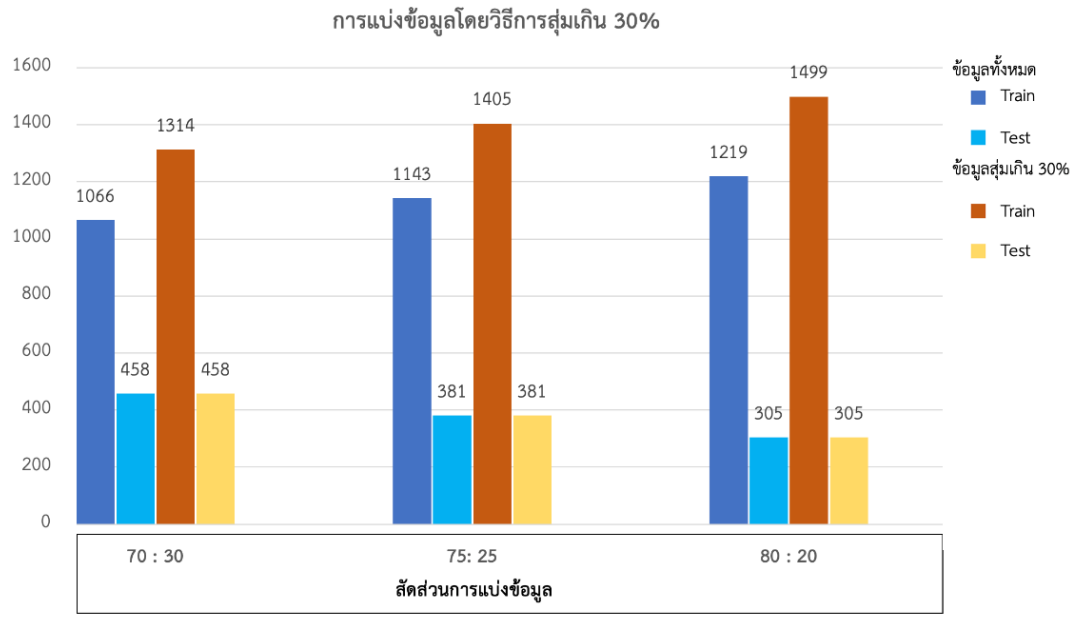
22(ก) ต้นไม้ตัดสินใจ C4.5 22(ข) ต้นไม้ตัดสินใจ C5.0 22(ค) Random forest

รูปที่ 22 ค่า AUC ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ

จากตารางที่ 4 แสดงผลลัพธ์ของการทดสอบโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยใช้การแบ่งข้อมูลเป็นสัดส่วนต่าง ๆ (70:30, 75:25, 80:20) ในการฝึกหัดและทดสอบโมเดล และวัดประสิทธิภาพด้วยค่าความถูกต้อง, ค่า AUC และ Confusion Metric โดยใช้ข้อมูลหลังการทำความสะอาด (clean data) ซึ่งแบ่งเป็นผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านมจำนวน 1,343 ระเบียน และผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านมจำนวน 181 ระเบียน สำหรับแต่ละอัลกอริทึมและการแบ่งข้อมูลที่แตกต่างกัน ผลการทดสอบแสดงให้เห็นว่าอัลกอริทึมต้นไม้ตัดสินใจ C5.0 มีค่าความถูกต้อง (accuracy) ที่สูงที่สุดในทุกกรณีของการแบ่งข้อมูล มีค่า AUC เป็น 0.5 ในทุกกรณี และมี True Positives (TP) และ True Negatives (TN) ที่แตกต่างกันในแต่ละรูปแบบการทดสอบ กล่าวคือเป็นการทำนายแบบสุ่มหรือไม่มีความแตกต่างในการทำนายระหว่างคลาสที่ต้องการทำนาย แสดงให้เห็นว่า C5.0 ไม่สามารถแยกแยะคลาสได้อย่างมีประสิทธิภาพกับข้อมูลที่ใช้ในการทดสอบ และการทำนายเป็นการสุ่มในการเลือกคลาส เป็นไปได้ที่จะไม่สามารถทำนายคลาสที่ต้องการแยกแยะได้อย่างถูกต้อง ซึ่งคล้ายกับอัลกอริทึมต้นไม้ตัดสินใจ C4.5 ซึ่งมีค่าความถูกต้องที่ไม่ต่างกันมากนักและมีค่า AUC ต่ำกว่าในบางกรณี ส่วนวิธี Random forest ให้ค่าความถูกต้องประมาณ 86% - 89% แต่ AUC จะมีค่าน้อยกว่าอัลกอริทึมต้นไม้ตัดสินใจ C5.0 ในบางกรณี เนื่องจาก Random forest มีความสามารถในการลดการเกิดการเรียนรู้มากเกินไป (overfitting) และมีความสามารถในการจัดการกับข้อมูลที่มีความซับซ้อนได้ดีกว่าอัลกอริทึมต้นไม้ตัดสินใจ C5.0 แต่ค่า AUC ที่ต่ำกว่าอาจเกิดขึ้นเนื่องจากการสุ่มที่มาจากการสร้างต้นไม้และคำตอบของแต่ละต้นใน Random forest ที่มีความแตกต่างกันในบางครั้งทำให้ค่า AUC ลดลงในบางกรณีเมื่อเปรียบเทียบกับอัลกอริทึมต้นไม้ตัดสินใจ C5.0 ที่สร้างต้นไม้แค่หนึ่งต้น ซึ่งในสถานการณ์ข้างต้นที่กล่าวมานี้เรียกว่าข้อมูลไม่สมดุล (class imbalance) ซึ่งอาจทำให้ค่าความถูกต้องมีค่าสูงมาก แต่ความสามารถในการทำนายคลาสที่น้อยมากและอาจไม่ดีเท่าที่ควร เนื่องจากโมเดลมักจะจำแนกคลาสที่มีจำนวนมาก เพราะมีข้อมูลมากกว่าในคลาสนั้น ในส่วนถัดไปจะนำเสนอวิธีการแก้ปัญหาข้อมูลไม่สมดุล โดยวิธีการสุ่มเกิน (oversampling) 30%

4.2 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 30%

ในส่วนนี้เป็นการหาแบบจำลองโดยข้อมูลที่ทำการสุ่มเกิน 30% โดยการแบ่งข้อมูลออกเป็นสัดส่วน 70:30, 75:25 และ 80:20 จากนั้นทำการสุ่มข้อมูลชุดฝึกหัด (training data) เพิ่ม 30% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,314 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,405 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,499 ระเบียน ดังรูปที่ 23 และแสดงผลลัพธ์ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูล (split data) แบบต่าง ๆ ดังตารางที่ 5



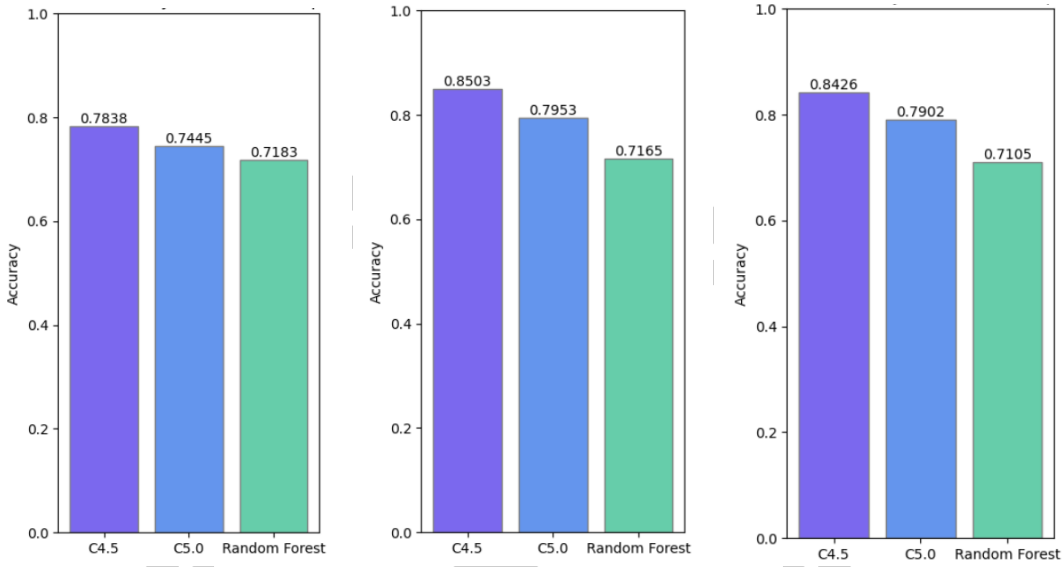
รูปที่ 23 จำนวนข้อมูลฝึกหัดและข้อมูลทดสอบที่ใช้ในการวิเคราะห์ข้อมูล โดยวิธีสุ่มเกิน (oversampling) 30%

ตารางที่ 5 แสดงค่าความถูกต้อง, ค่า AUC และ Confusion Metrix ของการสุ่มข้อมูลฝึกหัดเพิ่ม 30% ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูลแบบต่าง ๆ

อัลกอริทึม	สัดส่วนการแบ่งข้อมูล	Confusion Metric		Accuracy	AUC
C4.5	70 : 30	7(TP)	49(FP)	78.38	0.65
		50(FN)	352(TN)		
	75 : 25	6	13	**85.03**	0.62
		44	318		
	80 : 20	8	31	84.26	0.60
		32	234		
C5.0	70 : 30	13	73	74.45	0.57
		44	328		
	75 : 25	10	38	79.53	0.60
		40	293		
	80 : 20	10	33	79.02	0.62
		31	231		
Random forest	70 : 30	19	91	71.83	0.70
		38	310		
	75 : 25	13	71	71.65	0.68
		37	260		
	80 : 20	12	60	71.05	**0.69**
		28	204		

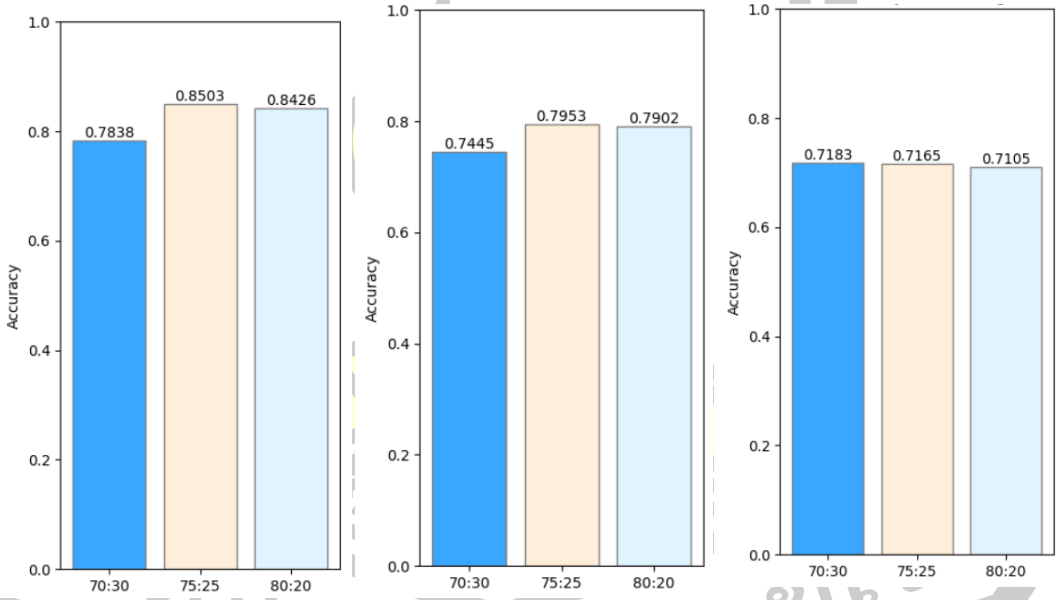
หมายเหตุ TP คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (true positive)
FN คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (false negative)
TN คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (true negative)
FP คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (false positive)

ในส่วนนี้จะแสดงค่าความถูกต้องและค่า AUC ของแต่ละวิธีของการวัดประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ดังรูปที่ 24-27



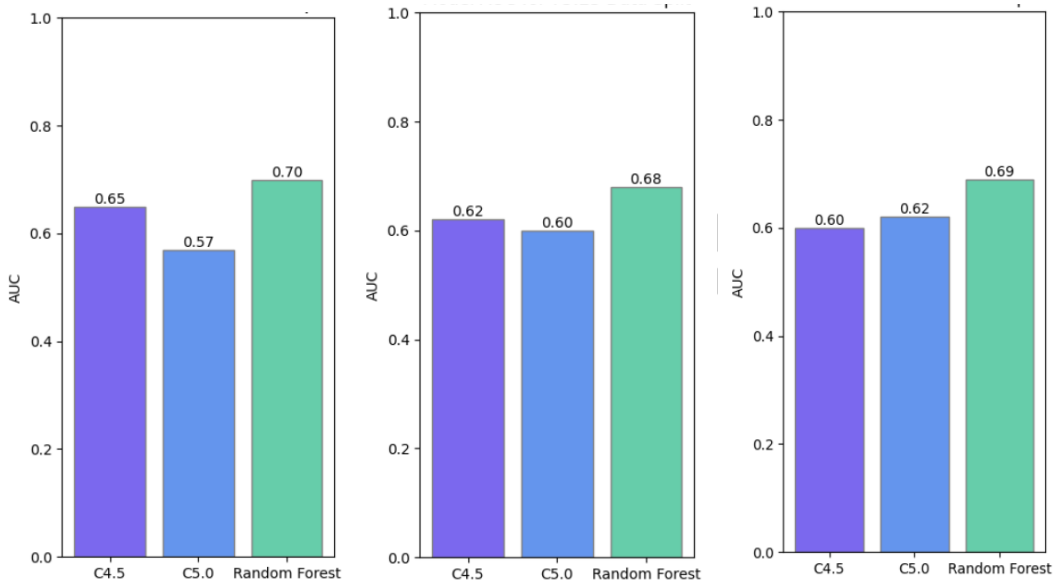
24(ก) การแบ่งข้อมูล 70:30 24(ข) การแบ่งข้อมูล 75:25 24(ค) การแบ่งข้อมูล 80:20

รูปที่ 24 ค่าความถูกต้องการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 30% ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



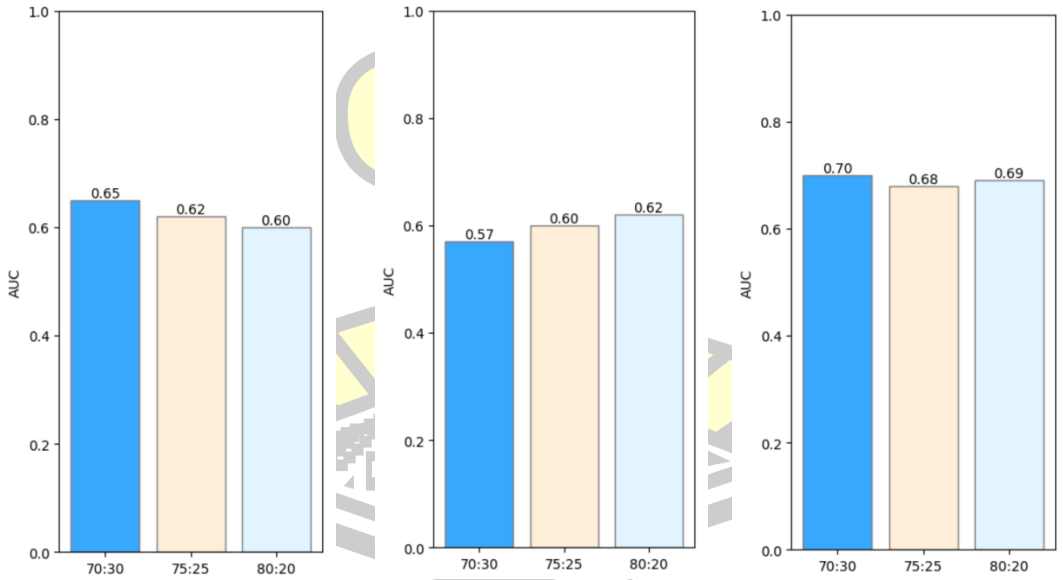
25(ก) ต้นไม้ตัดสินใจ C4.5 25(ข) ต้นไม้ตัดสินใจ C5.0 25(ค) Random forest

รูปที่ 25 ค่าความถูกต้องของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 30%



26(ก) การแบ่งข้อมูล 70:30 26(ข) การแบ่งข้อมูล 75:25 26(ค) การแบ่งข้อมูล 80:20

รูปที่ 26 ค่า AUC ของการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 30% ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



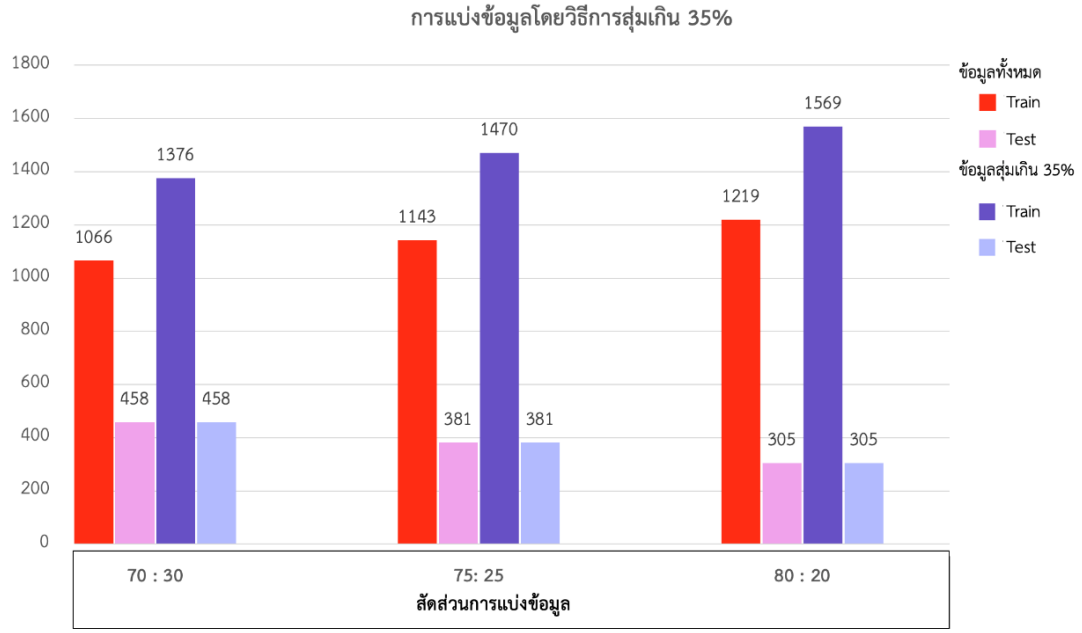
27(ก) ต้นไม้ตัดสินใจ C4.5 27(ข) ต้นไม้ตัดสินใจ C5.0 27(ค) Random forest

รูปที่ 27 ค่า AUC ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 30%

จากตารางที่ 5 แสดงผลลัพธ์ของการทดสอบโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยใช้การแบ่งข้อมูลเป็นสัดส่วนต่าง ๆ (70:30, 75:25, 80:20) ในการฝึกหัดและทดสอบโมเดล และวัดประสิทธิภาพด้วยค่าความถูกต้อง, ค่า AUC และ Confusion Metric โดยข้อมูลที่นำมาทำการวิเคราะห์เป็นข้อมูลชุดฝึกหัดที่ทำการสุ่มเพิ่ม 30% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,314 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,405 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,499 ระเบียน ในกรณีแบ่งข้อมูลเป็นสัดส่วน 70:30 อัลกอริทึมต้นไม้ตัดสินใจ C4.5 แสดงค่าความถูกต้องสูงสุดที่ 78.38% และค่า AUC ที่ 0.65 ซึ่งแสดงถึงความเหมาะสมในการจำแนกข้อมูลอยู่ในเกณฑ์มาตรฐานต่ำ ในขณะที่อัลกอริทึมต้นไม้ตัดสินใจ C5.0 มีค่าความถูกต้องประมาณ 74.45% แต่ค่า AUC อยู่ที่ 0.57 ซึ่งอาจเกิดจากลักษณะของข้อมูลที่ซับซ้อน ทำให้มีความยากในการแยกแยะคลาสหรือคุณลักษณะที่ใช้ในการสร้างต้นไม้ โดยที่วิธี Random forest ให้ค่าความถูกต้องอยู่ที่ 71.83% และค่า AUC อยู่ที่ 0.70 ซึ่งแสดงถึงความเหมาะสมในการจำแนกข้อมูลอยู่ในเกณฑ์มาตรฐานตัวแบบส่วนใหญ่ ซึ่งอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 เป็นอัลกอริทึมที่สร้างจากต้นไม้ต้นเดียว รวมทั้งใช้ข้อมูลทั้งหมดในการฝึกและไม่มีการสุ่มที่มาจากการสร้างต้นไม้หลายต้น ทำให้ค่าความถูกต้องสูงแต่ค่า AUC กลับสวนทางกัน ส่งผลให้การจำแนกคลาสไม่มีประสิทธิภาพ ในกรณีแบ่งข้อมูลเป็นสัดส่วน 75:25 อัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ยังคงแสดงค่าความถูกต้องค่อนข้างสูงอยู่ที่ 79% - 85% และค่า AUC อยู่ที่ 0.60 - 0.62 โดยที่วิธี Random forest ให้ค่าความถูกต้องและค่า AUC ไม่ต่างจากการแบ่งข้อมูลเป็นสัดส่วน 70:30 มากนัก และในกรณีแบ่งข้อมูลเป็นสัดส่วน 80:20 อัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีค่าความถูกต้องประมาณ 79% - 84% แต่ค่า AUC มีค่า 0.60 – 0.62 ในขณะที่วิธี Random forest มีค่าความถูกต้องประมาณ 71% และมีค่า AUC สูงถึง 0.69 ซึ่งจะเห็นได้ว่าการแบ่งข้อมูลมีผลต่อการฝึกและทดสอบโมเดล ซึ่งอาจทำให้ค่าประสิทธิภาพของโมเดลมีความแตกต่างกัน ผลการทดสอบทั้งหมดแสดงให้เห็นว่าวิธี Random forest มีค่าความถูกต้องและค่า AUC ที่สูงมากเช่นกัน โดยเฉพาะในกรณีการแบ่งข้อมูลสัดส่วน 70:30 และในส่วนถัดไปจะนำเสนอวิธีการแก้ปัญหาข้อมูลไม่สมดุล โดยวิธีการสุ่มเกิน (oversampling) 35% โดยเพิ่มจำนวนข้อมูลในกลุ่มที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม เพื่อให้ข้อมูลสมดุลขึ้นและช่วยในการเพิ่มประสิทธิภาพของโมเดลในการจำแนก

4.3 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 35%

ในส่วนนี้เป็นการหาแบบจำลองโดยข้อมูลที่ทำการสุ่มเกิน 35% โดยการแบ่งข้อมูลออกเป็นสัดส่วน 70:30, 75:25 และ 80:20 จากนั้นทำการสุ่มข้อมูลชุดฝึกหัด (training data) เพิ่ม 35% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,376 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,470 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,569 ระเบียน ดังรูปที่ 28 และแสดงผลลัพธ์ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูล (split data) แบบต่าง ๆ ดังตารางที่ 6



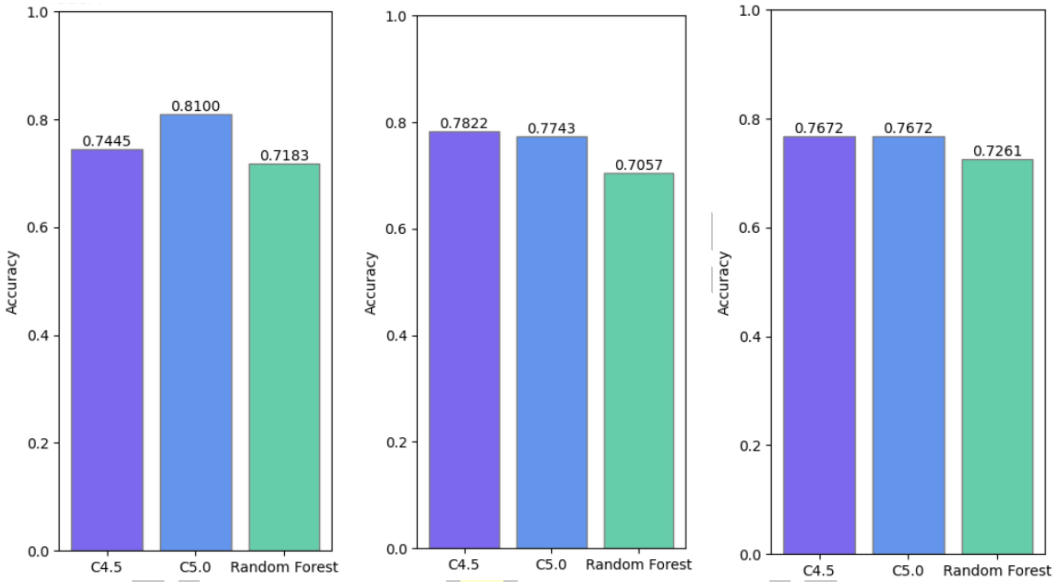
รูปที่ 28 จำนวนข้อมูลฝึกหัดและข้อมูลทดสอบที่ใช้ในการวิเคราะห์ข้อมูล โดยวิธีสุ่มเกิน (oversampling) 35%

ตารางที่ 6 แสดงค่าความถูกต้อง, ค่า AUC และ Confusion Metrix ของการสุ่มข้อมูลฝึกหัดเพิ่ม 35% ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูลแบบต่าง ๆ

อัลกอริทึม	สัดส่วนการแบ่งข้อมูล	Confusion Metric		Accuracy	AUC
C4.5	70 : 30	16(TP)	76(FP)	74.45	0.58
		41(FN)	325(TN)		
	75: 25	13	45	78.22	0.60
		38	285		
	80 : 20	15	45	76.72	0.64
		26	219		
C5.0	70 : 30	41	71	**81.00**	0.60
		16	330		
	75: 25	11	47	77.43	0.62
		39	284		
	80 : 20	9	39	76.72	0.61
		32	225		
Random forest	70 : 30	20	92	71.83	0.70
		37	309		
	75: 25	19	82	70.57	0.69
		31	252		
	80 : 20	18	61	72.61	**0.71**
		22	202		

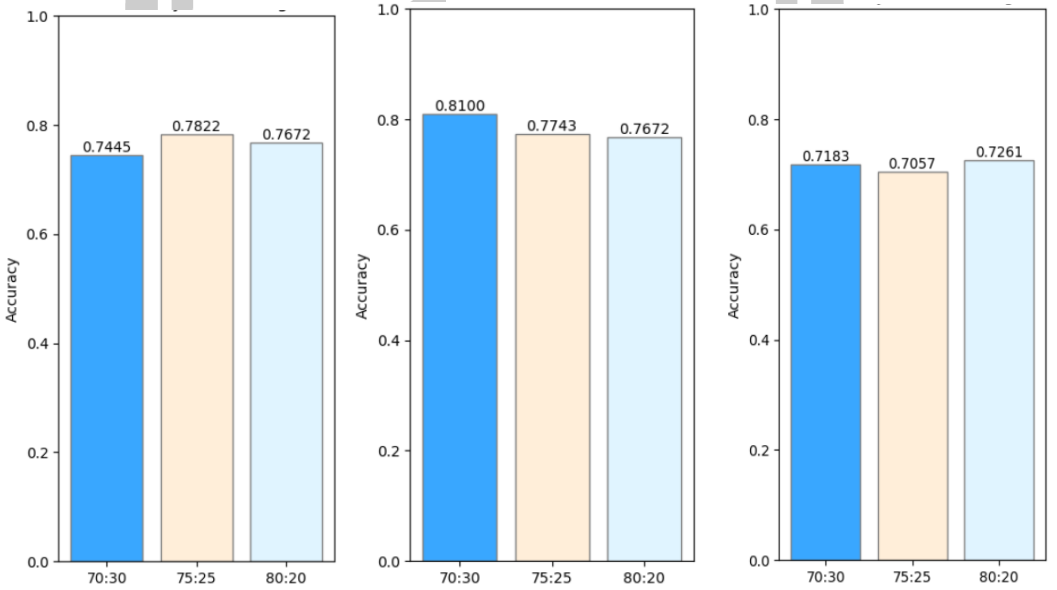
หมายเหตุ TP คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (true positive)
FN คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (false negative)
TN คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (true negative)
FP คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (false positive)

ในส่วนนี้จะแสดงค่าความถูกต้องและค่า AUC ของแต่ละวิธีของการวัดประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ดังรูปที่ 29-32



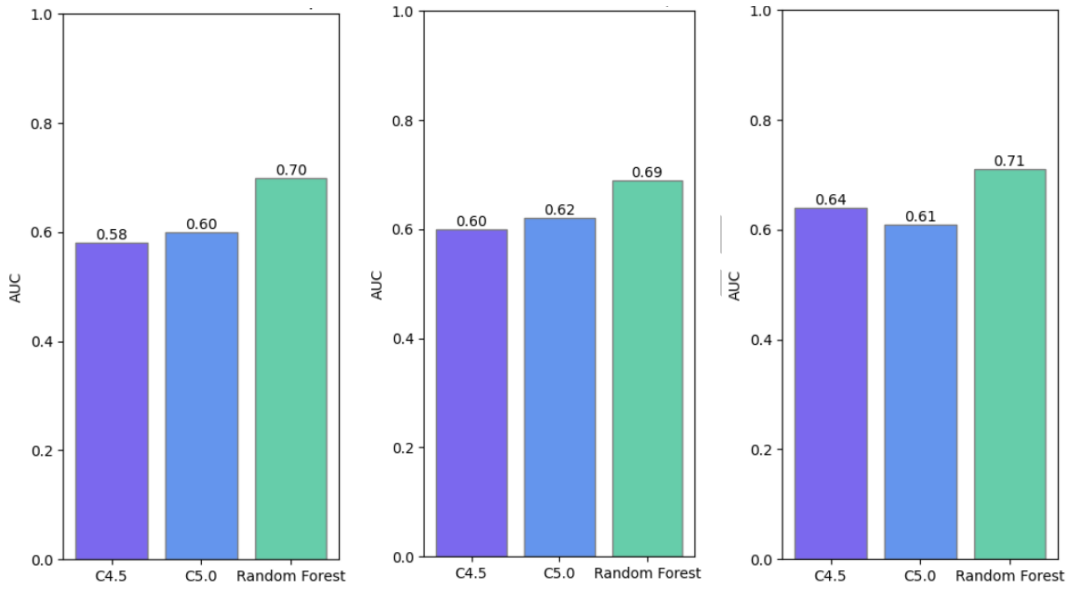
29(ก) การแบ่งข้อมูล 70:30 29(ข) การแบ่งข้อมูล 75:25 29(ค) การแบ่งข้อมูล 80:20

รูปที่ 29 ค่าความถูกต้องการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 35% ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



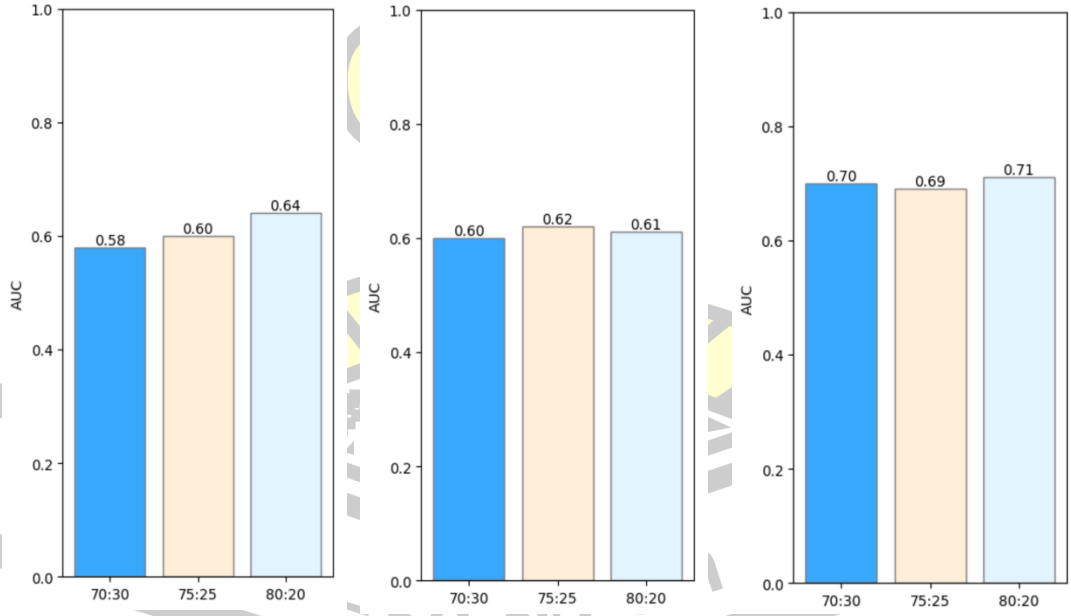
30(ก) ต้นไม้ตัดสินใจ C4.5 30(ข) ต้นไม้ตัดสินใจ C5.0 30(ค) Random forest

รูปที่ 30 ค่าความถูกต้องของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 35%



31(ก) การแบ่งข้อมูล 70:30 31(ข) การแบ่งข้อมูล 75:25 31(ค) การแบ่งข้อมูล 80:20

รูปที่ 31 ค่า AUC ของการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 35% ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



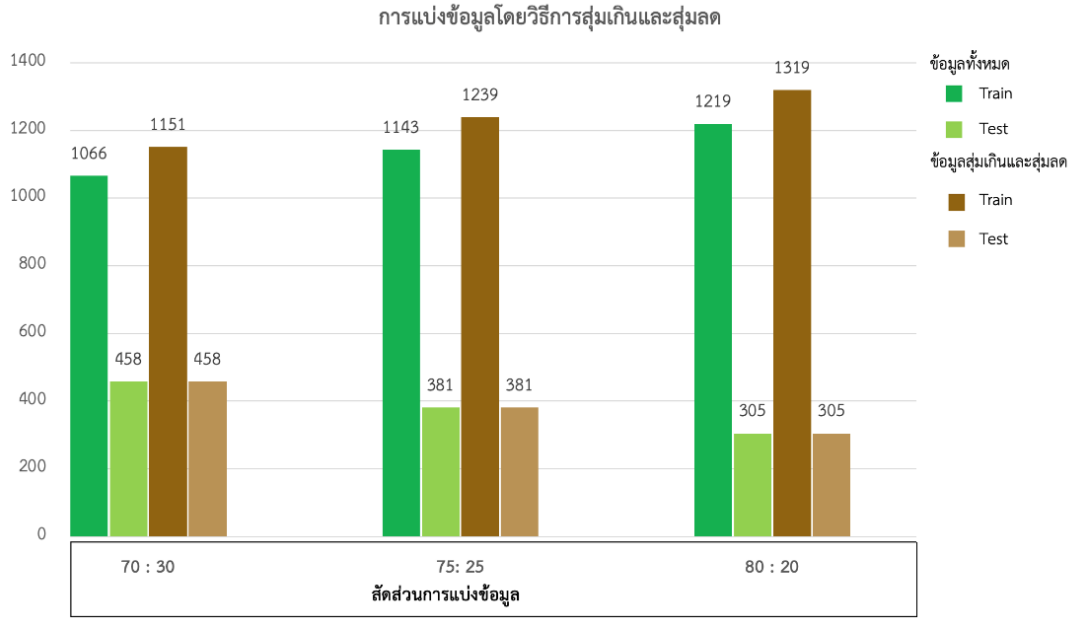
32(ก) ต้นไม้ตัดสินใจ C4.5 32(ข) ต้นไม้ตัดสินใจ C5.0 32(ค) Random forest

รูปที่ 32 ค่า AUC ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มข้อมูลฝึกหัดเพิ่ม 35%

จากตารางที่ 6 แสดงผลลัพธ์ของการทดสอบโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยใช้การแบ่งข้อมูลเป็นสัดส่วนต่าง ๆ (70:30, 75:25, 80:20) ในการฝึกหัดและทดสอบโมเดล และวัดประสิทธิภาพด้วยค่าความถูกต้อง , ค่า AUC, และ Confusion Metric โดยข้อมูลที่นำมาทำการวิเคราะห์เป็นข้อมูลฝึกหัดที่ทำการสุ่มเพิ่ม 35% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,376 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,470 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,569 ระเบียน สำหรับการแบ่งข้อมูลเป็นสัดส่วน 70:30 วิธี Random forest มีค่าความถูกต้องอยู่ที่ 71.83% และค่า AUC ที่ 0.70 ซึ่งแสดงถึงความเหมาะสมในการจำแนกข้อมูลอยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ เนื่องจากวิธี Random forest มีความสามารถในการจัดการกับข้อมูลที่ไม่สมดุล โดยสามารถจำแนกข้อมูลของกลุ่มความเสี่ยงสูงและความเสี่ยงต่ำได้ ส่วนอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีค่าความถูกต้องอยู่ในระดับประมาณ 74% - 81% แต่ค่า AUC อยู่ประมาณ 0.58 – 0.60 ซึ่งอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 สร้างต้นไม้ต้นเดียวและใช้ข้อมูลทั้งชุดในการฝึก ซึ่งอาจทำให้การจำแนกคลาสไม่มีประสิทธิภาพเมื่อเทียบกับการสร้างต้นไม้หลายต้น ในการแบ่งข้อมูลเป็นสัดส่วน 75:25 วิธี Random forest ยังคงแสดงค่าความถูกต้องและค่า AUC ไปในทิศทางเดียวกัน โดยมีค่าความถูกต้องอยู่ที่ 70.57% และค่า AUC ที่ 0.69 ซึ่งอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีค่าความถูกต้องประมาณ 77% - 78% แต่ค่า AUC อยู่ระหว่าง 0.60 - 0.62 และสำหรับการแบ่งข้อมูลเป็นสัดส่วน 80:20 วิธี Random forest แสดงค่าความถูกต้องประมาณ 72.61% และค่า AUC ที่ 0.71 ส่วนอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีค่าความถูกต้องในระดับประมาณ 76% และค่า AUC อยู่ที่ 0.61 – 0.64 จากผลการทดสอบที่กล่าวมาข้างต้น จะเห็นว่าวิธี Random forest เป็นวิธีที่มีประสิทธิภาพในการจำแนกข้อมูลในทุกกรณีที่ทดสอบ โดยมีค่าความถูกต้องและค่า AUC ที่สูงและเป็นไปในทิศทางเดียวกัน เมื่อเทียบกับอีก 2 อัลกอริทึม ซึ่งการใช้ข้อมูลสุ่มเกินเข้าไปเป็นการเพิ่มความสามารถในการจำแนกข้อมูลของโมเดลได้ และในส่วนถัดไปจะนำเสนอวิธีการเพื่อปรับปรุงประสิทธิภาพของโมเดลในการจำแนกข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเกิน (oversampling) และวิธีสุ่มลด (undersampling)

4.4 การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) และวิธีสุ่มลด (undersampling)

ในส่วนนี้เป็นการหาแบบจำลองโดยข้อมูลที่ทำการสุ่มเกินและสุ่มลด โดยการแบ่งข้อมูลออกเป็นสัดส่วน 70:30, 75:25 และ 80:20 จากนั้นทำการสุ่มข้อมูลชุดฝึกหัด (training data) สุ่มเกินและสุ่มลดในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,151 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,239 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,319 ระเบียน ดังรูปที่ 33 และแสดงผลลัพธ์ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูล (split data) แบบต่าง ๆ ดังตารางที่ 7



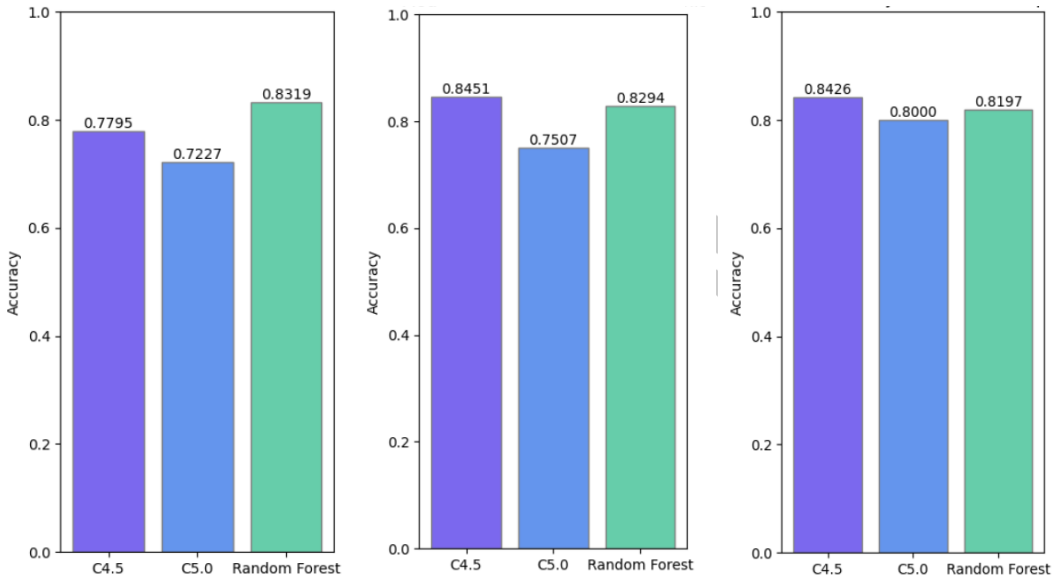
รูปที่ 33 จำนวนข้อมูลฝึกหัดและข้อมูลทดสอบที่ใช้ในการวิเคราะห์ข้อมูล โดยวิธีสุ่มเกินและวิธีสุ่มลด

ตารางที่ 7 แสดงค่าความถูกต้อง, ค่า AUC และ Confusion Metrix ของการสุ่มข้อมูลเพิ่มและการสุ่มข้อมูลลด ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest และการแบ่งข้อมูลแบบต่าง ๆ

อัลกอริทึม	สัดส่วนการแบ่งข้อมูล	Confusion Metric		Accuracy	AUC
C4.5	70 : 30	9(TP)	53(FP)	77.95	0.62
		48(FN)	348(TN)		
	75: 25	10	19	**84.51**	0.59
		40	312		
	80 : 20	9	16	84.26	**0.67**
		32	248		
C5.0	70 : 30	14	84	72.27	0.62
		43	317		
	75: 25	10	55	75.07	0.59
		40	276		
	80 : 20	11	32	80.00	0.61
		29	233		
Random forest	70 : 30	4	24	83.19	0.61
		53	377		
	75: 25	5	20	82.94	0.65
		45	311		
	80 : 20	3	17	81.97	**0.67**
		38	247		

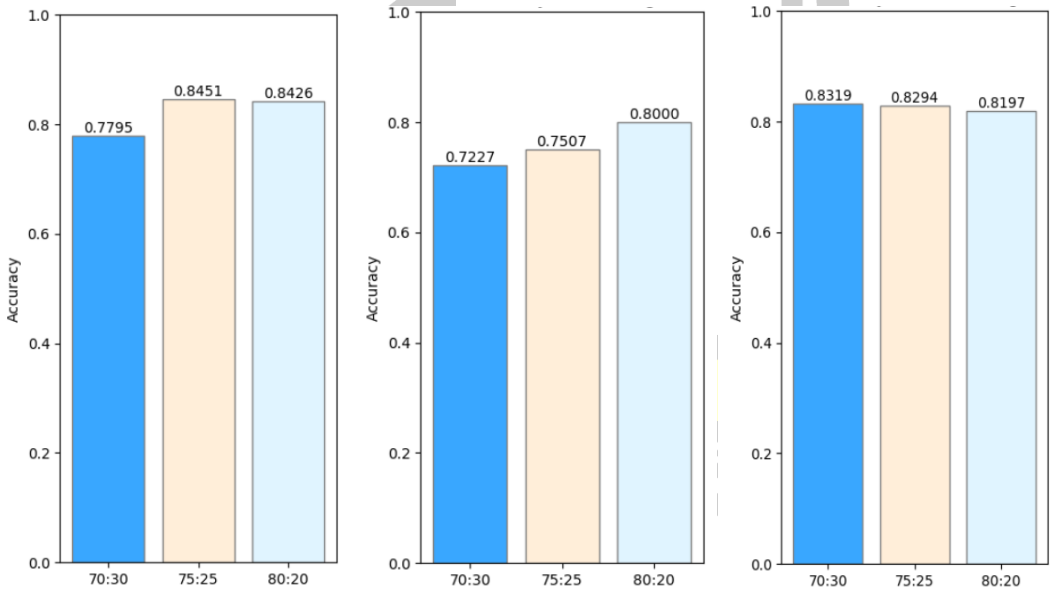
หมายเหตุ TP คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (true positive)
FN คือ ผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (false negative)
TN คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และทำนายว่า เป็นผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม (true negative)
FP คือ ผู้ที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม แต่ทำนายว่า เป็นผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (false positive)

ในส่วนนี้จะแสดงค่าความถูกต้อง และค่า AUC ของแต่ละวิธีของการวัดประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ดังรูปที่ 34-37



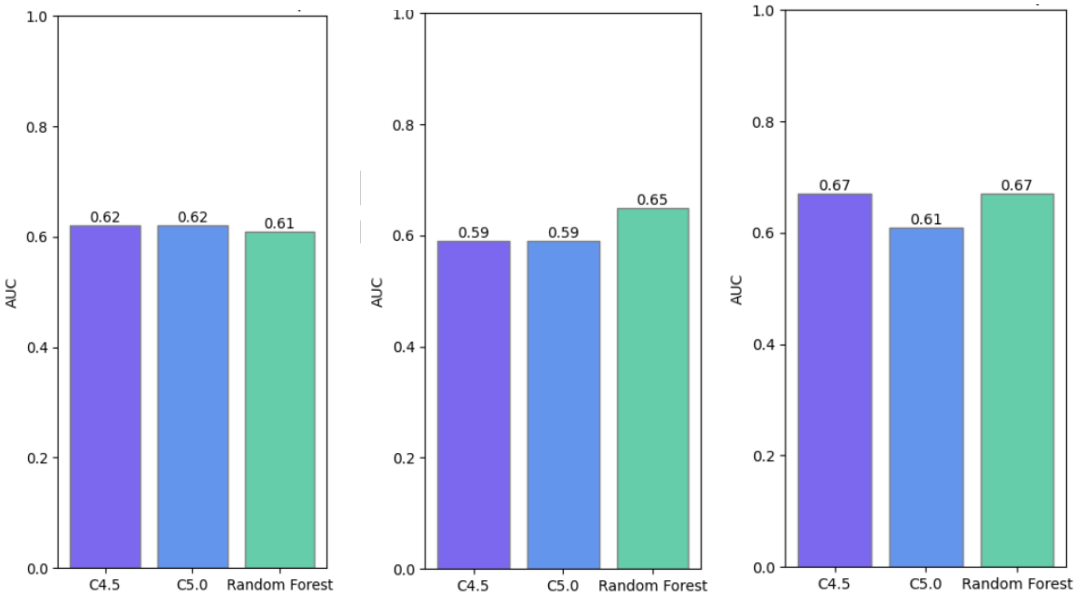
34(ก) การแบ่งข้อมูล 70:30 34(ข) การแบ่งข้อมูล 75:25 34(ค) การแบ่งข้อมูล 80:20

รูปที่ 34 ค่าความถูกต้องการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มเพิ่มข้อมูลและสุ่มลดข้อมูลฝึกหัด ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



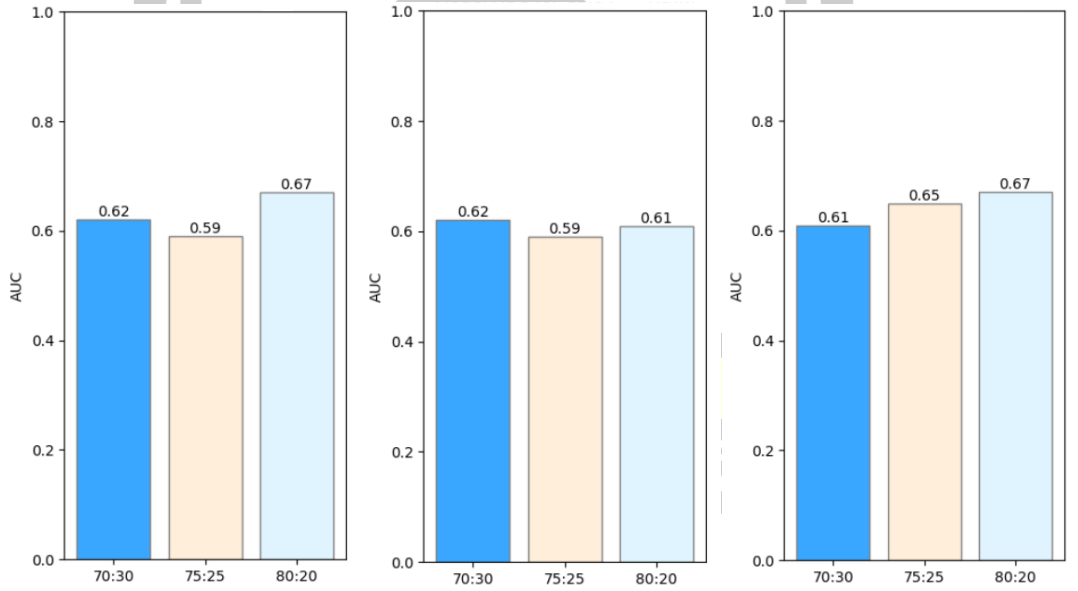
35(ก) ต้นไม้ตัดสินใจ C4.5 35(ข) ต้นไม้ตัดสินใจ C5.0 35(ค) Random forest

รูปที่ 35 ค่าความถูกต้องของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มเพิ่มข้อมูลและสุ่มลดข้อมูลฝึกหัด



36(ก) การแบ่งข้อมูล 70:30 36(ข) การแบ่งข้อมูล 75:25 36(ค) การแบ่งข้อมูล 80:20

รูปที่ 36 ค่า AUC การแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มเพิ่มข้อมูลและสุ่มลดข้อมูลฝึกหัด ด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest



37(ก) ต้นไม้ตัดสินใจ C4.5 37(ข) ต้นไม้ตัดสินใจ C5.0 37(ค) Random forest

รูปที่ 37 ค่า AUC ของอัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest ด้วยการแบ่งข้อมูลเป็นชุดข้อมูลฝึกและชุดข้อมูลทดสอบ โดยการสุ่มเพิ่มข้อมูลและสุ่มลดข้อมูลฝึกหัด

จากตารางที่ 7 แสดงผลลัพธ์ของการทดสอบโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และวิธี Random forest โดยใช้การแบ่งข้อมูลเป็นสัดส่วนต่าง ๆ (70:30, 75:25, 80:20) ในการฝึกหัดและทดสอบโมเดล และวัดประสิทธิภาพด้วยค่าความถูกต้อง, ค่า AUC และ Confusion Metric โดยข้อมูลที่นำมาทำการวิเคราะห์เป็นข้อมูลที่ทำการสุ่มเพิ่มและลดข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,151 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,239 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,319 ระเบียน พบว่า วิธี Random forest แสดงค่าความถูกต้องประมาณ 81% - 83% และค่า AUC ประมาณ 0.61 - 0.67 ซึ่งแตกต่างกันไปในแต่ละการแบ่งสัดส่วนของข้อมูล ส่วนอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ก็มีประสิทธิภาพในการจำแนกข้อมูล โดยมีค่าความถูกต้องประมาณ 72% - 84% และค่า AUC ประมาณ 0.59 - 0.67 ซึ่งการสุ่มข้อมูลเพิ่มและการสุ่มข้อมูลลดมีผลที่ไม่เหมือนกันในแต่ละโมเดลและในแต่ละกรณี ในบางกรณีการสุ่มข้อมูลเพิ่มอาจช่วยเพิ่มความแม่นยำของโมเดล ในขณะที่ในบางกรณีการสุ่มข้อมูลลดอาจทำให้ผลลัพธ์แย่ลง เนื่องจากการลดข้อมูลทำให้ข้อมูลที่มีความสำคัญสูญเสียไป อาจเป็นเหตุให้โมเดลเกิดการเรียนรู้ที่ไม่สมบูรณ์และแสดงผลลัพธ์ที่แย่ การแบ่งข้อมูลเป็นสัดส่วนต่าง ๆ มีผลต่อประสิทธิภาพของโมเดล โดยทั่วไปแล้วการแบ่งข้อมูล 80:20 ให้ผลลัพธ์ที่ดีกว่าเมื่อเทียบกับการแบ่งข้อมูล 70:30 และ 75:25 ในข้อมูลที่ทำการศึกษา ซึ่งอาจเป็นเพราะสัดส่วนข้อมูลการฝึกที่มากกว่าช่วยให้โมเดลมีโอกาสเรียนรู้และทดสอบข้อมูลที่มีความหลากหลายมากขึ้น

จากที่กล่าวมาทั้งหมด พบว่า วิธี Random forest ร่วมกับการสุ่มข้อมูลฝึกหัดเพิ่ม 35% ให้ค่าความถูกต้องและค่า AUC สูงที่สุด โดยมีค่าความถูกต้อง 72.61% และค่า AUC อยู่ที่ 0.71

4.5 การหาแบบจำลองที่ดีที่สุดในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึม Random forest โดยวิธีสุ่มเกิน (oversampling) 35%

ในส่วนนี้เป็นการหา K-fold cross-validation, ความลึกของต้นไม้ (tree depth) และจำนวนต้นไม้ (number of trees) เพื่อหาค่าที่ให้ประสิทธิภาพที่ดีที่สุด โดยที่ K-fold cross-validation ที่กำหนดคือ 3, 5, 7, 9 และ 10 จำนวนต้นไม้ที่กำหนดคือ 50, 75, 100, 150, 200, 300, 400, 500 และ 1000 ส่วนความลึกของต้นไม้ที่กำหนดคือ 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 และ 19 จะแสดงดังตารางที่ 8 – ตารางที่ 12

ตารางที่ 8 ผลลัพธ์การทดสอบค่าความลึกต้นไม้และจำนวนต้นไม้ในแบบจำลอง Random forest ด้วย k-fold cross-validation = 3 พร้อมค่าเฉลี่ยความถูกต้อง (mean accuracy) และค่าเฉลี่ย AUC (mean AUC)

ความลึกต้นไม้ (max dept)		จำนวนต้นไม้ (n)								
		50	75	100	150	200	300	400	500	1000
5	Mean Accuracy	0.7031	0.7031	0.7004	0.7004	0.7018	0.7011	0.7011	0.7018	0.7024
	Mean AUC	0.6326	0.6330	0.6310	0.6286	0.6288	0.6335	0.6320	0.6320	0.6314
6	Mean Accuracy	0.6997	0.7018	0.7018	0.7011	0.7011	0.7058	0.7051	0.7051	0.7065
	Mean AUC	0.6388	0.6438	0.6456	0.6438	0.6433	0.6458	0.6456	0.6463	0.6457
7	Mean Accuracy	0.7078	0.7051	0.7085	0.7045	0.7038	0.7031	0.7031	0.7038	0.7031
	Mean AUC	0.6622	0.6586	0.6594	0.6629	0.6607	0.6611	0.6607	0.6606	0.6601
8	Mean Accuracy	0.7038	0.7018	0.7045	0.7031	0.7038	0.7065	0.7065	0.7078	0.7099
	Mean AUC	0.6725	0.6749	0.6723	0.6729	0.6729	0.6733	0.6720	0.6722	0.6720
9	Mean Accuracy	0.7112	0.6991	0.6984	0.7011	0.7031	0.6930	0.7018	0.7038	0.7038
	Mean AUC	0.6802	0.6799	0.6823	0.6814	0.6798	0.6804	0.6783	0.6807	0.6774
10	Mean Accuracy	0.6889	0.6923	0.6916	0.6896	0.6869	0.6856	0.6856	0.6862	0.6876
	Mean AUC	0.6801	0.6782	0.6791	0.6777	0.6774	0.6784	0.6780	0.6787	0.6801
11	Mean Accuracy	0.6903	0.6876	0.6923	0.6883	0.6849	0.6856	0.6829	0.6822	0.6842
	Mean AUC	0.6814	0.6791	0.6802	0.6798	0.6781	0.6805	0.6805	0.6807	0.6807
12	Mean Accuracy	0.6930	0.6910	0.6889	0.6856	0.6856	0.6856	0.6842	0.6835	0.6822
	Mean AUC	0.6809	0.6801	0.6817	0.6814	0.6793	0.6812	0.6814	0.6818	0.6808
13	Mean Accuracy	0.6883	0.6889	0.6876	0.6842	0.6849	0.6829	0.6808	0.6822	0.6829
	Mean AUC	0.6822	0.6801	0.6813	0.6814	0.6805	0.6804	0.6806	0.6805	0.6808
14	Mean Accuracy	0.6903	0.6896	0.6876	0.6829	0.6842	0.6822	0.6808	0.6815	0.6829
	Mean AUC	0.6815	0.6799	0.6803	0.6802	0.6796	0.6809	0.6808	0.6803	0.6805
15	Mean Accuracy	0.6910	0.6896	0.6876	0.6829	0.6842	0.6822	0.6808	0.6822	0.6829
	Mean AUC	0.6813	0.6791	0.6798	0.6798	0.6796	0.6809	0.6808	0.6803	0.6805
16	Mean Accuracy	0.6903	0.6889	0.6876	0.6829	0.6835	0.6822	0.6808	0.6822	0.6829
	Mean AUC	0.6816	0.6791	0.6801	0.6797	0.6796	0.6809	0.6810	0.6803	0.6806
17	Mean Accuracy	0.6903	0.6889	0.6876	0.6829	0.6835	0.6822	0.6808	0.6822	0.6829
	Mean AUC	0.6816	0.6791	0.6801	0.6797	0.6796	0.6809	0.6810	0.6803	0.6806
18	Mean Accuracy	0.6903	0.6889	0.6876	0.6829	0.6835	0.6822	0.6808	0.6822	0.6829
	Mean AUC	0.6816	0.6791	0.6801	0.6797	0.6796	0.6809	0.6810	0.6803	0.6806
19	Mean Accuracy	0.6903	0.6889	0.6876	0.6829	0.6835	0.6822	0.6808	0.6822	0.6829
	Mean AUC	0.6816	0.6791	0.6801	0.6797	0.6796	0.6809	0.6810	0.6803	0.6806

ตารางที่ 9 ผลลัพธ์การทดสอบค่าความลึกต้นไม้และจำนวนต้นไม้ในแบบจำลอง Random forest ด้วย k-fold cross-validation = 5 พร้อมค่าเฉลี่ยความถูกต้อง (mean accuracy) และค่าเฉลี่ย AUC (mean AUC)

ความลึกต้นไม้ (max dept)		จำนวนต้นไม้ (n)								
		50	75	100	150	200	300	400	500	1000
5	Mean Accuracy	0.7071	0.7078	0.7065	0.7071	0.7078	0.7071	0.7071	0.7064	0.7071
	Mean AUC	0.6527	0.6510	0.6478	0.6507	0.6507	0.6502	0.6495	0.6489	0.6503
6	Mean Accuracy	0.7078	0.7112	0.7112	0.7145	0.7125	0.7118	0.7105	0.7112	0.7112
	Mean AUC	0.6591	0.6627	0.6614	0.6644	0.6636	0.6654	0.6643	0.6650	0.6660
7	Mean Accuracy	0.7166	0.7179	0.7166	0.7179	0.7166	0.7139	0.7145	0.7199	0.7172
	Mean AUC	0.6782	0.6759	0.6744	0.6786	0.6803	0.6794	0.6780	0.6779	0.6788
8	Mean Accuracy	0.7179	0.7206	0.7173	0.7179	0.7193	0.7152	0.7159	0.7159	0.7152
	Mean AUC	0.6859	0.6893	0.6887	0.6879	0.6878	0.6878	0.6870	0.6880	0.6891
9	Mean Accuracy	0.7105	0.7098	0.7112	0.7119	0.7112	0.7132	0.7105	0.7092	0.7105
	Mean AUC	0.6982	0.6979	0.6982	0.6970	0.6962	0.6959	0.6961	0.6960	0.6958
10	Mean Accuracy	0.6970	0.7017	0.7011	0.7011	0.7024	0.7058	0.7051	0.7058	0.7051
	Mean AUC	0.6978	0.6974	0.6971	0.6987	0.6992	0.6987	0.6982	0.6992	0.6994
11	Mean Accuracy	0.6990	0.6990	0.6984	0.6977	0.6984	0.7011	0.7004	0.7017	0.6984
	Mean AUC	0.6993	0.6994	0.6977	0.6986	0.7002	0.7003	0.7002	0.7000	0.7010
12	Mean Accuracy	0.6990	0.6957	0.6957	0.6977	0.6984	0.6983	0.7011	0.7017	0.6997
	Mean AUC	0.7013	0.6999	0.7004	0.7008	0.7017	0.7019	0.7007	0.7005	0.7011
13	Mean Accuracy	0.6970	0.6950	0.6943	0.6997	0.6970	0.6970	0.7004	0.7004	0.6997
	Mean AUC	0.7002	0.6999	0.6992	0.7008	0.7017	0.7016	0.7008	0.7004	0.7011
14	Mean Accuracy	0.6990	0.6977	0.6963	0.6990	0.6977	0.6977	0.7004	0.6997	0.6990
	Mean AUC	0.7016	0.7005	0.6998	0.7019	0.7019	0.7022	0.7010	0.7004	0.7011
15	Mean Accuracy	0.6990	0.6977	0.6957	0.6983	0.6970	0.6970	0.7004	0.6997	0.6997
	Mean AUC	0.7020	0.7006	0.6999	0.7017	0.7015	0.7023	0.7012	0.7004	0.7015
16	Mean Accuracy	0.6997	0.6984	0.6957	0.6983	0.6970	0.6970	0.7004	0.6997	0.6997
	Mean AUC	0.7020	0.7004	0.6999	0.7018	0.7014	0.7021	0.7011	0.7006	0.7015
17	Mean Accuracy	0.6997	0.6984	0.6957	0.6983	0.6970	0.6970	0.7004	0.6997	0.6997
	Mean AUC	0.7020	0.7006	0.6999	0.7018	0.7014	0.7021	0.7011	0.7006	0.7014
18	Mean Accuracy	0.6997	0.6984	0.6957	0.6983	0.6970	0.6970	0.7004	0.6997	0.6997
	Mean AUC	0.7020	0.7006	0.6999	0.7018	0.7014	0.7021	0.7011	0.7006	0.7014
19	Mean Accuracy	0.6997	0.6984	0.6957	0.6983	0.6970	0.6970	0.7004	0.6997	0.6997
	Mean AUC	0.7020	0.7006	0.6999	0.7018	0.7014	0.7021	0.7011	0.7006	0.7014

ตารางที่ 10 ผลลัพธ์การทดสอบค่าความลึกต้นไม้และจำนวนต้นไม้ในแบบจำลอง Random forest ด้วย k-fold cross-validation = 7 พร้อมค่าเฉลี่ยความถูกต้อง (mean accuracy) และค่าเฉลี่ย AUC (mean AUC)

ความลึกต้นไม้ (max dept)		จำนวนต้นไม้ (n)								
		50	75	100	150	200	300	400	500	1000
5	Mean Accuracy	0.7099	0.7078	0.7078	0.7092	0.7092	0.7085	0.7078	0.7078	0.7085
	Mean AUC	0.6556	0.6539	0.6545	0.6572	0.6588	0.6584	0.6574	0.6578	0.6561
6	Mean Accuracy	0.7099	0.7099	0.7051	0.7085	0.7078	0.7092	0.7105	0.7092	0.7092
	Mean AUC	0.6659	0.6636	0.6653	0.6677	0.6675	0.6684	0.6689	0.6689	0.6683
7	Mean Accuracy	0.7139	0.7132	0.7146	0.7153	0.7153	0.7159	0.7166	0.7132	0.7153
	Mean AUC	0.6842	0.6845	0.6844	0.6842	0.6859	0.6861	0.6863	0.6879	0.6877
8	Mean Accuracy	0.7187	0.7186	0.7186	0.7173	0.7193	0.7173	0.7173	0.7173	0.7173
	Mean AUC	0.6950	0.6982	0.6962	0.6977	0.6984	0.6996	0.6999	0.6990	0.6999
9	Mean Accuracy	0.7106	0.7160	0.7173	0.7173	0.7166	0.7153	0.7146	0.7173	0.7146
	Mean AUC	0.7083	0.7066	0.7073	0.7072	0.7077	0.7067	0.7079	0.7068	0.7079
10	Mean Accuracy	0.7112	0.7119	0.7133	0.7119	0.7173	0.7146	0.7153	0.7153	0.7173
	Mean AUC	0.7068	0.7084	0.7077	0.7075	0.7084	0.7084	0.7075	0.7078	0.7086
11	Mean Accuracy	0.7065	0.6984	0.7058	0.7092	0.7099	0.7072	0.7078	0.7072	0.7078
	Mean AUC	0.7065	0.7080	0.7065	0.7075	0.7079	0.7073	0.7076	0.7081	0.7086
12	Mean Accuracy	0.7038	0.6997	0.7017	0.7024	0.7119	0.7024	0.7058	0.7126	0.7092
	Mean AUC	0.7074	0.7085	0.7079	0.7087	0.7078	0.7079	0.7077	0.7086	0.7082
13	Mean Accuracy	0.7011	0.6984	0.6984	0.6997	0.7024	0.6997	0.7038	0.7038	0.7024
	Mean AUC	0.7075	0.7070	0.7069	0.7085	0.7079	0.7083	0.7080	0.7082	0.7075
14	Mean Accuracy	0.7024	0.6990	0.6990	0.7017	0.7031	0.6997	0.7038	0.7031	0.7024
	Mean AUC	0.7060	0.7071	0.7062	0.7085	0.7081	0.7084	0.7083	0.7087	0.7075
15	Mean Accuracy	0.7024	0.6977	0.6984	0.7011	0.7038	0.7004	0.7044	0.7038	0.7031
	Mean AUC	0.7060	0.7072	0.7062	0.7087	0.7082	0.7084	0.7081	0.7086	0.7076
16	Mean Accuracy	0.7024	0.6977	0.6984	0.7004	0.7031	0.6997	0.7044	0.7038	0.7031
	Mean AUC	0.7061	0.7072	0.7064	0.7085	0.7081	0.7085	0.7080	0.7086	0.7076
17	Mean Accuracy	0.7024	0.6977	0.6984	0.7004	0.7031	0.6997	0.7044	0.7038	0.7031
	Mean AUC	0.7061	0.7072	0.7064	0.7085	0.7081	0.7085	0.7080	0.7087	0.7076
18	Mean Accuracy	0.7024	0.6977	0.6984	0.7004	0.7031	0.6997	0.7044	0.7038	0.7031
	Mean AUC	0.7061	0.7072	0.7064	0.7085	0.7081	0.7085	0.7080	0.7087	0.7076
19	Mean Accuracy	0.7024	0.6977	0.6984	0.7004	0.7031	0.6997	0.7044	0.7038	0.7031
	Mean AUC	0.7061	0.7072	0.7064	0.7085	0.7081	0.7085	0.7080	0.7087	0.7076

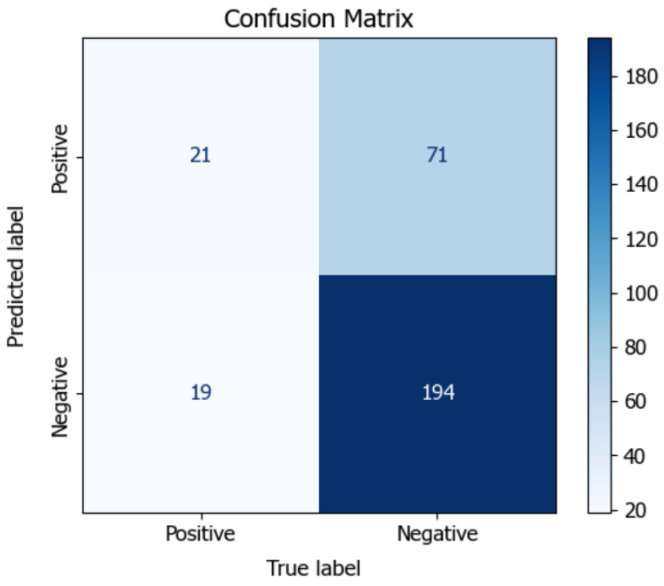
ตารางที่ 11 ผลลัพธ์การทดสอบค่าความลึกต้นไม้และจำนวนต้นไม้ในแบบจำลอง Random forest ด้วย k-fold cross-validation = 9 พร้อมค่าเฉลี่ยความถูกต้อง (mean accuracy) และค่าเฉลี่ย AUC (mean AUC)

ความลึกต้นไม้ (max dept)		จำนวนต้นไม้ (n)								
		50	75	100	150	200	300	400	500	1000
5	Mean Accuracy	0.7051	0.7064	0.7071	0.7071	0.7085	0.7085	0.7085	0.7085	0.7078
	Mean AUC	0.6505	0.6539	0.6544	0.6540	0.6529	0.6546	0.6532	0.6534	0.6545
6	Mean Accuracy	0.7105	0.7105	0.7105	0.7105	0.7105	0.7092	0.7098	0.7098	0.7098
	Mean AUC	0.6684	0.6684	0.6673	0.6677	0.6655	0.6671	0.6662	0.6659	0.6647
7	Mean Accuracy	0.7193	0.7193	0.7200	0.7193	0.7186	0.7200	0.7200	0.7193	0.7173
	Mean AUC	0.6857	0.6880	0.6860	0.6845	0.6824	0.6828	0.6818	0.6826	0.6803
8	Mean Accuracy	0.7186	0.7206	0.7213	0.7180	0.7193	0.7186	0.7180	0.7166	0.7173
	Mean AUC	0.6885	0.6938	0.6914	0.6907	0.6917	0.6924	0.6928	0.6932	0.6933
9	Mean Accuracy	0.7139	0.7166	0.7173	0.7146	0.7153	0.7159	0.7159	0.7166	0.7173
	Mean AUC	0.6982	0.7011	0.6985	0.6962	0.6988	0.6993	0.7005	0.7003	0.7000
10	Mean Accuracy	0.7153	0.7173	0.7119	0.7132	0.7139	0.7166	0.7146	0.7146	0.7153
	Mean AUC	0.7002	0.7055	0.7035	0.7022	0.7029	0.7037	0.7039	0.7036	0.7027
11	Mean Accuracy	0.7146	0.7132	0.7146	0.7112	0.7092	0.7139	0.7132	0.7146	0.7132
	Mean AUC	0.7013	0.7036	0.7024	0.7028	0.7029	0.7026	0.7020	0.7019	0.7024
12	Mean Accuracy	0.7119	0.7132	0.7071	0.7038	0.7031	0.7072	0.7065	0.7058	0.7105
	Mean AUC	0.7025	0.7043	0.7029	0.7016	0.7023	0.7029	0.7026	0.7022	0.7023
13	Mean Accuracy	0.7112	0.7085	0.7105	0.7045	0.7024	0.7058	0.7078	0.7058	0.7092
	Mean AUC	0.7032	0.7047	0.7030	0.7015	0.7026	0.7035	0.7029	0.7025	0.7016
14	Mean Accuracy	0.7044	0.7071	0.7078	0.7038	0.7011	0.7051	0.7058	0.7045	0.7099
	Mean AUC	0.7014	0.7038	0.7022	0.7015	0.7013	0.7041	0.7027	0.7018	0.7010
15	Mean Accuracy	0.7044	0.7065	0.7071	0.7031	0.7011	0.7051	0.7058	0.7038	0.7099
	Mean AUC	0.7028	0.7049	0.7037	0.7014	0.7018	0.7043	0.7027	0.7017	0.7016
16	Mean Accuracy	0.7044	0.7065	0.7071	0.7031	0.7011	0.7051	0.7058	0.7038	0.7099
	Mean AUC	0.7029	0.7050	0.7037	0.7014	0.7017	0.7043	0.7027	0.7019	0.7016
17	Mean Accuracy	0.7044	0.7065	0.7071	0.7031	0.7011	0.7051	0.7058	0.7038	0.7099
	Mean AUC	0.7029	0.7050	0.7037	0.7014	0.7017	0.7043	0.7027	0.7019	0.7015
18	Mean Accuracy	0.7044	0.7065	0.7071	0.7031	0.7011	0.7051	0.7058	0.7038	0.7099
	Mean AUC	0.7029	0.7050	0.7037	0.7014	0.7017	0.7043	0.7027	0.7019	0.7015
19	Mean Accuracy	0.7044	0.7065	0.7071	0.7031	0.7011	0.7051	0.7058	0.7038	0.7099
	Mean AUC	0.7029	0.7050	0.7037	0.7014	0.7017	0.7043	0.7027	0.7019	0.7015

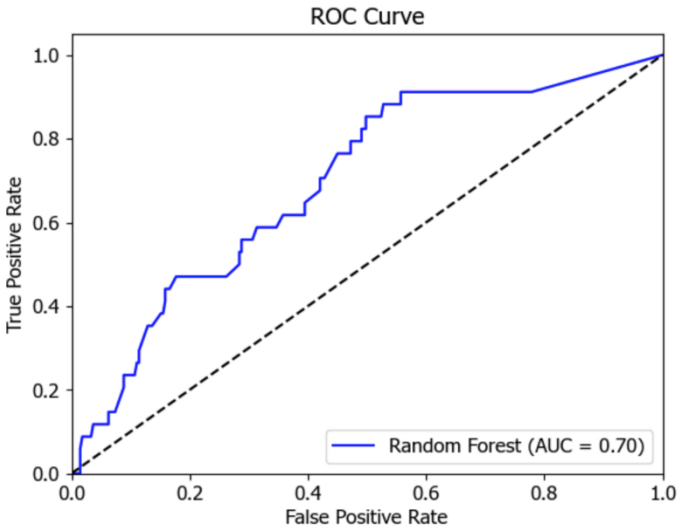
ตารางที่ 12 ผลลัพธ์การทดสอบค่าความลึกต้นไม้และจำนวนต้นไม้ในแบบจำลอง Random forest ด้วย k-fold cross-validation = 10 พร้อมค่าเฉลี่ยความถูกต้อง (mean accuracy) และค่าเฉลี่ย AUC (mean AUC)

ความลึกต้นไม้ (max dept)		จำนวนต้นไม้ (n)								
		50	75	100	150	200	300	400	500	1000
5	Mean Accuracy	0.7071	0.7071	0.7091	0.7085	0.7058	0.7058	0.7078	0.7085	0.7071
	Mean AUC	0.6490	0.6469	0.6502	0.6495	0.6515	0.6523	0.6512	0.6513	0.6520
6	Mean Accuracy	0.7078	0.7112	0.7125	0.7138	0.7125	0.7145	0.7145	0.7132	0.7125
	Mean AUC	0.6672	0.6665	0.6665	0.6652	0.6655	0.6669	0.6657	0.6661	0.6651
7	Mean Accuracy	0.7159	0.7179	0.7152	0.7179	0.7186	0.7159	0.7145	0.7152	0.7165
	Mean AUC	0.6805	0.6806	0.6800	0.6812	0.6818	0.6822	0.6825	0.6823	0.6823
8	Mean Accuracy	0.7138	0.7172	0.7179	0.7159	0.7152	0.7159	0.7145	0.7152	0.7152
	Mean AUC	0.6908	0.6919	0.6919	0.6930	0.6919	0.6935	0.6941	0.6957	0.6955
9	Mean Accuracy	0.7165	0.7172	0.7172	0.7186	0.7172	0.7192	0.7186	0.7179	0.7186
	Mean AUC	0.6957	0.6948	0.6982	0.6994	0.6994	0.7001	0.7012	0.7008	0.7012
10	Mean Accuracy	0.7118	0.7213	0.7152	0.7206	0.7206	0.7199	0.7179	0.7186	0.7192
	Mean AUC	0.6996	0.7003	0.7014	0.7037	0.7034	0.7050	0.7053	0.7051	0.7052
11	Mean Accuracy	0.7145	0.7125	0.7112	0.7125	0.7166	0.7193	0.7172	0.7179	0.7166
	Mean AUC	0.7024	0.7037	0.7040	0.7055	0.7038	0.7043	0.7060	0.7066	0.7062
12	Mean Accuracy	0.7152	0.7105	0.7064	0.7125	0.7092	0.7172	0.7145	0.7159	0.7139
	Mean AUC	0.7034	0.7038	0.7026	0.7043	0.7043	0.7060	0.7066	0.7059	0.7055
13	Mean Accuracy	0.7132	0.7112	0.7058	0.7105	0.7065	0.7125	0.7125	0.7125	0.7119
	Mean AUC	0.7015	0.7034	0.7026	0.7039	0.7030	0.7061	0.7064	0.7061	0.7047
14	Mean Accuracy	0.7132	0.7098	0.7064	0.7112	0.7071	0.7132	0.7139	0.7132	0.7071
	Mean AUC	0.7019	0.7025	0.7024	0.7043	0.7033	0.7065	0.7072	0.7074	0.7053
15	Mean Accuracy	0.7145	0.7098	0.7064	0.7119	0.7071	0.7125	0.7132	0.7132	0.7071
	Mean AUC	0.7023	0.7028	0.7028	0.7042	0.7037	0.7065	0.7076	0.7073	0.7055
16	Mean Accuracy	0.7152	0.7098	0.7064	0.7119	0.7085	0.7139	0.7139	0.7132	0.7071
	Mean AUC	0.7019	0.7028	0.7028	0.7042	0.7035	0.7070	0.7076	0.7074	0.7054
17	Mean Accuracy	0.7152	0.7098	0.7064	0.7119	0.7085	0.7139	0.7139	0.7132	0.7071
	Mean AUC	0.7022	0.7028	0.7028	0.7042	0.7036	0.7070	0.7076	0.7074	0.7054
18	Mean Accuracy	0.7152	0.7098	0.7064	0.7119	0.7085	0.7139	0.7139	0.7132	0.7071
	Mean AUC	0.7022	0.7028	0.7028	0.7042	0.7036	0.7070	0.7076	0.7074	0.7054
19	Mean Accuracy	0.7152	0.7098	0.7064	0.7119	0.7085	0.7139	0.7139	0.7132	0.7071
	Mean AUC	0.7022	0.7028	0.7028	0.7042	0.7036	0.7070	0.7076	0.7074	0.7054

จากตารางที่ 8 - ตารางที่ 12 เป็นการหาตัวแบบจำลองที่ดีที่สุดจากวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 35% ที่ให้ผลลัพธ์ที่ดีที่สุด โดยจะทำการหาจำนวนต้นไม้, 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 ผลลัพธ์แสดงให้เห็นว่าค่าที่ดีที่สุดคือ จำนวนต้นไม้เท่ากับ 150, ความลึกของต้นไม้เท่ากับ 15 และ K-fold cross-validation เท่ากับ 7 ซึ่งการปรับค่าพารามิเตอร์เหล่านี้ช่วยในการควบคุมความซับซ้อนของโมเดล โดยเฉพาะความลึกของต้นไม้ การเลือกค่าความลึกที่เหมาะสมสามารถป้องกันโมเดลจากการเรียนรู้ข้อมูลมากเกินไป (overfitting) หรือการไม่เรียนรู้เพียงพอ (underfitting) ซึ่งจะช่วยให้โมเดลมีประสิทธิภาพมากที่สุดในการทำนายข้อมูลที่ไม่เคยเห็นมาก่อน รวมทั้งการทดลองค่าพารามิเตอร์ต่าง ๆ ในขั้นตอนการฝึกและประเมินโมเดลสามารถช่วยประหยัดเวลาและทรัพยากรในกระบวนการพัฒนาโมเดล โดยไม่ต้องสร้างและทดลองทุก ๆ ค่าพารามิเตอร์ที่เป็นไปได้ หลังจากค้นพบค่าพารามิเตอร์ที่ดีที่สุดทำการฝึกและทดสอบโมเดล Random forest ด้วยการใช้ค่าพารามิเตอร์เหล่านี้ และประเมินประสิทธิภาพของโมเดลโดยใช้ข้อมูลที่ไม่เคยเห็นมาก่อน เพื่อให้แน่ใจว่าโมเดลทำงานได้ดีที่สุด พบว่าให้ค่าความถูกต้องในการทำนายคลาสต่าง ๆ บนชุดข้อมูลทดสอบที่ไม่เคยเห็นมาก่อนเท่ากับ 70.49% ดังรูปที่ 38 และค่า AUC อยู่ที่ 0.70 ดังรูปที่ 39



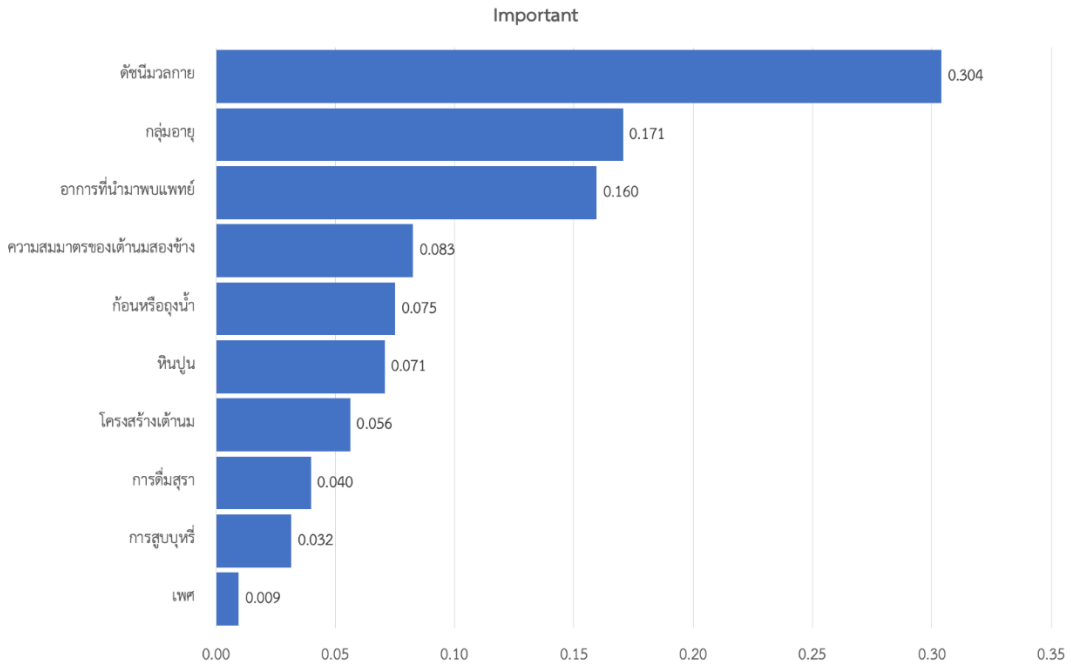
รูปที่ 38 Confusion Metrix ของวิธี Random forest ด้วยวิธีสุ่มเกิน (oversampling) 35% โดยจำนวนต้นไม้เท่ากับ 150, ความลึกของต้นไม้เท่ากับ 15 และ K-fold cross-validation เท่ากับ 7



รูปที่ 39 ค่า AUC ของวิธี Random forest ด้วยวิธีสุ่มเกิน (oversampling) 35% โดยจำนวนต้นไม้เท่ากับ 150, ความลึกของต้นไม้เท่ากับ 15 และ K-fold cross-validation เท่ากับ 7

4.6 ปัจจัยที่ส่งผลในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม

การจำแนกโรคมะเร็งเต้านมโดยใช้การหา K-fold ที่ดีที่สุดและการค้นหา Feature Importance เป็นกระบวนการทางวิทยาการข้อมูล (data science) ที่ใช้ในการวิเคราะห์ข้อมูลเพื่อตรวจสอบปัจจัยที่ใช้ในการจำแนกโรคมะเร็งเต้านม โดยการหา Feature Importance คือ การหาคุณสมบัติหรือคุณลักษณะของข้อมูลที่มีผลในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม คุณสมบัติที่มีค่าสูงจะมีผลในการบ่งบอกว่ามีความสัมพันธ์กับการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมมาก ส่วนคุณสมบัติที่มีค่าต่ำสามารถเรียกว่าไม่มีผลหรือมีผลน้อยในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ซึ่งปัจจัยที่ส่งผลต่อการจำแนกโรคมะเร็งเต้านม ได้แก่ ดัชนีมวลกาย, กลุ่มอายุ, อาการที่นำมาพบแพทย์, ความสมมาตรของเต้านมสองข้าง, ก้อนหรือถุงน้ำ, หินปูน, โครงสร้างเต้านม, การดื่มสุรา, การสูบบุหรี่ และเพศ ตามลำดับ แสดงดังรูปที่ 40 โดยที่ดัชนีมวลกายของผู้ที่มีน้ำหนักเกินสามารถจำแนกผู้ป่วยที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมได้ดีที่สุด และกลุ่มอายุ 33-45 ปี สามารถจำแนกที่ผู้ป่วยมีความเสี่ยงในการเป็นโรคมะเร็งเต้านมได้ดีที่สุด



รูปที่ 40 ปัจจัยที่ส่งผลในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม

ผลการศึกษาแสดงให้เห็นว่าวิธี Random forest มีค่าความถูกต้องและค่า AUC ที่ดีกว่าเมื่อเทียบกับอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม ผลลัพธ์นี้สนับสนุนสมมติฐานที่ว่า Random forest มีประสิทธิภาพสูงกว่าวิธีอัลกอริทึมต้นไม้ตัดสินใจ อีกทั้งผลการศึกษาแสดงให้เห็นว่าดัชนีมวลกาย (BMI) และอายุเป็นสองในห้าปัจจัยที่มีความสำคัญมากที่สุดในการจำแนกโรคมะเร็งเต้านม ข้อมูลนี้สอดคล้องกับสมมติฐานที่ว่าดัชนีมวลกายและอายุเป็นปัจจัยสำคัญในการจำแนกความเสี่ยงของการเป็นโรคมะเร็งเต้านม

บทที่ 5 สรุปผล อภิปรายผล และข้อเสนอแนะ

จากผลการวิเคราะห์การเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมด้วยอัลกอริทึมต้นไม้ตัดสินใจ C4.5. C5.0 และวิธี Random forest และการแบ่งข้อมูลแบบต่าง ๆ สามารถสรุปผลการศึกษา อภิปรายผล และข้อเสนอแนะได้ดังนี้

5.1 สรุปผล

5.2 อภิปรายผล

5.3 ข้อเสนอแนะ

5.1 สรุปผล

จากการศึกษาการเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ (decision tree algorithm) ในการจำแนกประเภทโรคมะเร็งเต้านม (breast cancer) และศึกษาปัจจัยเสี่ยงที่ส่งผลในการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม โดยใช้ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม จากคณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2553 ถึง พ.ศ. 2565 จำนวน 1,845 ระเบียน มีตัวแปรอิสระ (independent variable) ทั้งหมด 10 ตัวแปร ได้แก่ 1.เพศ (sex) 2. อายุ (age) 3. ดัชนีมวลกาย (BMI) 4. สูบบุหรี่ (smoking) 5. ดื่มสุรา (binge) 6. อาการที่นำมาพบแพทย์ (chief complaint) 7. ก้อนหรือถุงน้ำ (mass or cyst) 8. ความสมมาตรของเต้านมสองข้าง (asymmetries) 9. แคลเซียม (calcification) และ 10. โครงสร้างเต้านม (architectural distortion) และตัวแปรตาม (dependent variable) ได้แก่ ผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม และผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม มีข้อมูลทั้งหมด 1,845 ระเบียน โดยพบว่ามีค่าที่ขาดหายไป (missing value) 1.6% หลังทำความสะอาดข้อมูล (data cleaning) ซึ่งจะทำให้เหลือข้อมูลทั้งหมด 1,524 ระเบียน ซึ่งมีข้อมูลผู้ป่วยที่มีความเสี่ยงต่ำในการเป็นโรคมะเร็งเต้านม จำนวน 1,343 ระเบียน และข้อมูลผู้ป่วยที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม จำนวน 181 ระเบียน จากผลการศึกษาพบว่า อัลกอริทึมต้นไม้ตัดสินใจ C4.5, C5.0 และ Random forest ให้ค่าความถูกต้อง (accuracy) ค่อนข้างสูงอยู่ระหว่าง 74% - 89% แต่ค่าเกณฑ์ในการทำนาย AUC (area under ROC curve) กลับสวนทางกันอยู่ในช่วง 0.50 - 0.67 ซึ่งถือว่าค่อนข้างต่ำ เนื่องจากการทำนายโมเดลไม่สามารถแยกกลุ่ม (class) ได้ดีพอ โดยเฉพาะต่อการทำนายคลาสที่มีจำนวนน้อยกว่า (minority class) โมเดลอาจมีความลำเอียงในการทำนายและมักทำนายคลาสที่มีจำนวนมากกว่า (majority class) กล่าวคือเป็นการทำนายแบบสุ่มหรือไม่มีความแตกต่างในการทำนายระหว่างคลาสที่ต้องการทำนาย และการทำนายเป็นการสุ่มในการเลือกคลาสเป็นไปได้ที่จะไม่สามารถทำนายคลาสที่ต้องการแยกแยะได้อย่างถูกต้อง โดยปัญหาที่พบนี้เรียกว่าปัญหาข้อมูลไม่สมดุล (class imbalance) เพื่อทำการแก้ไขปัญหาข้อมูลไม่สมดุล ผู้วิจัยได้ทำการสุ่ม

เพิ่มข้อมูลกลุ่มที่มีจำนวนน้อยเพิ่ม โดยการแบ่งข้อมูลออกเป็นสัดส่วน 70:30, 75:25 และ 80:20 จากนั้นทำการสุ่มข้อมูลชุดฝึกหัด (training data) เพิ่ม 30% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,314 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,405 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,499 ระเบียน พบว่าอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ให้ค่าที่ไม่ต่างจากเดิมมากนัก โดยค่าความถูกต้องอยู่ที่ 74% - 85% และค่า AUC อยู่ที่ 0.57 - 0.65 แต่วิธี Random forest ให้ค่าความถูกต้องและค่า AUC ไปในทิศทางเดียวกัน โดยค่าความถูกต้อง อยู่ที่ประมาณ 71% และค่า AUC อยู่ที่ 0.68 – 0.70 ถือว่าอยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ ซึ่งผลลัพธ์ที่ได้ขึ้นอยู่กับการแบ่งข้อมูลออกเป็นชุดฝึก (training set) และชุดทดสอบ (test set) ถ้าความแตกต่างในชุดฝึกและชุดทดสอบมีความแตกต่างกันมาก ๆ อาจทำให้โมเดลมีประสิทธิภาพในการทำนายที่แตกต่างกัน ทำให้ค่าการทำนายมีผลที่แตกต่างกันเนื่องจากความแตกต่างในข้อมูลที่ใช้ในการทดสอบและวัดประสิทธิภาพ แม้ว่าผู้วิจัยจะได้เพิ่มข้อมูลในกลุ่มที่มีจำนวนน้อยแล้ว แต่การเพิ่ม 30% อาจยังไม่เพียงพอที่จะทำให้ข้อมูลมีความสมดุลขึ้น ผู้วิจัยจึงได้ทำการสุ่มเพิ่มข้อมูลฝึกหัดเพิ่ม 35% โดยการแบ่งข้อมูลออกเป็นสัดส่วน 70:30, 75:25 และ 80:20 จากนั้นทำการสุ่มข้อมูลชุดฝึกหัด (training data) เพิ่ม 35% ในแต่ละสัดส่วนของการแบ่งข้อมูล โดยสัดส่วนการแบ่งข้อมูล 70:30 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,066 ระเบียน เป็น 1,376 ระเบียน สัดส่วนการแบ่งข้อมูล 75:25 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,143 ระเบียน เป็น 1,470 ระเบียน และสัดส่วนการแบ่งข้อมูล 80:20 ข้อมูลชุดฝึกหัดเพิ่มขึ้นจาก 1,219 ระเบียน เป็น 1,569 ระเบียน สำหรับการแบ่งข้อมูล 70:30 พบว่าวิธี Random forest แสดงค่าความถูกต้องที่ 71.83% และ AUC ที่ 70% ขณะที่ C4.5 และ C5.0 มีค่าค่าความถูกต้อง อยู่ที่ 74% - 81% และ AUC อยู่ที่ 0.58 – 0.60 ซึ่งต่ำกว่าวิธี Random forest สำหรับการแบ่งข้อมูล 75:25 วิธี Random forest ยังคงแสดงผลลัพธ์ไปในทิศทางเดียวกัน ด้วยค่าความถูกต้องที่ 70.57% และค่า AUC ที่ 0.69 สำหรับการแบ่งข้อมูล 80:20 วิธี Random forest แสดงผลดีที่สุดด้วยค่าค่าความถูกต้องที่ 72.61% และค่า AUC อยู่ที่เป็น 0.71 ซึ่งวิธี Random forest แสดงผลลัพธ์ที่ดีที่สุดในทุกสัดส่วนการแบ่งข้อมูล เมื่อดูจากค่าความถูกต้องและค่า AUC ร่วมกัน อย่างไรก็ตามแม้ว่าผู้วิจัยจะเพิ่มข้อมูลในกลุ่มที่มีจำนวนน้อยแล้ว แต่ยังไม่ทราบว่าการเพิ่มข้อมูลนั้นมีผลต่อความแม่นยำของโมเดลหรือไม่ ผู้วิจัยจึงทำการสุ่มข้อมูลเพิ่มและลดข้อมูล (oversampling and undersampling) ซึ่งเป็นวิธีการจัดการกับปัญหาข้อมูลไม่สมดุล (class imbalance) ที่ได้รับความนิยมอีกหนึ่งวิธี โดยหลังการสุ่มข้อมูลเพิ่มและลดข้อมูล พบว่าวิธี Random forest ให้ค่าความถูกต้องระหว่าง 81% - 83% และมีค่า AUC อยู่ที่ 0.61 – 0.67 ซึ่งเป็นค่าที่สูงที่สุดในทุกวิธีการแบ่งข้อมูล สำหรับอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ให้ค่าความถูกต้องอยู่ที่ 72% - 84% และค่า AUC อยู่ที่ 0.59 – 0.67 โดย

วิธี Random forest ในการสุ่มข้อมูลเพิ่มและลดข้อมูลสามารถนำไปใช้ในการจำแนกความเสี่ยงของการเป็นโรคมะเร็งเต้านมได้ เนื่องจากให้ค่าความถูกต้องและค่า AUC ไปในทิศทางเดียวกันแต่ยังน้อยการการสุ่มข้อมูลฝึกหัดเพิ่ม 35% แต่ในการลดข้อมูลอาจทำให้ข้อมูลที่สำคัญถูกลดลงและข้อมูลสูญหายไป ทำให้โมเดลเสียโอกาสในการเรียนรู้จากข้อมูลทั้งหมดที่มีในคลาสที่มีจำนวนมาก และเนื่องจากข้อมูลที่มีจำนวนน้อยแล้ว การลดข้อมูลอาจไม่เป็นทางเลือกที่ดีเพราะอาจทำให้ข้อมูลน้อยเกินไปที่จะฝึกฝนโมเดลให้มีประสิทธิภาพ

วิธีการแก้ปัญหาข้อมูลไม่สมดุลทั้งหมดที่กล่าวมาจะเห็นว่าการสุ่มข้อมูลฝึกหัดเพิ่ม (oversampling) 35% ร่วมกับวิธี Random forest ให้ผลลัพธ์ที่ดีกว่าเมื่อเทียบวิธีต้นไม้ตัดสินใจ C4.5 และ C5.0 เนื่องจาก Random forest เป็นอัลกอริทึมที่สร้างหลายต้นไม้และทำการโหวตผลลัพธ์จากแต่ละต้นไม้เพื่อตัดสินใจ ทำให้เกิดความหลากหลายในการตัดสินใจ และลดความแปรปรวนที่อาจเกิดขึ้นจากการใช้ต้นไม้ตัดสินใจเพียงต้นเดียวที่เกิดจากความไม่แน่นอนหรือความผิดพลาดของข้อมูล เมื่อพิจารณาเฉพาะวิธีการสุ่มข้อมูลเพิ่ม (oversampling) 35% สามารถสรุปได้ว่าการเพิ่มจำนวนข้อมูลของคลาสที่มีจำนวนน้อยเพื่อให้มีความสมดุลมากขึ้นจะช่วยในการปรับปรุงประสิทธิภาพของโมเดล โดยเฉพาะเมื่อใช้ร่วมกับวิธี Random forest ซึ่งมีโครงสร้างของการทำงานที่รองรับความหลากหลายของข้อมูลและสามารถจัดการกับความไม่แน่นอนในข้อมูลได้อย่างมีประสิทธิภาพ ในขณะที่อัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ก็เป็นเครื่องมือที่มีประสิทธิภาพในการจัดการข้อมูลและทำนายผล แต่เมื่อเทียบกับวิธี Random forest ในสถานการณ์ที่มีข้อมูลไม่สมดุล วิธี Random forest มีแนวโน้มที่จะให้ผลการทำนายที่ดีกว่า นอกจากนี้วิธี Random forest ยังมีข้อดีในการจัดการกับคุณลักษณะ (features) ที่มีคุณลักษณะต่างกัน การหลีกเลี่ยงการเกิดการเรียนรู้ข้อมูลเกินไป (overfitting) และสามารถแสดงภาพรวมของความสำคัญของแต่ละปัจจัยได้

จากนั้นทำการหาตัวแบบจำลองที่ดีที่สุดจากวิธี Random forest โดยวิธีสุ่มเกิน (oversampling) 35% ที่ให้ผลลัพธ์ที่ดีที่สุด โดยจะทำการหาจำนวนต้นไม้, ความลึกของต้นไม้ และ K-fold cross-validation เพื่อประเมินประสิทธิภาพของโมเดลในแต่ละค่าพารามิเตอร์ ซึ่งจำนวนต้นไม้ที่กำหนดคือ 50, 75, 100, 150, 200, 300, 400, 500 และ 1000 ความลึกของต้นไม้ที่กำหนดคือ 4, 5, 6, ..., 19 ส่วนสุดท้าย K-fold cross-validation ที่กำหนดคือ 3, 5, 7, 9 และ 10 ผลลัพธ์แสดงให้เห็นว่าค่าที่ดีที่สุดคือ จำนวนต้นไม้เท่ากับ 150, ความลึกของต้นไม้เท่ากับ 15 และ K-fold cross-validation เท่ากับ 7 และประเมินประสิทธิภาพของโมเดลโดยใช้ข้อมูลที่ไม่เคยเห็นมาก่อนเพื่อให้แน่ใจว่าโมเดลทำงานได้ดีที่สุด พบว่าให้ค่าความถูกต้องในการทำนายคลาสต่าง ๆ บนชุดข้อมูลทดสอบที่ไม่เคยเห็นมาก่อนเท่ากับ 70.49% และค่า AUC อยู่ที่ 0.70 ซึ่งปัจจัยที่ส่งผลในการจำแนกโรคมะเร็งเต้านม 5 อันดับแรก ได้แก่ ดัชนีมวลกาย, กลุ่มอายุ, อาการที่นำมาพบแพทย์, ความสมมาตรของเต้านมสองข้าง และก้อนหรือถุงน้ำ ตามลำดับ

5.2 อภิปรายผล

จากผลการวิจัยการเปรียบเทียบประสิทธิภาพอัลกอริทึมต้นไม้ตัดสินใจในการจำแนกโรคมะเร็งเต้านมโดยการสังเคราะห์ข้อมูลจากเวชระเบียนของผู้ป่วยที่มีก้อนเนื้อบริเวณเต้านม ในส่วนของการเปรียบเทียบประสิทธิภาพวิธีต้นไม้ตัดสิน C4.5 ,C5.0 และวิธี Random forest พบว่าข้อมูลที่ใช้ในการจำแนกคลาสมีจำนวนของคลาสมากน้อยไม่เท่ากัน (class imbalance) ซึ่งอาจทำให้โมเดลที่ได้สร้างขึ้นมีความสามารถในการทำนายคลาสที่มีจำนวนตัวอย่างมาก มากกว่าคลาสที่มีจำนวนตัวอย่างน้อย ๆ นำมาซึ่งผลลัพธ์ที่ไม่เสถียรและไม่แม่นยำในคลาสที่มีจำนวนตัวอย่างน้อย โดยปัญหานี้เกิดขึ้นได้ทั่วไปโดยเฉพาะข้อมูลทางการแพทย์ เนื่องจากการตรวจจับโรคหรือความเสี่ยงของโรคที่มีอัตราการเกิดต่อประชากรต่ำ ทำให้คลาสของผู้ป่วยหรือผู้ที่มีความเสี่ยงสูงมีจำนวนน้อยเมื่อเปรียบเทียบกับคลาสของผู้ที่ไม่มีโรคหรือมีความเสี่ยงต่ำ เพื่อแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้ใช้เทคนิคการสุ่มเพิ่ม (oversampling) เพื่อเพิ่มจำนวนตัวอย่างในคลาสที่น้อยเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน และเทคนิคการสุ่มลด (undersampling) ลดตัวอย่างในคลาสที่มีจำนวนมากลงเพื่อทำให้จำนวนตัวอย่างในทุกคลาสเท่ากันหรือใกล้เคียงกัน รวมทั้งการแบ่งข้อมูล (split data) แบบต่าง ๆ ทำให้ค่าการทำนายมีผลต่างกันเนื่องจากความแตกต่างในชุดข้อมูลที่ถูกใช้ในการสร้างและทดสอบโมเดล และความแตกต่างในแบบจำลองที่ใช้ในการวิเคราะห์ข้อมูล จากการแก้ปัญหาจะเห็นว่าอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ซึ่งให้ผลลัพธ์ที่ไม่ต่างจากเดิมและผลลัพธ์ที่ได้ไม่ต่างกันมากนัก ไม่ว่าจะเป็นการใช้เทคนิคการสุ่มข้อมูลฝึกหัดเพิ่ม 30%, เทคนิคการสุ่มข้อมูลฝึกหัดเพิ่ม 35% และเทคนิคการสุ่มเพิ่มและเทคนิคการสุ่มลด อาจเกิดขึ้นเพราะสาเหตุต่าง ๆ ที่มีผลต่อประสิทธิภาพของโมเดล เนื่องจากอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 เป็นวิธีการสร้างต้นไม้ตัดสินใจซึ่งมีความเหมือนกันในหลักการการสร้างต้นไม้ตัดสินใจโดยการทำงานของเทคนิคการสุ่มข้อมูลเพิ่มหรือการสุ่มลดข้อมูลอาจไม่มีผลต่อต้นไม้ตัดสินใจถ้าข้อมูลมีความซับซ้อนสูง และการเพิ่มข้อมูลฝึกหัด 30% และ 35% หรือการลดข้อมูลอาจจะไม่มีผลต่อประสิทธิภาพโมเดลมากนัก เนื่องจากการเปลี่ยนแปลงนี้อาจเป็นเพียงแค่การเปลี่ยนแปลงเล็กน้อยที่ไม่สามารถส่งผลต่อประสิทธิภาพโมเดล ส่วนวิธี Random forest ให้ค่าความถูกต้อง และค่า AUC ที่ดีขึ้นเมื่อเปรียบเทียบกับอัลกิรีทึม C4.5 และ C5.0 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) ซึ่งช่วยลดความผิดพลาดและ

เพิ่มความแม่นยำของโมเดล ในทางตรงกันข้ามอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 มีเพียงต้นไม้ต้นเดียวซึ่งมีความหลากหลายที่น้อยกว่า

การสุ่มข้อมูลเพิ่ม (oversampling) เป็นเทคนิคที่ใช้เพื่อจัดการกับปัญหาข้อมูลไม่สมดุล ซึ่งเป็นการเพิ่มจำนวนข้อมูลของคลาสที่มีข้อมูลน้อย เพื่อให้คลาสนั้นมีข้อมูลใกล้เคียงกับคลาสที่มากกว่าในชุดข้อมูล ซึ่งเพิ่มการเรียนรู้ของคลาสที่เป็นข้อจำกัด ในกรณีที่คลาสหนึ่งมีข้อมูลน้อย การสุ่มเพิ่มสามารถเพิ่มประสิทธิภาพของโมเดลในการจำแนกคลาสนั้น ๆ และลดปัญหาของการเรียนรู้ไม่สมดุล ช่วยให้มีการเรียนรู้ที่สมดุลกับคลาสที่น้อยกว่า โดยโมเดลที่ฝึกด้วยข้อมูลที่มีการสุ่มเพิ่มอาจแสดงผลลัพธ์ที่ดีกว่าในบางสถานการณ์ เมื่อเทียบกับการใช้ข้อมูลไม่สมดุลแบบเดิม ถึงแม้ว่าการสุ่มข้อมูลเพิ่ม (oversampling) เป็นวิธีการที่มีประสิทธิภาพในการจัดการกับปัญหาข้อมูลไม่สมดุล แต่ยังต้องพิจารณาข้อเสียและข้อจำกัดด้วย เนื่องจากว่าการสุ่มเพิ่มเป็นการเพิ่มข้อมูลที่เหมือนหรือคล้ายกัน อาจทำให้โมเดลเกิดการเรียนรู้ข้อมูลมากเกินไปบนข้อมูลเหล่านั้น และอาจทำให้โมเดลไม่สามารถทำนายผลลัพธ์ที่แม่นยำกับข้อมูลที่ไม่เคยเห็นมาก่อนได้ รวมถึงเพิ่มความซับซ้อนของการคำนวณ เพิ่มระยะเวลาในการฝึกสอนและการทดสอบของโมเดล และการสุ่มข้อมูลเพิ่มโดยการสร้างข้อมูลใหม่อาจมีโอกาสทำให้ข้อมูลที่ถูกสร้างขึ้นมาไม่สะท้อนความเป็นจริงของการกระจายของข้อมูลในชุดข้อมูลจริง

การลดข้อมูล (undersampling) เป็นเทคนิคที่ใช้ในการจัดการกับปัญหาข้อมูลไม่สมดุล โดยการลดจำนวนข้อมูลในคลาสที่มีข้อมูลมากเพื่อให้มีความสมดุลกับคลาสที่มีข้อมูลน้อยขึ้น โดยการลดข้อมูลช่วยลดความซับซ้อนของโมเดล เนื่องจากมีจำนวนข้อมูลน้อยลง ซึ่งสามารถลดเวลาที่ใช้ในการฝึกสอนและทดสอบโมเดลได้ และยังช่วยให้โมเดลมีโอกาสในการจำแนกคลาสที่มีข้อมูลน้อยแม่นยำขึ้น เนื่องจากมีจำนวนข้อมูลที่เพียงพอสำหรับการฝึกสอน อีกทั้งช่วยลดการใช้ทรัพยากรคอมพิวเตอร์ที่เกินจำเป็น โดยที่โมเดลที่ฝึกด้วยข้อมูลที่มีการลดข้อมูลมักมีความเสถียรมากกว่าในสถานการณ์ที่ข้อมูลไม่สมดุลเนื่องจากมีการเรียนรู้จำนวนคลาสที่น้อยลง แต่อย่างไรก็ตามการลดข้อมูลมีข้อจำกัด เมื่อมีการลดข้อมูล โมเดลที่ได้จากการฝึกอาจจะไม่เสถียรหรือมีประสิทธิภาพต่ำเมื่อนำไปใช้กับข้อมูลจริงที่ไม่ได้ถูกลดขนาด เพราะการลดข้อมูลอาจทำให้ข้อมูลบางส่วนหรือสารสนเทศที่สำคัญต่อการทำนายของโมเดลหายไป ส่งผลให้โมเดลไม่สามารถเรียนรู้แนวโน้มหรือความสัมพันธ์ทั้งหมดที่มีในข้อมูลจริงได้ รวมทั้งอาจไม่สามารถสะท้อนความเป็นจริงของข้อมูลในเซ็ตข้อมูลทั้งหมด ซึ่งส่งผลให้โมเดลที่ฝึกด้วยข้อมูลที่ถูกลดขนาดไม่สามารถทำนายผลลัพธ์ของข้อมูลจริงที่มีขนาดใหญ่ได้

จากผลการศึกษาที่ได้กล่าวมาทั้งหมด พบว่าวิธี Random forest ร่วมกับเทคนิคการสุ่มเพิ่ม (oversampling) 35% เป็นวิธีที่ดีที่สุดสำหรับชุดข้อมูลมะเร็งเต้านมที่ทำการศึกษา ซึ่งสอดคล้องกับงานวิจัยของศวิกร บันลือทรัพย์ และวราพร จิระพันธุ์ทอง (2565) โดยข้อมูลที่ทำการศึกษาเป็นข้อมูลทางด้านการแพทย์เหมือนกัน ได้ทำการศึกษาอัลกอริทึมการเรียนรู้ของเครื่องสำหรับการทำนายการคัดแยกผู้ป่วย COVID-19 ได้ศึกษาข้อมูลผู้ป่วย จำนวน 1,608,923 ราย จากกรมควบคุมโรค โดยใช้

อัลกอริทึมการจำแนกข้อมูลทั้งหมด 3 แบบได้แก่ Random forest, Neural network และ Naive bayes พบว่าวิธี Random forest มีค่าความถูกต้องเท่ากับ 93.33% ซึ่งเป็นอัลกอริทึมที่ดีที่สุดจากทั้ง 3 แบบ โดยสามารถสรุปได้ว่าวิธี Random forest คือวิธีที่แม่นยำและมีประสิทธิภาพที่สุดสำหรับข้อมูลทางการแพทย์ในทั้ง 2 กรณี เนื่องจากข้อมูลทางการแพทย์มักมีความซับซ้อนและมีหลายมิติทำให้วิธี Random forest ซึ่งเป็นการตัดสินใจจากต้นไม้หลาย ๆ ต้น สามารถจัดการความซับซ้อนนี้ได้ดี จากเหตุผลที่กล่าวมาวิธี Random forest เป็นวิธีที่เหมาะสมและมีประสิทธิภาพสำหรับการจำแนกและทำนายข้อมูลทางการแพทย์ ไม่ว่าจะเป็นการคัดแยกผู้ป่วย COVID-19 หรือผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านมที่ทำการศึกษา อีกทั้งยังสอดคล้องกับงานวิจัยของวิชญ์วิสิฐ เกสรสิทธิ์ และวิชิต หล่อจีระชุณห์กุล (2561) ได้ใช้วิธีการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน โดยได้ใช้วิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวานจำนวน 4 วิธี คือ วิธีสุ่มเกิน (oversampling), วิธีสุ่มลด (undersampling), วิธีผสมผสาน (hybrid method) และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยอัลกอริทึมต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด ซึ่งกล่าวได้ว่าวิธีการสุ่มข้อมูลเป็นวิธีที่ช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูล โดยโมเดลจะมีข้อมูลมากขึ้นสำหรับการเรียนรู้และการจำแนกคลาสที่น้อยกว่า และเพิ่มความถูกต้องในการจำแนกคลาสที่น้อยกว่าได้อีกด้วย จากนั้นผู้วิจัยได้นำวิธี Random forest ร่วมกับเทคนิคการสุ่มเพิ่ม (oversampling) 35% ในการแบ่งสัดส่วนข้อมูล 80:20 ที่ให้ค่าความถูกต้อง และค่า AUC มากที่สุด มาทำการหาจำนวนต้นไม้, ความลึกของต้นไม้ และ K-fold cross-validation เพื่อดูค่าที่เปลี่ยนไปในแต่ละรอบ ผลการศึกษาแสดงให้เห็นว่าจำนวนต้นไม้ที่ดีที่สุดคือ 150 ความลึกของต้นไม้คือ 15 และ K-Fold ที่ดีที่สุดคือ 7 หลังจากค้นพบค่าพารามิเตอร์ที่ดีที่สุดทำการฝึกและทดสอบโมเดล Random forest ด้วยการใช้ค่าพารามิเตอร์เหล่านี้ และประเมินประสิทธิภาพของโมเดลโดยใช้ข้อมูลที่ไม่เคยเห็นมาก่อนโดยเพื่อให้แน่ใจว่าโมเดลทำงานได้ดีที่สุด พบว่าโมเดลให้ค่าความถูกต้องในการทำนายคลาสต่าง ๆ บนชุดข้อมูลทดสอบที่ไม่เคยเห็นมาก่อนเท่ากับ 70.49% และค่า AUC เท่ากับ 0.70 ถือว่าอยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ ซึ่งปัจจัยที่ส่งผลในการจำแนกโรคมะเร็งเต้านม 5 อันดับแรก ได้แก่ ดัชนีมวลกาย, กลุ่มอายุ, อาการที่นำมาพบแพทย์, ความสมมาตรของเต้านมสองข้าง และก้อนหรือถุงน้ำ ตามลำดับ ซึ่งสอดคล้องกับงานวิจัยของภรณี เหล่าอิทธิ และนภา ปริญญานิติกูล (2559) ศึกษาเกี่ยวกับมะเร็งเต้านม พบว่าผู้หญิงที่อ้วนหรือมีภาวะน้ำหนักเกินมาตรฐาน (ดัชนีมวลกายเกินมาตราฐาน) โดยเฉพาะในผู้หญิงวัยหลังหมดประจำเดือนแล้ว, การสูบบุหรี่ และดื่มเครื่องดื่มที่มีแอลกอฮอล์ก็เพิ่มโอกาสความเสี่ยงของการเกิดโรคมะเร็งเต้านม รวมไปถึงอายุที่มากขึ้นทำให้มีความเสี่ยงในการเป็นมะเร็งเต้านมได้เช่นเดียวกัน รวมไปถึงงานวิจัยของอุมาพร นันทธีโร (2555) ศึกษาเกี่ยวกับอาการนำมาพบแพทย์ของผู้มาตรวจมะเร็งเต้านม โดยพบว่าอาการนำมาพบแพทย์ เช่น

คลำเจอก้อน, เจ็บเต้านม มีความสัมพันธ์กับขนาดของก้อนเนื้อ (Mass) เพราะเนื่องจากการเกิดของโรคมะเร็งเต้านมของผู้ป่วยมักไม่มีอาการผิดปกติในระยะแรก จึงมีความจำเป็นและสำคัญอย่างยิ่งที่ต้องทำการตรวจค้นหามะเร็งเต้านมในระยะเริ่มต้นของผู้มาตรวจมะเร็งเต้านมทั้งในกลุ่มที่มีอาการและไม่มีอาการ โดยพบผู้ที่มีความเสี่ยงสูงในการเป็นโรคมะเร็งเต้านม (BIRADS 4-6) และมีอาการนำมาพบแพทย์ คิดเป็นร้อยละ 96.3

ดังนั้น สำหรับการแก้ปัญหาข้อมูลไม่สมดุลในงานวิจัยนี้เป็นอีกหนึ่งแง่ที่สำคัญที่ช่วยให้โมเดลทำนายได้แม่นยำและเสถียร ซึ่งเป็นประโยชน์ในการวิเคราะห์โรคและการตรวจสอบความเสี่ยงในโดเมนทางการแพทย์ ซึ่งเป็นงานที่ความถูกต้องและน่าเชื่อถือมีความสำคัญเป็นอย่างยิ่ง การจัดการกับปัญหาความไม่สมดุลข้อมูลและการแบ่งข้อมูลออกเป็นชุดฝึกและชุดทดสอบเป็นปัจจัยสำคัญในการทำนาย การสุ่มข้อมูลเพิ่มและลดข้อมูลอาจช่วยปรับปรุงความแม่นยำของโมเดล แต่อาจมีผลให้ข้อมูลเสียหายและเพิ่มเวลาในการประมวลผล ควรพิจารณาความสมดุลของข้อมูลและการแบ่งข้อมูลให้อยู่ในเกณฑ์ที่เหมาะสมในแต่ละกรณีการวิเคราะห์ข้อมูลนั้น ๆ ซึ่งเป็นสิ่งสำคัญในการสร้างโมเดลที่มีประสิทธิภาพในการจำแนกข้อมูลทางการแพทย์ ฉะนั้นการเลือกอัลกอริทึมที่เหมาะสมจะขึ้นอยู่กับความต้องการของงานและลักษณะของข้อมูล และอาจจะต้องพิจารณาเพิ่มเติมเกี่ยวกับความเหมาะสมของข้อมูลและอัลกอริทึมในงานที่มีลักษณะแตกต่างกัน ซึ่งประสิทธิภาพของโมเดลยังขึ้นอยู่กับขนาดข้อมูลทั้งหมดที่ทำการศึกษา ถ้ามีข้อมูลมากพอและสามารถแบ่งเป็นสัดส่วนใหญ่ ๆ สำหรับการทดสอบและสำหรับการฝึกก็อาจทำให้โมเดลมีประสิทธิภาพดีขึ้น เนื่องจากโมเดลมีโอกาสเรียนรู้จากข้อมูลมากขึ้น ซึ่งอาจช่วยให้โมเดลสามารถจับความสัมพันธ์ในข้อมูลได้ดีขึ้น รวมถึงช่วยให้การวัดประสิทธิภาพของโมเดลมีความเสถียรมากขึ้น เนื่องจากมีข้อมูลมากที่ใช้ในการฝึกสอน แต่การมีข้อมูลมากไม่ได้หมายความว่าเป็นการแก้ปัญหาข้อมูลไม่สมดุลจำเป็นต้องใช้วิธีการสุ่มข้อมูลเพิ่มหรือลดข้อมูลในบางกรณีที่คลาสมีข้อมูลน้อยมาก การสุ่มข้อมูลเพิ่มหรือลดข้อมูลอาจยังจำเป็นอยู่เพื่อให้โมเดลมีความสมดุลและสามารถจำแนกคลาสนั้นได้อย่างแม่นยำมากขึ้น

จากผลการศึกษา วิธี Random forest แสดงประสิทธิภาพที่ดีกว่าอัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ C5.0 ในหลายด้าน โดยเฉพาะอย่างยิ่งหลังจากการปรับปรุงข้อมูลไม่สมดุลด้วยวิธีสุ่มข้อมูลฝึกหัดเพิ่ม 35% ช่วยให้ข้อมูลมีความสมดุลมากขึ้น การปรับปรุงนี้ช่วยให้วิธี Random forest สามารถแยกแยะคลาสที่มีจำนวนน้อยลงได้ดีขึ้น และทำให้ค่าความถูกต้องและค่า AUC สูงขึ้น เมื่อเทียบกับอัลกอริทึมต้นไม้ตัดสินใจที่ยังคงแสดงผลลัพธ์ไม่แตกต่างจากเดิมมากนัก แสดงให้เห็นว่าวิธี Random forest มีความสามารถปรับตัวได้ดีกับข้อมูลที่มีความซับซ้อนและไม่สมดุล รวมทั้งมีประสิทธิภาพดีกว่าอัลกอริทึมต้นไม้ตัดสินใจซึ่งสอดคล้องกับสมมติฐาน อีกทั้งผลการวิเคราะห์ข้อมูลแสดงให้เห็นว่าดัชนีมวลกายและอายุเป็นปัจจัยสำคัญที่มีผลต่อการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม โดยผลการวิจัยแสดงให้เห็นว่าดัชนีมวลกายเป็นปัจจัยที่มีค่า Feature Importance

สูงที่สุดในการจำแนกความเสี่ยงของโรคมะเร็งเต้านม และพบว่าอายุโดยเฉพาะกลุ่มอายุ 33-45 ปี เป็นอีกหนึ่งปัจจัยที่สามารถจำแนกความเสี่ยงของโรคมะเร็งเต้านมได้ดีที่สุด ซึ่งสอดคล้องกับสมมติฐานที่ว่า ดัชนีมวลกายและอายุเป็นปัจจัยสำคัญที่มีผลต่อการจำแนกผู้ที่มีความเสี่ยงในการเป็นโรคมะเร็งเต้านม

5.3 ข้อเสนอแนะ

5.3.1 ข้อเสนอแนะจากการวิจัย

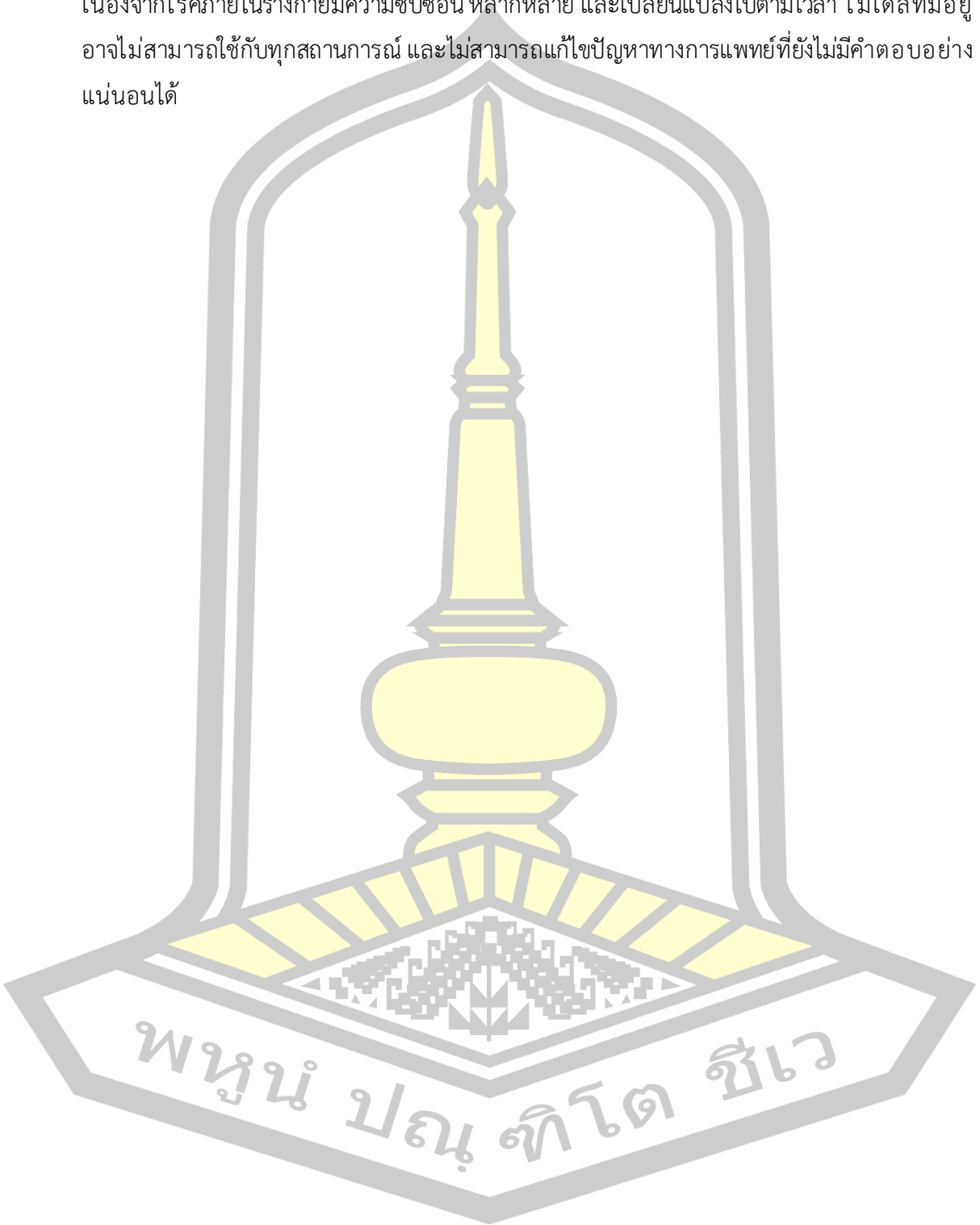
จากปัญหาข้อมูลไม่สมดุล การแก้ปัญหาข้อมูลไม่สมดุลเป็นขั้นตอนสำคัญในการปรับปรุงคุณภาพของข้อมูลและการวิจัยข้อมูลทางการแพทย์ เนื่องจากการตรวจจับโรคหรือความเสี่ยงของโรคที่มีอัตราการเกิดต่อประชากรต่ำ ทำให้คลาสของผู้ป่วยหรือผู้ที่มีความเสี่ยงสูงมีจำนวนน้อยเมื่อเปรียบเทียบกับคลาสของผู้ที่ไม่มีโรคหรือมีความเสี่ยงต่ำ การใช้เทคนิคที่เหมาะสมเพื่อปรับปรุงความแม่นยำของโมเดลและการจัดการข้อมูลอย่างมีประสิทธิภาพเป็นสิ่งสำคัญ เพื่อให้การวิจัยทางการแพทย์มีผลสัมฤทธิ์ที่ดีที่สุด สามารถทำนายโรคอย่างแม่นยำและมีประสิทธิภาพมากขึ้นในอนาคต

5.3.2 ข้อเสนอแนะเพื่อการวิจัยครั้งต่อไป

1. ข้อมูลที่นำมาทำการศึกษาเป็นข้อมูลจากโรงพยาบาลสุทธาเวช คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคามซึ่งเป็นข้อมูลจากโรงพยาบาลเดียว การใช้ข้อมูลจากโรงพยาบาลเดียวอาจทำให้ข้อมูลไม่สามารถแทนความหลากหลายของประชากรโรคมะเร็งเต้านมในประเทศไทยได้แบบสมบูรณ์ เนื่องจากอาจมีรูปแบบการดูแลรักษาที่แตกต่างกัน หากผู้วิจัยท่านอื่นสนใจศึกษาเกี่ยวกับโรคมะเร็งเต้านมควรพิจารณาเพิ่มข้อมูลจากแหล่งข้อมูลที่หลากหลาย เพื่อให้ข้อมูลที่ใช้ในการศึกษาเป็นตัวแทนที่ดีที่สุดของประชากรโรคมะเร็งเต้านมในประเทศไทย ทำให้การวิจัยเป็นไปอย่างถูกต้องและเชื่อถือได้มากยิ่งขึ้นสำหรับการพยากรณ์โรคมะเร็งเต้านมและปัจจัยเสี่ยงที่เกี่ยวข้อง และจะสามารถนำข้อมูลไปใช้ในการวางแผนรับมือและปรับปรุงการดูแลรักษาให้มีประสิทธิภาพมากขึ้นได้ในอนาคต

2. การเพิ่มตัวแปรที่เกี่ยวข้องมากขึ้น อาจช่วยให้โมเดลมีประสิทธิภาพที่ดีขึ้น โดยเฉพาะในกรณีที่ตัวแปรมีอิทธิพลต่อความเสี่ยงหรือพฤติกรรมของผู้ป่วยมะเร็งเต้านม ตัวแปรเหล่านี้อาจมีความสัมพันธ์กับเป้าหมายของงานและช่วยให้โมเดลมีความแม่นยำมากขึ้น เช่น ความเครียด, ประวัติคนในครอบครัว หรือสถานะสมรส เป็นต้น ซึ่งข้อมูลเหล่านี้อาจจะต้องวางแผนในการเก็บรวบรวมข้อมูลเป็นระยะเวลานาน เนื่องจากในปัจจุบันข้อมูลในส่วนนี้เข้าถึงได้ยากและยังไม่ได้มีการเผยแพร่ให้ทุกคนสามารถเข้าถึงข้อมูลได้

3. การค้นคว้าและพัฒนาโมเดลใหม่ ๆ ในด้านการแพทย์เป็นกระบวนการที่สำคัญและจำเป็น เนื่องจากโรคภายในร่างกายมีความซับซ้อน หลากหลาย และเปลี่ยนแปลงไปตามเวลา โมเดลที่มีอยู่อาจไม่สามารถใช้กับทุกสถานการณ์ และไม่สามารถแก้ไขปัญหาทางการแพทย์ที่ยังไม่มีคำตอบอย่างแน่นอนได้



บรรณานุกรม

บรรณานุกรม

Berrar, D., Granzow, M., Dubitzky, W., Stilgenbauer, S., Wilgenbus, K. D. H., Lichter, P., & Eils, R. (2001). New insights in clinical impact of molecular genetic data by knowledge-driven data mining. *In Proceedings of the Second International Conference on Systems Biology* (pp. 275-281). 148.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation.* Prentice-Hall, Inc..

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. *SPSS inc*, 9(13), 1-73.

Fentiman, I. (2009). Male breast cancer: a review. *Ecancermedicalscience*. 2009; 3: 140. External Resources Pubmed/Medline (NLM) Crossref (DOI).

Frey, L., Fisher, D., Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003, November). Identifying Markov blankets with decision tree induction. *In Third IEEE International Conference on Data Mining* (pp. 59-66). IEEE.

Govindarajan, M. (2007). Text mining technique for data mining application. *International Journal of Computer and Information Engineering*, 1(11), 3473-3478.

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. *In Feature extraction: Foundations and applications* (pp. 137-165). Berlin, Heidelberg: Springer Berlin Heidelberg.

Marqués, A. I., García, V., & Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7), 1060-1070. https://doi.org/10.1057/jors.2012.120

Ménard, C., Johann, D., Lowenthal, M., Muanza, T., Sproull, M., Ross, S., ... & Camphausen, K. (2006). Discovering clinical biomarkers of ionizing radiation exposure with serum proteomic analysis. *Cancer Research*, 66(3), 1844-1850.

Ministry of Public Health. (2005). Report on health behavior assessment: *selfexamination of womenaged 35 years and over*. P. A. C. o. Thailand.

Mosayebi, A., Mojaradi, B., Bonyadi Naeini, A., & Khodadad Hosseini, S. H. (2020). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PloS one*, 15(10), e0237658.

Nwokocha, N., Kabari, L., & Agaba, F. (2019). Prediction of Breast Cancer Disease Using Decision Tree Algorithm. *International Journal of Innovative Information Systems & Technology Research*, 7(1), 34-38.

Rajeswari, S., & Suthendran, K. (2019). C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*, 156, 530-539.

Ray, R., Abdullah, A. A., Mallick, D. K., & Dash, S. R. (2019, November). Classification of benign and malignant breast cancer using supervised machine learning algorithms based on image and numeric datasets. *In Journal of Physics: Conference Series* (Vol. 1372, No. 1, p. 012062). IOP Publishing.

Saabith, A. L. S., Sundararajan, E., & Bakar, A. A. (2014). Comparative study on different classification techniques for breast cancer dataset. *Int J Comput Sc Mob Comput*, 3(10), 185-191.

Saad, H., & Nagarur, N. (2020). Data mining techniques in predicting breast cancer. *Applied Sci.*, 20(4). https://doi.org/https://doi.org/10.3923/jas.2020.124.133

Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J., & Dogan, B. E. (2017). BI-RADS® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3), 179-190.

Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of breast cancer using decision tree data mining technique. *International Journal of Computer Applications*, 98(10), 16-24.

Venkatesan, E. V., & Velmurugan, T. (2015). Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*, 8(29), 1-8.

Yi, M. (2019). A complete guide to scatter plots. *Retrieved February*, 25, 2021.

คณะแพทยศาสตร์ ศิริราชพยาบาล. มะเร็งเต้านมเรื่องที่น่ารู้และข้อควรปฏิบัติสำหรับผู้ป่วยมะเร็งเต้านม. กรุงเทพฯ: ภาควิชาศัลยศาสตร์; 2555.

จิณพิชญ์ชา มะมม. (2555). ความก้าวหน้าในการดูแลรักษาผู้ป่วยมะเร็งเต้านม. *วารสารสภาการพยาบาล*, 23(2), 11-25.

จิตตินัดด์ หวานนท์. (2533). ประวัติโรคมะเร็งเต้านม. *วารสารวิจัยวิทยาศาสตร์การแพทย์*, 4(2), 145-148.

ชัดชัย แก้วตา และอัจฉรา มหาวีรวัฒน์. (2553). การวินิจฉัยคดีด้วยเทคนิคต้นไม้ตัดสินใจ. *The 3rd*

ณัฏฐพร นันทิวัฒนา. (2563, 23 กุมภาพันธ์). *มะเร็งเต้านม มะเร็งอันดับ 1 ของผู้หญิง.* โรงพยาบาลศิครินทร์. https://www.sikarin.com/doctor-articles/โรคมะเร็งเต้านม-มะเร็ง

ณัฏฐพร นันทิวัฒนา. (2563, 5 สิงหาคม). *5 อันดับมะเร็งที่พบบ่อยในคนไทย.* โรงพยาบาลศิครินทร์. https://www.sikarin.com/health/มะเร็ง-อาการโรคมะเร็ง

ณัฐวุฒิ ศรีวิบูลย์. (2559). การเปรียบเทียบประสิทธิภาพอัลกอริธึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง. *Creative Science*, 8(3), 344-352.

ทิพพารัตน์ แสนคาร และธนัช กนกเทศ. (2019). ความรู้เกี่ยวกับมะเร็งเต้านมการป้องกันและวิธีการตรวจเต้านมด้วยตนเอง. *วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย ฉบับวิทยาศาสตร์และเทคโนโลยี*, 13(3), 14-21.

ปัทมาพร พึ่งบุญ. (2555). มะเร็งเต้านมในระหว่างตั้งครรภ์. *วารสารสภาการพยาบาล*, 21(4), 32-42.

พิมพ์จุฑา วราทรัพย์, จิราวัลย์ จิตรถเวช และวิชิต หล่อจีระชุณห์กุล. (2561). ปัจจัยเสี่ยงที่มีความสัมพันธ์ต่อการเกิดโรคมะเร็ง. *วารสารวิทยาศาสตร์ลาดกระบัง*, 27(2), 1-14.

ภรณี เหล่าอิทธิ และนภา ปริญญานิติกูล. (2559). มะเร็งเต้านม: ระบาดวิทยา การป้องกัน และแนวทางการตรวจคัดกรอง. *จุฬาลงกรณ์เวชสาร*, 60(5), 497-507.

ภรัณยา ปาลวิสุทธิ์. (2559). การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อย. *วารสารเทคโนโลยีสารสนเทศ*, 12(1), 54-63.

มุทิตา หวังคิด, อาริกา ธรรมโน และอาริต ธรรมโน. (2563). การพยากรณ์โรคมะเร็งเต้านมด้วยอัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง. *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, 10(2), 1-9.

วิชญ์วิสิฐ เกสรสิทธิ์ และวิชิต หล่อจีระชุณห์กุล. (2561). การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน. วารสารวิจัย มหาวิทยาลัยขอนแก่น (ฉบับบัณฑิตศึกษา), 18(3), 11-21.

ศิวกร บรรลือทรัพย์ และ วราพร จิระพันธุ์ทอง. (2022). อัลกอริธึมการเรียนรู้ของเครื่องสำหรับการทำนายการคัดแยกผู้ป่วย COVID-19. *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, 12(1).

สถาบันมะเร็งแห่งชาติ กรมการแพทย์ กระทรวงสาธารณสุข. (2563). *มะเร็งคืออะไร. https://www.nci.go.th/th/New_web/index.html*

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

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

หะสัน มูหาหมัด. (2555, 20 กุมภาพันธ์). *ระยะของมะเร็งเต้านม*. ThaiBreastCancer.com. http://www.thaibreastcancer.com/ca-131/

อัจจิมา มณฑาพันธุ์. (2562). การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม. *วารสารแพทยสารทหารอากาศ*, 65(2), 49-56.

อุกฤษฎ์ ศรีสุข และจารี ทองคำ. (2564). การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับอุบัติการณ์ของผู้ป่วย. *วารสารวิทยาศาสตร์และเทคโนโลยีมหาวิทยาลัยมหาสารคาม*, 40(2), 157-163.

อุปมา เลี้ยงสว่างวงศ์. (2541). โรคมะเร็ง. *วารสารวิจัยวิทยาศาสตร์การแพทย์*, 12(1), 43-54.

อุมาพร นันทธีโร. (2555). อาการนำมาพบแพทย์ของผู้มาตรวจมะเร็งเต้านม. *วารสารการแพทย์โรงพยาบาลศรีสะเกษ สุรินทร์ บุรีรัมย์*, 27(3), 251-262.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น*.(2). หสม. ดาต้า คิวบ์.

ประวัติผู้เขียน

ชื่อ	นายชัยยันต์ สุขหมั่น
วันเกิด	วันที่ 2 กันยายน 2540
สถานที่เกิด	อำเภอสนม จังหวัดสุรินทร์
สถานที่อยู่ปัจจุบัน	เดอะแกรนด์ เรวดี ซอย 2 101 ถ. ติวานนท์ ตำบลตลาดขวัญ อำเภอเมืองนนทบุรี นนทบุรี 11000
ตำแหน่งหน้าที่การงาน	นักเทคโนโลยีสารสนเทศ
สถานที่ทำงานปัจจุบัน	กลุ่มพัฒนาระบบข่าวกรองและเฝ้าระวังโรคไม่ติดต่อ กองระบาดวิทยา กรมควบคุมโรค กระทรวงสาธารณสุข ,88/21 ถนน ติวานนนท์ ตำบลตลาดขวัญ อำเภอเมือง จังหวัดนนทบุรี 11000
ประวัติการศึกษา	พ.ศ. 2563 วิทยาศาสตรบัณฑิต (วท.บ.) สาขา คณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม พ.ศ. 2566 วิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขา วิทยาการจัดการสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม