การเปรียบเทียบประสิทธิภาพของตัวแบบการเรียนรู้ของเครื่องสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด

Comparing the Performance of Machine Learning Models for Classifying Lung Cancer Patients

ศวิตา ทองขุนวงศ์[1], ภัคพล สวัสกมล[2*]

Sawita Tongkunwong[1] , Phakkhaphon Sawatkamon[2*]

[1]โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี มหาวิทยาลัยเทคโนโลยีสุรนารี

[2]สาขาวิชาคณิตศาสตร์และภูมิสารสนเทศ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

[1] Suranaree University of Technology Hospital, Suranaree University of Technology

[2]School of Mathematical Science and Geoinformatics, Institute of Science, Suranaree University of Technology

[*] Corresponding Author: E-mail address: Sawiphak2424@gmail.com

บทคัดย่อ

จุดมุ่งหมายของงานวิจัยในครั้งนี้คือ การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอดด้วยการเรียนรู้ของเครื่อง ประกอบด้วย 5 ตัวแบบ ได้แก่ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรสเนเบอร์ และ ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน ในการศึกษาครั้งนี้ได้ใช้ข้อมูลทุติยภูมิจากเว็บไซต์ Kaggle.com จำนวน 309 ข้อมูล และใช้ค่าความแม่นยำเป็นตัวเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด ผลการศึกษาพบว่า ตัวแบบที่มีประสิทธิภาพสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด คือ ตัวแบบป่าสุ่ม มีค่าความแม่นยำมากสุด คือ 90.32% ส่วนตัวแบบต้นไม้ตัดสินใจ ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรสเนเบอร์ และตัวแบบซัพพอร์ตเวกเตอร์แมชชีน มีค่าความแม่นยำในการจำแนกผู้ป่วยโรคมะเร็งปอด เท่ากับ 87.10%, 83.87%, 80.65% และ 87.10% ตามลำดับ

คำสำคัญ: โรคมะเร็งปอด, การเรียนรู้ของเครื่อง, ป่าสุ่ม

ABSTRACT

The objective of this research is to compare the performance of machine learning models for lung cancer patient classification. The study includes five different models: decision tree, random forest, naive Bayes, k-nearest neighbors, and support vector machine. The research utilizes secondary data from the Kaggle.com website, consisting of 309 records, and uses accuracy as the metric to compare the performance of these models. The study result showed that the most efficient model for lung cancer patient classification is the random forest model,

Received	25-02-2024
Revised	30-05-2024
Accepted	28-06-2024

achieving the highest accuracy of 90.32%. Meanwhile, the decision tree, naive Bayes, k-nearest neighbors, and support vector machine models demonstrated classification accuracies of 87.10%, 83.87%, 80.65%, and 87.10%, respectively, in identifying lung cancer patients.

Keywords: lung cancer, machine learning, random forest

1. บทนำ

โรคมะเร็งปอดเป็นหนึ่งปัญหาสุขภาพที่มีความรุนแรงและส่งผลกระทบต่อประชากรในประเทศไทย โดยกรมการแพทย์สถาบันมะเร็งแห่งชาติ ให้ข้อมูลว่าโรคมะเร็งปอดเป็นสาเหตุการเสียชีวิตอันดับที่ 2 ของคนไทย และมีอัตราการเสียชีวิตในผู้หญิงสูงกว่าผู้ชาย [1] มะเร็งปอดเกิดจากเซลล์ที่อยู่ในปอดมีการเจริญเติบโตที่ไม่ปกติ เนื่องมาจากการเปลี่ยนแปลงทางพันธุกรรมของเซลล์ปกติ และไม่สามารถควบคุมการเจริญเติบโตนี้ได้ จึงทำให้เกิดก้อนเนื้อในปอด โดยปัจจัยที่ทำให้เกิดจากเปลี่ยนแปลงทางพันธุกรรมของเซลล์นั้นเกิดจากหลายปัจจัย เช่น การสูบบุหรี่ การสูดดมควันบุหรี่ การสูดดมมลพิษทางอากาศ การมีประวัติคนในครอบครัวเคยเป็นโรคมะเร็งปอด (ปัจจัยทางพันธุกรรม)

การตรวจพบโรคมะเร็งปอดในระยะเริ่มต้นของผู้ป่วยทำได้ยาก เนื่องจากผู้ป่วยมักไม่มีอาการแสดงที่บ่งบอกถึงความผิดปกติของร่างกาย แต่ผู้ป่วยมักจะพบความผิดปกติของร่างกายในระยะลุกลามของโรคมะเร็งปอด เช่น ไอเรื้อรัง ไอมีเสมหะ หายใจมีเสียงหวีด เจ็บหน้าอก เหนื่อยง่าย เป็นต้น โรคมะเร็งปอดจะแบ่งออกเป็น 2 ชนิด คือ มะเร็งปอดเซลล์เล็ก และ มะเร็งปอดที่ไม่ใช่เซลล์เล็ก โดยมะเร็งปอดเซลล์เล็ก เป็นเซลล์มะเร็งที่เจริญเติบโตและแพร่กระจายในร่างกายได้อย่างรวดเร็ว จึงทำให้มีอัตราการเสียชีวิตที่สูง วิธีการการรักษานั้นมักจะใช้วิธีฉายรังสี ส่วนมะเร็งปอดที่ไม่ใช่เซลล์เล็ก เป็นเซลล์มะเร็งที่เจริญเติบโตและแพร่กระจายในร่างกายได้ช้ากว่า หากทราบหรือตรวจพบในระยะเริ่มต้นของโรคจะมีโอกาสรักษาให้หายได้ [2]

การทราบถึงระยะของโรคมะเร็งปอดนั้นมีความสำคัญต่อการรักษาผู้ป่วยอย่างมาก เนื่องจากถ้าเรามีเครื่องมือในการคัดกรองผู้ป่วยโรคมะเร็งปอดในระยะเริ่มต้น จะทำให้ผู้ป่วยนั้นไม่เสียโอกาสในการรักษาโรคมะเร็งปอดได้ ดังนั้น ในการวิจัยนี้จึงได้ใช้ความรู้ทางด้านคณิตศาสตร์เข้ามาช่วยสร้างเครื่องมือในการสร้างตัวแบบจำแนกผู้ป่วยโรคมะเร็งปอด และทำการเปรียบเทียบประสิทธิภาพการทำนายของตัวแบบเพื่อหาตัวแบบในการทำนายที่ดีที่สุดเพื่อใช้ในการจำแนกผู้ป่วยโรคมะเร็งปอดโดยทำการเปรียบตัวแบบทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบป่าสุ่ม, ตัวแบบต้นไม้ตัดสินใจ, ตัวแบบนาอีฟเบย์, ตัวแบบเคเนียร์เรสเนเบอร์ และ ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน

การเรียนรู้ของเครื่องเป็นส่วนหนึ่งของปัญญาประดิษฐ์ เป็นเทคโนโลยีที่ทันสมัยที่จะสามารถเข้ามาช่วยสนับสนุนด้านการแพทย์ ซึ่งตัวแบบทั้ง 5 ตัวแบบ จัดเป็นการเรียนรู้แบบมีผู้สอน คือ ตัวแบบจะเน้นการเรียนรู้จากกลุ่มข้อมูลชุดตัวอย่างเพื่อให้ได้ตัวแบบทำนายมีความแม่นยำในการจำแนกผู้ป่วยโรคมะเร็งปอด นอกจากนี้ได้มีการนำความรู้การเรียนรู้ของเครื่องไปประยุกต์ในงานวิจัยต่าง ๆ เช่น การพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล เป็นงานวิจัยที่ได้มีการเปรียบเทียบประสิทธิภาพแบบจำลองสำหรับการพยากรณ์ผู้ป่วยโรคเบาหวาน ทั้งหมด 5 แบบจำลอง ผลการวิจัย

พบว่า เทคนิคป่าสุ่ม (random forest) ให้ค่าความถูกต้องในการทำนายผลเป็นโรคเบาหวานมากที่สุดที่ร้อยละ 99.75 [5] สำหรับงานวิจัยเปรียบเทียบประสิทธิภาพของตัวแบบการเรียนรู้ของเครื่องสำหรับการทำนายราคาพริกขี้หนูในจังหวัดนครศรีธรรมราช ตัวแบบป่าสุ่มเป็นตัวแบบที่เหมาะสมสำหรับนำไปพัฒนาตัวแบบสำหรับทำนายราคาพริกขี้หนู ในพื้นที่ของจังหวัดนครศรีธรรมราช ผลจากการทำนายพบว่าราคาทำนายมีราคาใกล้เคียงกับความจริง [6] สำหรับงานวิจัยวิธีการเรียนรู้ของเครื่องเพื่อทำนายการรับผลงานวิจัยทางวิชาการ พบว่า ตัวแบบป่าสุ่มมีค่าความแม่นยำในการทำนายที่ร้อยละ 81 [7]

1.1 ทฤษฎีที่เกี่ยวข้อง

1.1.1. ตัวแบบต้นไม้ตัดสินใจ (decision tree) เป็นอัลกอริทึมที่ใช้ในการจำแนกข้อมูล โดยมีลักษณะเลียนแบบการตัดสินใจของมนุษย์ โดยที่ผลลัพธ์ที่ได้จะมีสองกลุ่มหรืออาจมากกว่าสองกลุ่ม ตัวแบบต้นไม้ตัดสินใจจะเริ่มจาก โหนดราก (root node) และทำการแบ่งข้อมูลออกเป็นส่วนย่อย ๆ ด้วยการตัดสินใจแต่ละโหนด โดยมีการใช้เงื่อนไขเพื่อทำการแบ่งข้อมูลออกเป็นสองกลุ่มและทำการกระจายข้อมูลไปยังโหนดย่อย (child node) สองโหนดนั้นจะมีเงื่อนไขต่อไปอีกและกระจายข้อมูลต่อไปในโหนดย่อยของมันเรื่อย ๆ จนกระจายข้อมูลไปถึงโหนดใบ (leaf node) ที่ไม่มีการแบ่งข้อมูลต่อไปอีกแล้ว ซึ่งเป็นการให้ผลลัพธ์ของกระบวนการตัดสินใจ การสร้างตัวแบบจะเริ่มจากโหนดรากเป็นอันดับแรก เริ่มจากการคำนวณหาข้อมูลที่เหมาะสมที่จะเป็นโหนดราก ซึ่งจะพิจารณาจากค่าความได้เปรียบในการแยกข้อมูล (information gain) ที่มากที่สุด ซึ่งได้มาจากการคำนวณค่าเอนโทรปี (entropy) เพื่อให้การจำแนกของข้อมูลอยู่ในกลุ่มเดียวกันมากที่สุด เมื่อเราได้โหนดรากแล้ว ก็จะสร้างโหนดย่อยในลำดับต่อไปจนกระทั่งถึงโหนดใบ และได้ต้นไม้ตัดสินใจที่สมบูรณ์ ซึ่งค่าเอนโทรปี $\left(E(S)\right)$ และ ค่าความได้เปรียบในการแยกข้อมูล $\left(I(S,a)\right)$ สามารถคำนวณได้ดังสมการที่ (1) และ (2)

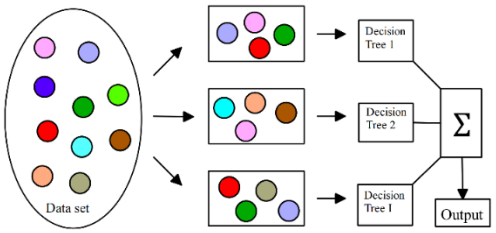
$$E(S) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \quad (1)$$

โดยที่ $P(x_i)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ x_i และมีทั้งหมด n เหตุการณ์

$$I(S,a) = E(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} E(S_v) \quad (2)$$

โดยที่ v คือ ค่าของลักษณะประจำ $\left(Values(A)\right)$,
$S_v = \{s | s \in S, Value(s,a) = v\}$ และ
$|S_v|$ คือ จำนวนข้อมูลในแต่ละกลุ่มย่อย S_v
$|S|$ คือ จำนวนข้อมูลทั้งหมด

1.1.2. ตัวแบบป่าสุ่ม (random forest) เป็นอัลกอริทึมที่ใช้ในการทำนายในรูปแบบของการจำแนกข้อมูล และการพยากรณ์ โดยหลักการทำงานของป่าสุ่มจะเป็นการสุ่มเอาข้อมูลซึ่งเป็นการสุ่มข้อมูลแบบใส่กลับคืน จากนั้นนำข้อมูลที่สุ่มได้ไปสร้างเป็นต้นไม้ตัดสินใจหลาย ๆ ต้น ซึ่งต้นไม้แต่ละต้นจะได้รับข้อมูล (data set) ที่แตกต่างกัน เมื่อทำการทำนายก็ให้ต้นไม้แต่ละต้นทำการตัดสินใจ โดยผลลัพธ์ที่ถูกโหวตมากสุดจะถูกเลือกเป็นผลลัพธ์ที่ทำนายออกมา



ภาพที่ 1 โครงสร้างการทำงานของป่าสุ่ม

1.1.3. ตัวแบบนาอีฟเบย์ (Naive Bayes) เป็นตัวแบบการจำแนกข้อมูลที่อาศัยหลักการความน่าจะเป็น (probability) ตามทฤษฎีบทของเบย์ (bayes theorem) มาใช้ในการจำแนกข้อมูล (classification) ซึ่งหลักการของนาอีฟเบย์จะใช้การคำนวณหาความน่าจะเป็นในการทำนายผลลัพธ์โดยจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ซึ่งมีความเหมาะสมกับกรณีที่เซตตัวอย่างมีเป็นจำนวนมากและคุณสมบัติ (attribute) ของตัวอย่างที่เป็นอิสระต่อกัน ซึ่งมีสมการคำนวณดังสมการที่ (3)

$$P(A|B)=\frac{P(B|A)\times P(A)}{P(B)} \qquad (3)$$

โดยที่ $P(A|B)$ คือ ความน่าจะเป็นของเหตุการณ์ A เมื่อมีข้อมูล B (posterior probability)

$P(B|A)$ คือ ความน่าจะเป็นของเหตุการณ์ B เมื่อรู้ว่าเหตุการณ์ A เป็นจริง (likelihood)

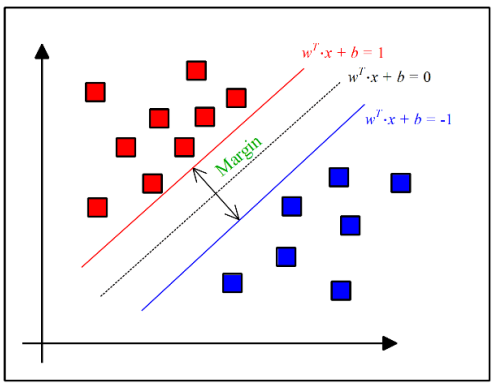
$P(A)$ คือ ความน่าจะเป็นของเหตุการณ์ A (prior probability)

$P(B)$ คือ ความน่าจะเป็นของเหตุการณ์ B (evidence probability)

1.1.4. ตัวแบบเคเนียร์เรสเนเบอร์ (K-nearest neighbors) เป็นอัลกอริทึมที่ใช้ในการจำแนกข้อมูลที่ไม่ค่อยมีความซับซ้อน จึงได้รับความนิยมเป็นอันดับแรก ๆ ในการแก้ปัญหาการจำแนกประเภท แนวคิดของวิธีเคเนียร์เรสเนเบอร์จะทำการจำแนกข้อมูลที่มีคุณสมบัติใกล้เคียงกัน K ตัว และค่า K ส่วนใหญ่จะกำหนดให้เป็นเลขคี่ การจำแนกกลุ่มให้กับข้อมูลชุดทดสอบจะทำการคำนวณระยะทางที่ใกล้กับข้อมูลชุดทดสอบมากที่สุด K ตัว จากนั้นข้อมูลชุดทดสอบจะจัดอยู่ในกลุ่มข้อมูลที่ปรากฏเยอะที่สุดใน K ตัว สำหรับงานวิจัยนี้ใช้สูตรการคำนวณระยะทางแบบยูคลิด (Euclidean distance) ซึ่งระยะทางแบบยูคลิดจากจุด $P(p_1,p_2,\ldots,p_n)$ และจุด $Q(q_1,q_2,\ldots,q_n)$ ในปริภูมิ N มิติสามารถคำนวณได้ตามสมการที่ (4)

$$d(P,Q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \qquad (4)$$

1.1.5. ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) เป็นอัลกอริทึมที่ใช้ในการจำแนกประเภทของข้อมูล (classification) และการทำนาย (regression) แนวคิดสำหรับการจำแนกข้อมูล คือ ตัวแบบจะพยายามลดความผิดพลาดในการจำแนกข้อมูลและหาระยะขอบที่มากที่สุด (maximum margin) ของระนาบตัดสินใจ (decision hyperplane) ในการแบ่งกลุ่มข้อมูล ดังภาพที่ 2



ภาพที่ 2 ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน

สำหรับปัญหาการจำแนกข้อมูลนั้น กำหนดให้ $\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots,(x_n,y_n)\}$ เป็นข้อมูลชุดฝึก โดยที่ $x_i \in R^n$ และ $y_i \in \{-1,1\}$ สำหรับทุก ๆ $i=1,2,3,\ldots,n$ เส้นแบ่งกลุ่มข้อมูลจะเป็นดังสมการ (5)

$$w^T \cdot x + b = 0 \quad (5)$$

เมื่อ w คือ เวกเตอร์น้ำหนัก (weight vector) และ b คือ ค่าไบแอส (bias) จากนั้นเส้นแบ่งกลุ่มที่ดีที่สุด คือ เส้นแบ่งที่มีระยะขอบมากสุด ซึ่งสามารถคำนวณได้จากสมการ (6) และ สมการ (7)

$$w^T \cdot x + b = 1 \quad (6)$$

$$w^T \cdot x + b = -1 \quad (7)$$

สำหรับการจำแนกกลุ่มเมื่อป้อนข้อมูล (input) x_j จำแนกอยู่ในกลุ่มใดพิจารณาจาก ถ้า $w^T \cdot x_j + b \geq 1$ จะจัดอยู่ในกลุ่ม $y=1$ แต่ถ้า $w^T \cdot x_j + b \leq -1$ จะจัดอยู่ในกลุ่ม $y=-1$

ในกรณีที่ข้อมูลไม่สามารถแบ่งกลุ่มได้ด้วยเส้นตรง ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนจะมีวิธีการใช้ฟังก์ชันเคอร์เนล (kernel function) มาช่วยประยุกต์ในการแก้ปัญหาการจำแนกกลุ่มได้ [3]

1.1.6. ตัววัดประสิทธิภาพของตัวแบบ (performance) ในงานวิจัยนี้เราใช้ค่าความแม่นยำ (accuracy) เป็นตัววัดประสิทธิภาพของตัวแบบ ซึ่งค่าความแม่นยำเป็นการคำนวณจากสัดส่วนของจำนวนข้อมูลที่เกิดจากการจำแนกของตัวแบบที่สามารถจำแนกได้ถูกต้องต่อข้อมูลทั้งหมดที่นำมาทดสอบ มีสูตรการคำนวณดังสมการที่ (8) และ ผลจากการจำแนกข้อมูลสามารถแสดงในรูปของตารางเมทริกซ์สับสน (confusion matrix) ดังตารางที่ 1

$$\text{Accuracy} = \left(\frac{TP+TN}{TP+FN+TN+FP}\right)\times 100 \quad (8)$$

เมื่อ

TP (true positive) คือ ตัวแบบทำนายว่าเป็นโรคมะเร็งปอดและตรงกับข้อมูลจริง

TN (true negative) คือ ตัวแบบทำนายว่าไม่เป็นโรคมะเร็งปอดซึ่งตรงกับข้อมูลจริง

FP (false positive) คือ ตัวแบบทำนายว่าเป็นโรคมะเร็งปอดซึ่งไม่ตรงกับข้อมูลจริง

FN (false negative) คือ ตัวแบบทำนายว่าไม่เป็นโรคมะเร็งปอดซึ่งไม่ตรงกับข้อมูลจริง

ตารางที่ 1 แสดงเมทริกซ์สับสน

ค่าความจริง (Actual)	**ค่าทำนาย(Prediction)**	
	บวก (Positive)	ลบ (Negative)
บวก (Positive)	TP = True Positive	FN = False Negative
ลบ (Negative)	FP = False Positive	TN = True Negative

1.2 วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพตัวแบบการเรียนรู้ของเครื่องเพื่อใช้สำหรับการจำแนกผู้ป่วยโรคมะเร็งปอดทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรสเนเบอร์ และ ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน

2. วิธีดำเนินการวิจัย

2.1 กรอบแนวคิดวิจัย

ตัวแปรที่ศึกษามีทั้งหมด 15 ตัวแปร ประกอบด้วยตัวแปรต้น 14 ตัวแปร เป็นตัวแปรที่ใช้ศึกษาเพื่อจำแนกผู้ป่วยโรคมะเร็งปอด ได้แก่ สูบบุหรี่, โรคเล็บเหลือง, ความวิตกกังวล, แรงกดดันอิทธิพลรอบข้าง, กลุ่มโรคเรื้อรัง, อ่อนเพลีย, โรคภูมิแพ้, หายใจมีเสียงหวีด, ดื่มแอลกอฮอล์, มีอาการไอ,

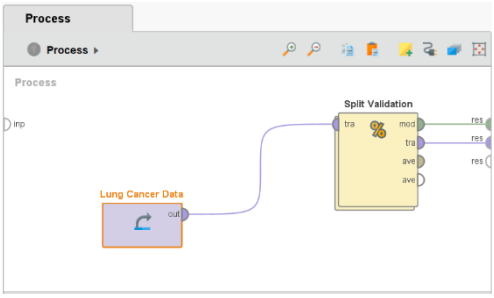
หายใจถี่ตื่น, กลืนอาหารลำบาก, มีอาการเจ็บหน้าอก ข้อมูลของตัวแปรต้นเป็นการเก็บข้อมูลที่เป็นลักษณะการสอบถามอาการซึ่งการเก็บข้อมูลจะเก็บแทนด้วยตัวเลขดังนี้ 1 คือ ไม่ใช่ 2 คือ ใช่ และข้อมูลอายุเป็นการเก็บข้อมูลแบบจำนวนเต็มบวกมีหน่วยเป็นปี ส่วนตัวแปรตาม คือ ผลการจำแนกผู้ป่วยโรคมะเร็งปอด โดยที่ 1 คือ ผู้ป่วยที่ไม่เป็นโรคมะเร็งปอด และ 2 คือ ผู้ป่วยโรคมะเร็งปอด

2.2 วิธีดำเนินการวิจัย

2.2.1. เก็บรวบรวมข้อมูล การตรวจหาโรคมะเร็งปอด(Lung Cancer Detection) จากเว็บไซต์ Kaggle.com [4] ซึ่งเป็นเว็บไซต์สาธารณะที่เปิดให้มีการนำข้อมูลมาศึกษาและพัฒนาตัวแบบทางด้านปัญญาประดิษฐ์ รวมถึงเปิดให้ผู้ที่สนใจในข้อมูลดังกล่าวนำไปศึกษาในด้านที่สนใจ จากการดาวน์โหลดข้อมูลจากเว็บไซต์ดังกล่าวพบว่ามีจำนวนข้อมูลทั้งหมดเท่ากับ 309 ข้อมูล ไฟล์ข้อมูลเป็นไฟล์สกุล csv

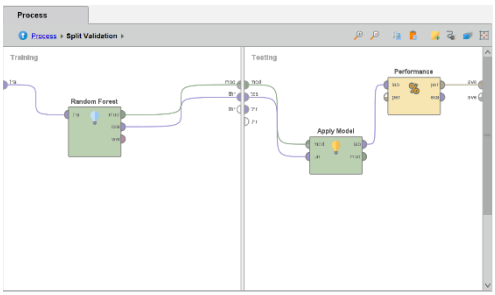
2.2.2. ทำความสะอาดข้อมูล (data cleansing) คือ เป็นการจัดการกับการขาดหายของข้อมูลที่มีการขาดหายไปด้วยค่าของข้อมูลที่ปรากฏเยอะสุดหรือลบข้อมูลที่มีการขาดหายออกจากชุดข้อมูลทั้งหมด เพื่อให้ได้ข้อมูลที่เหมาะสมในการสร้างตัวแบบเพื่อเปรียบเทียบประสิทธิภาพสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด

2.2.3. แบ่งข้อมูลออกเป็นสองส่วน โดยใช้คำสั่ง split validation ในโปรแกรม RapidMiner ซึ่งคำสั่งดังกล่าวเป็นตัวดำเนินการ (operator) แบ่งข้อมูลออกเป็นสองส่วนแบบสุ่ม ส่วนแรกจะถูกนำไปเป็นข้อมูลชุดฝึก (training data) เพื่อใช้ในการสร้างตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอดด้วยการเรียนรู้ของเครื่อง ส่วนที่สองจะเป็นข้อมูลชุดทดสอบ (test data) เพื่อใช้ในการประเมินความแม่นยำของตัวแบบ ซึ่งในงานวิจัยนี้ได้กำหนด ข้อมูลชุดฝึกจำนวน 278 ข้อมูล และข้อมูลชุดทดสอบจำนวน 31 ข้อมูล เมื่อคิดเป็นอัตราส่วนแล้วมีค่าเท่ากับ 9:1



ภาพที่ 3 การใช้คำสั่ง Split Validation ในโปรแกรม RapidMiner

3.2.4. สร้างตัวแบบการทำนายสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอดด้วยการเรียนรู้ของเครื่องทั้งหมด 5 ตัวแบบได้แก่ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรียสเนเบอร์ และ ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน ด้วยโปรแกรม RapidMiner



ภาพที่ 4 การสร้างตัวแบบป่าสุ่มของโปรแกรม RapidMiner

สำหรับตัวแบบอื่น ๆ เราจะใช้คำสั่งในช่อง Operators ของตัวโปรแกรม RapidMiner ซึ่งอยู่ทางด้านซ้ายมือของโปรแกรมแล้วพิมพ์คำสั่งต่อไปนี้

คำสั่ง Decision Tree สำหรับตัวแบบต้นไม้ตัดสินใจ คำสั่ง k-NN สำหรับตัวแบบเคเนียร์เรียสเนเบอร์ คำสั่ง Naive Bayes สำหรับตัวแบบนาอีฟเบย์ และคำสั่ง Support Vector Machine สำหรับตัวแบบซัพพอร์ตเวกเตอร์แมชชีน จากนั้นลากกล่องคำสั่งมาที่ช่อง Training คล้ายกับภาพที่ 3 แล้วกดประมวลผลโปรแกรม

2.2.5. นำข้อมูลชุดทดสอบมาใช้กับตัวแบบแต่ละตัวแบบ โปรแกรม RapidMiner จะแสดงผลการจำแนกผู้ป่วยโรคมะเร็งปอดในรูปแบบของตารางเมทริกซ์สับสน จากนั้นผลของการจำแนกในตารางเมทริกซ์สับสนมาคำนวณค่าความแม่นยำ เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของตัวแบบ

3. ผลการวิจัย

จากข้อมูลทั้งหมดจำนวน 309 ข้อมูล เมื่อเข้าสู่ขั้นตอนความสะอาดข้อมูล พบว่าข้อมูลทั้งหมดมีความเหมาะสมที่จะใช้ในการสร้างตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด โดยพบว่า ข้อมูลชุดนี้มีผู้ป่วยเป็นโรคมะเร็งปอดทั้งสิ้น 270 คน คิดเป็นร้อยละ 87.38 และทำการจำแนกข้อมูลด้านสุขภาพซึ่งถูกแสดงรายละเอียดในตารางที่ 2

ตารางที่ 2 จำนวนและร้อยละของข้อมูลทั้งหมดจำแนกตามข้อมูลด้านสุขภาพพร้อมทั้งแสดงค่าเฉลี่ยของอายุของข้อมูลทั้งหมดจำนวน 309 ข้อมูล

ข้อมูลทั่วไปและข้อมูลด้านสุขภาพ	จำนวน	ร้อยละ	ข้อมูลทั่วไปและข้อมูลด้านสุขภาพ	จำนวน	ร้อยละ
ประวัติการสูบบุหรี่			หายใจมีเสียงหวีด		
เคยสูบบุหรี่	174	56.31	มี	172	55.66
ไม่เคยสูบบุหรี่	135	43.69	ไม่มี	137	44.34
โรคเล็บเหลือง			ดื่มแอลกอฮอล์		
เป็น	176	56.96	ดื่ม	172	55.66
ไม่เป็น	133	43.04	ไม่ดื่ม	137	44.34
ความวิตกกังวล			มีอาการไอ		
มี	154	49.84	มี	179	57.93
ไม่มี	155	50.16	ไม่มี	130	42.07
แรงกดดันอิทธิพลรอบข้าง			หายใจถี่ตื่น		
มี	155	50.16	เป็น	198	64.08
ไม่มี	154	49.84	ไม่เป็น	111	35.92
กลุ่มโรคเรื้อรัง			กลืนอาหารลำบาก		
มี	156	50.49	เป็น	145	46.93
ไม่มี	153	49.51	ไม่เป็น	164	53.07
อ่อนเพลีย			มีอาการเจ็บหน้าอก		
มี	208	67.31	มี	172	55.66
ไม่มี	101	32.69	ไม่มี	137	44.34
โรคภูมิแพ้			มะเร็งปอด		
เป็น	172	55.66	เป็น	270	87.38
ไม่เป็น	137	44.34	ไม่เป็น	39	12.62
อายุ (ปี) ค่าเฉลี่ยเลขคณิตเท่ากับ 62.67 ปี อายุต่ำสุด 21 ปี อายุสูงสุด 87 ปี					

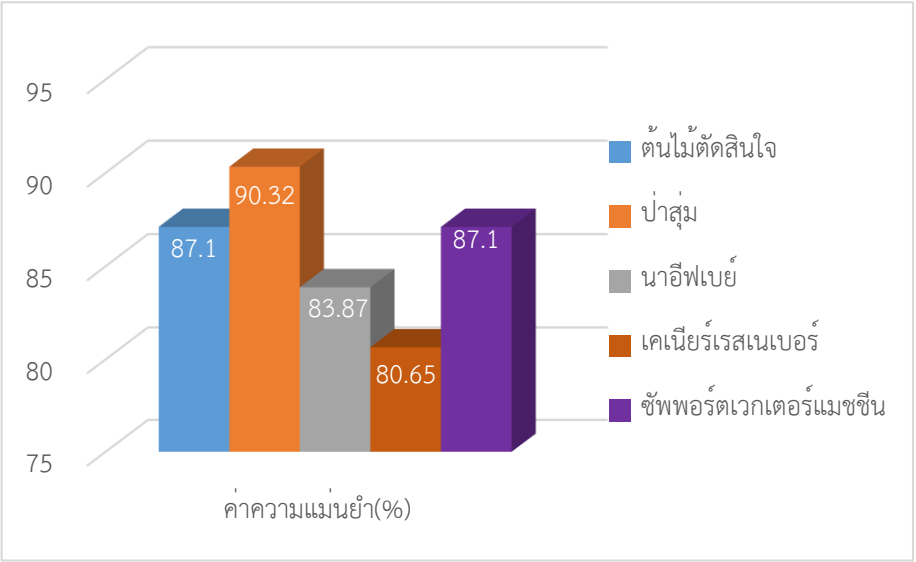
นำข้อมูลทั้งหมดจำนวน 309 ข้อมูล มาทำการฝึกสอนกับตัวแบบการเรียนรู้ของเครื่อง โดยเริ่มจากการแบ่งข้อมูลออกเป็น 2 กลุ่มแบบสุ่ม ดังคำอธิบายในหัวข้อ 2.2.3 สำหรับงานวิจัยนี้แบ่งเป็นข้อมูลชุดฝึกหัดจำนวน 278 ข้อมูล และข้อมูลชุดทดสอบจำนวน 31 ข้อมูล จากนั้นใช้โปรแกรม RapidMiner สร้างตัวแบบทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรสเนเบอร์ และตัวแบบซัพพอร์ตเวกเตอร์แมชชีน เพื่อทำการเปรียบเทียบประสิทธิภาพของแต่ละตัวแบบ และผลของการจำแนกผู้ป่วยโรคมะเร็งปอด โดยใช้ข้อมูลชุดทดสอบทั้งสิ้น 31 ข้อมูล ของแต่ละตัวแบบแสดงในรูปของเมทริกซ์สับสนดังตารางที่ 3

ตารางที่ 3 เมทริกซ์สับสนของตัวแบบ 5 ตัวแบบ

ตัวแบบ	ค่าทำนาย	ค่าจริง เป็นมะเร็งปอด	ค่าจริง ไม่เป็นมะเร็งปอด
ต้นไม้ตัดสินใจ	เป็นมะเร็งปอด	24	1
	ไม่เป็นมะเร็งปอด	3	3
ป่าสุ่ม	เป็นมะเร็งปอด	26	2
	ไม่เป็นมะเร็งปอด	1	2
นาอีฟเบย์	เป็นมะเร็งปอด	25	3
	ไม่เป็นมะเร็งปอด	2	1
เคเนียร์เรียสเนเบอร์	เป็นมะเร็งปอด	25	4
	ไม่เป็นมะเร็งปอด	2	0
ซัพพอร์ตเวกเตอร์แมชชีน	เป็นมะเร็งปอด	26	3
	ไม่เป็นมะเร็งปอด	1	1

จากตารางที่ 3 แสดงผลการจำแนกผู้ป่วยโรคมะเร็งปอดในรูปเมทริกซ์สับสน (confusion matrix) ซึ่งตารางเมทริกซ์สับสนเป็นตารางที่ถูกใช้ในการประเมินประสิทธิภาพของตัวแบบการเรียนรู้ของเครื่องที่ใช้ในการจำแนกข้อมูล จากนั้นคำนวณค่าความแม่นยำในการจำแนกผู้ป่วยโรคมะเร็งปอดของแต่ละตัวแบบตามสมการที่ (8) โดยค่าความแม่นยำแสดงถึงการเปรียบเทียบในภาพที่ 4

จากภาพที่ 5 ผลการประเมินค่าความแม่นยำของตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด พบว่า ตัวแบบป่าสุ่ม มีค่าความแม่นยำมากที่สุด คือ 90.32% ตัวแบบต้นไม้ตัดสินใจ ตัวแบบซัพพอร์ตเวกเตอร์แมชชีน มีค่าความแม่นยำเท่ากัน คือ 87.10% ตัวแบบนาอีฟเบย์ มีค่าความแม่นยำ คือ 83.87% และตัวแบบเคเนียร์เรสเนเบอร์ มีค่าแม่ยำ คือ 80.65%



ภาพที่ 5 การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอด

4. อภิปรายผลและสรุป

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพตัวแบบทำนายการจำแนกผู้ป่วยโรคมะเร็งปอดด้วยตัวแบบทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม ตัวแบบนาอีฟเบย์ ตัวแบบเคเนียร์เรสเนเบอร์ และตัวแบบซัพพอร์ตเวกเตอร์แมชชีน พบว่า สำหรับการใช้ข้อมูลชุดนี้ ตัวแบบป่าสุ่ม ถูกประเมินว่ามีประสิทธิภาพในการจำแนกผู้ป่วยโรคมะเร็งปอดมากที่สุดจากตัวแบบทั้งหมดที่นำมาศึกษา โดยมีความแม่นยำมากที่สุด คือ 90.32% ซึ่งสอดคล้องกับงานวิจัยเรื่องการพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล พบว่า ตัวแบบป่าสุ่มมีความแม่นยำในการทำนายผู้ป่วยเบาหวานมากที่สุดมีค่าความแม่นยำในการทำนาย 99.80% [5] เปรียบเทียบประสิทธิภาพของตัวแบบการเรียนรู้ของเครื่องสำหรับทำนายราคาพริกขี้หนูในจังหวัดนครศรีธรรมราช พบว่า การทำนายโดยใช้ต้นไม้ป่าสุ่มราคาพริกขี้หนูใกล้เคียงกับราคาจริง [6] การเรียนรู้ของเครื่องเพื่อทำนายการรับผลงานวิจัยทางวิชาการโดยใช้ตัวแบบป่าสุ่มมีความแม่นยำในการทำนาย 81% [7] แต่สำหรับงานวิจัย การจำแนกผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล และการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล พบว่า ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในการทำนายสูงสุด โดยมีความแม่นยำเท่ากับ 76.95% [8]

ข้อเสนอแนะสำหรับงานวิจัยครั้งนี้ เนื่องจากข้อมูลที่เรานำมาวิเคราะห์มีจำนวนผู้เป็นมะเร็งปอดมากกว่าจำนวนข้อมูลผู้ที่ไม่เป็นมะเร็งปอด ดังนั้นแนวทางการพัฒนางานวิจัย เสนอให้มีการเก็บรวบรวมข้อมูลของกลุ่มผู้ป่วยที่เป็นมะเร็งปอดและไม่เป็นมะเร็งปอดจำนวนเท่า ๆ กัน และอาจต้องมีการเก็บรวบรวมข้อมูลเฉพาะเจาะจงของประชากรไทย รวมถึงการสร้างแบบจำลองอื่น ๆ

เพื่อให้ได้ตัวแบบสำหรับการจำแนกผู้ป่วยโรคมะเร็งปอดที่มีประสิทธิภาพต่อไป

5. เอกสารอ้างอิง

[1] กรมการแพทย์. กรมการแพทย์เผยมะเร็งปอดเป็นสาเหตุการเสียชีวิตอันดับ 2 ของคนไทย. [อินเตอร์เน็ต]. 2565. [เข้าถึงเมื่อ 10 กันยายน 2566]. เข้าถึงได้จาก: http://pr.moph.go.th/? url=pr/detail /2/02/121812/.

[2] โรงพยาบาลบำรุงราษฎร์. มะเร็งปอด. [อินเตอร์เน็ต]. 2566. [เข้าถึงเมื่อ 10 กันยายน 2566]. เข้าถึงได้จาก: http://www.bumrungrad.com/th/conditions/lung-cancer.

[3] ปริญญา สงวนสัตย์. AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง. พิมพ์ครั้งที่ 1. นนทบุรี: สำนักพิมพ์ไอดีซี พรีเมียร์: 2562.

[4] Lung Cancer Detection. [อินเตอร์เน็ต]. 2565. [เข้าถึงเมื่อ 1 กันยายน 2566]. เข้าถึงได้จาก: https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection.

[5] กฤตกนก ศรีพิมพ์สอ, และกิตติพล วิแสง. การพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล. วารสารวิชาการการจัดการเทคโนโลยีเทคโนโลยีมหาวิทยาลัยราชภัฏมหาสารคาม. 2566;10(1): 51-63.

[6] โสภี แก้วชะฎา, สมพร เรืองอ่อน, และอุทัย คูหาพงศ์. เปรียบเทียบประสิทธิภาพของตัวแบบการเรียนรู้ของเครื่องสำหรับทำนายราคาพริกขี้หนูในจังหวัดนครศรีธรรมราช. วารสารวิชาการซายน์เทค มรภ.ภูเก็ต. 2565;6(2): 1-11.

[7] Skorikov, M., & Momen, S. (2020). Machine learning approach to predicting the acceptance of academic papers (pp113-116). In *The 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT).*

[8] รุ่งโรจน์ บุญมา, และนิเวศ จิระวิชิตชัย. การจำแนกผู้ป่วยโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล. วารสารวิชาการซายน์เทค มรภ.ภูเก็ต. 2562;3(2): 11-19.