Rattanakosin Journal of Science and Technology: RJST
2024, Volume 6 Issue 1: 34-42
Research Article
Online ISSN: 2673-0553

ตัวแบบทำนายมะเร็งปากมดลูกด้วยการเรียนรู้ของเครื่อง

Prediction model for cervical cancer by using machine learning

ภัคพล สวัสกมล[1] และ ศวิตา ทองขุนวงศ์[2*]
[1] สาขาวิชาคณิตศาสตร์และภูมิสารสนเทศ มหาวิทยาลัยเทคโนโลยีสุรนารี
[2]โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี มหาวิทยาลัยเทคโนโลยีสุรนารี
[*]E-mail: sawiphak2424@gmail.com

บทคัดย่อ

จุดมุ่งหมายของการศึกษาครั้งนี้ คือ เปรียบเทียบประสิทธิภาพตัวแบบทำนายมะเร็งปากมดลูกด้วยการเรียนรู้ของเครื่องด้วยวิธีต้นไม้ตัดสินใจ เกรเดียนท์บูตทรี ป่าสุ่ม และการเรียนรู้เชิงลึก โดยการศึกษาครั้งนี้ได้นำข้อมูลมาจากเว็ปไซต์ Kaggle.com จำนวน 110 ข้อมูล ประกอบด้วยตัวแปรต้นทั้งหมด 21 ตัวแปร ได้แก่ อายุ จำนวนคู่นอน อายุที่เริ่มเพศสัมพันธ์ครั้งแรก จำนวนครั้งการตั้งครรภ์ อัตราการสูบบุหรี่ต่อปี การใช้ยาคุมกำเนิด จำนวนปีที่ใช้ยาคุมกำเนิด การใส่ห่วงคุมกำเนิด จำนวนปีที่ใส่ห่วงคุมกำเนิด โรคติดต่อทางเพศสัมพันธ์ หูดที่อวัยวะเพศ หูดบริเวณปากมดลูก หูดบริเวณปากช่องคลอด โรคซิฟิลิซ การติดเชื้อบริเวณอุ้งเชิงกราน การติดเชื้อไวรัสเฮอร์พีส์บริเวณอวัยวะเพศ หูดข้าวสุก โรคเอดส์ การติดเชื้อไวรัสเอชไอวี การติดเชื้อไวรัสเอชพีวี และไวรัสตับอักเสบบี และตัวแปรตาม 1 ตัวแปร จำแนกออกเป็น 2 กลุ่ม คือ เป็นและไม่เป็นมะเร็งปากมดลูก

ผลการศึกษา พบว่า ตัวแบบทำนายมะเร็งปากมดลูกด้วยวิธีการเรียนรู้เชิงลึกมีประสิทธิภาพในการทำนายดีที่สุดโดยมีค่าความแม่นยำในการทำนายสูงสุดเท่ากับ 95.45%

คำสำคัญ: มะเร็งปากมดลูก การเรียนรู้ของเครื่อง การเรียนรู้เชิงลึก

Abstract

The aim of this study is to compare predictive performance of cervical cancer prediction models using machine learning techniques, including decision trees, gradient boosted trees, random forests, and deep learning. The study collected data from Kaggle.com, comprising 110

* Corresponding author, e-mail: sawiphak2424@gmail.com
Received 24 11 2023 / Revised: 27 01 2024 / Accepted: 15 02 2024

data points, with 21 independent variables, such as age, number of sexual partners, age of first sexual intercourse, number of pregnancies, smoking per year, contraceptive pill use, year of contraceptive pill use, intrauterine device use, years of intrauterine device use, sexually transmitted infections, genital herpes, genital warts, syphilis, cervical infection, herpes simplex virus infection, rice-sized cervical polyps, AIDS, HIV infection, HPV infection, and hepatitis B virus infection. The dependent variable is binary, classifying individuals into two groups: those with cervical cancer and those without.

The study results indicate that the deep learning-based cervical cancer prediction model performs with the highest predictive efficiency, achieving an accuracy rate of 95.45%.

Keywords: Cervical cancer, Machine learning, Deep Learning

1. ที่มาและความสำคัญ

จากสถิติล่าสุดเมื่อเดือนกุมภาพันธ์ 2566 ที่เปิดเผยโดยกระทรวงสาธารณสุขในวันมะเร็งโลก โรคมะเร็งจัดเป็นปัญหาสำคัญทางสาธารณสุข เนื่องจากประเทศไทยพบผู้ป่วยรายใหม่ปีละประมาณ 140,000 ราย และในจำนวนนี้มีผู้เสียชีวิตประมาณ 80,000 รายต่อปี โดยโรคมะเร็งปากมดลูกเป็นโรคมะเร็งที่พบมากที่สุดในกลุ่มโรคมะเร็งระบบสืบพันธุ์ของสตรี และจากข้อมูลสถิติล่าสุดในเดือนมีนาคม 2566 ของ (IARC) International Agency for Research on Cancer พบว่า ประเทศไทยมีผู้ป่วยโรคมะเร็งปากมดลูกปีละ 9,158 ราย และมีอัตราเสียชีวิต 4,705 รายต่อปี [1]

อวัยวะสืบพันธุ์ของสตรีจะมีทั้งส่วนที่มองเห็นได้จากภายนอกร่างกายและส่วนที่อยู่ลึกเข้าไปในบริเวณอุ้งเชิงกราน โดยหากดูสถิติโรคมะเร็งอวัยวะสืบพันธุ์ของสตรี โดยส่วนใหญ่มักจะเป็นตำแหน่งที่อยู่ลึกเข้าไปในร่างกายซึ่งไม่สามารถสังเกตได้จากภายนอก โดยเฉพาะปากมดลูกนั้นจะอยู่เข้าไปด้านในสุดของช่องคลอด ดังนั้นการวินิจฉัยมะเร็งปากมดลูกจึงมีความจำเป็นต้องซักประวัติความเสี่ยง เพื่อวิเคราะห์พฤติกรรมผู้ป่วย และทำการตรวจสุขภาพประจำปี ซึ่งการตรวจพบเจอโรคมะเร็งปากมดลูกในกรณีที่เป็นระยะก่อนมะเร็ง การแพทย์ปัจจุบันสามารถทำให้ผู้ป่วยมีโอกาสหายขาดและไม่ทำให้กลายเป็นมะเร็งปากมดลูกในอนาคต

เนื่องจากการนำเทคโนโลยีทางด้านการเรียนรู้ของเครื่อง (machine learning) มาช่วยในการวินิจฉัยทางด้านการแพทย์ จะช่วยคัดกรองผู้ป่วยที่มีความเสี่ยงต่อการเป็นโรคมะเร็งปากมดลูกมีประสิทธิภาพมากขึ้น และไม่เสียโอกาสที่จะเข้ารับการรักษาในตั้งแต่ระยะเริ่มต้นที่ตรวจพบ เนื่องจากเป็นเทคโนโลยีในปัจจุบันที่ทันสมัยและเน้นการเรียนรู้จากข้อมูลตัวอย่างเพื่อพัฒนาออกมาเป็นตัวแบบทำนาย ทำให้มีความแม่นยำในการวินิจฉัยโรคได้

2. วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพตัวแบบทำนายมะเร็งปากมดลูกด้วยวิธีการเรียนรู้ของเครื่องทั้งหมด 4 วิธี ได้แก่ ต้นไม้ตัดสินใจ (decision tree) ป่าสุ่ม (random forest) เกรเดียนท์บูตทรี (gradient boosted trees) และการเรียนรู้เชิงลึก (deep learning)

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การเรียนรู้ของเครื่องแบ่งออกเป็น 3 ประเภท ดังนี้ [6]

1. การเรียนรู้แบบมีผู้สอน (supervised learning) การเรียนรู้ชนิดนี้จะแบ่งออกเป็น 2 แบบ คือ การจำแนกประเภทข้อมูล (classification) และการถดถอย (regression)

2. การเรียนรู้แบบไม่ต้องมีผู้สอน (unsupervised learning) การเรียนรู้ชนิดนี้จะแบ่งออกเป็น 2 ประเภท คือ การจัดกลุ่ม (clustering) การหาความสัมพันธ์ (association)

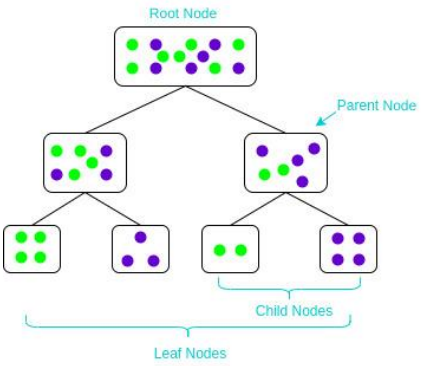
3. การเรียนรู้ที่ต้องอาศัยการป้อนผลลัพธ์กลับ (reinforcement learning) เป็นการเรียนรู้ที่ต้องอาศัยการป้อนกลับของผลลัพธ์กลับมาปรับปรุงการเรียนรู้ของตนเอง

ในการศึกษาครั้งนี้ การเรียนรู้ของเครื่องด้วยวิธีต้นไม้ตัดสินใจ วิธีป่าสุ่ม วิธีเกรเดียนท์บูตทรี และวิธีการเรียนรู้เชิงลึก จัดเป็นการเรียนรู้แบบมีผู้สอน

ตัวแบบสำหรับการจำแนกประเภท (models for classification)

1. ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ จะมีรูปแบบเป็นการจำลองการตัดสินใจของมนุษย์ เป็นการเรียนรู้จากคุณลักษณะของข้อมูลแล้วสร้างแผนผังการตัดสินใจที่มีความคล้ายกับต้นไม้ โดยผลลัพธ์ที่ได้จะมีสองกลุ่มหรืออาจมากกว่าสองกลุ่ม



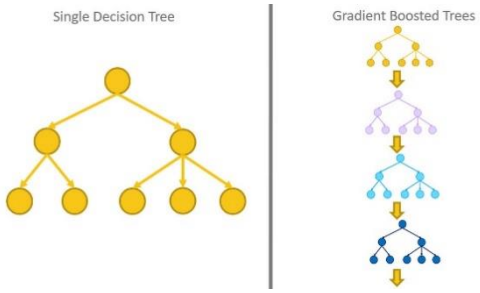
รูปที่ 1 แสดงโครงสร้างของต้นไม้ตัดสินใจ [2]

ต้นไม้ตัดสินใจ ประกอบไปด้วยโหนด (nodes) และรองโหนด (leaves) โหนดแรกที่เริ่มต้น เรียกว่า “โหนดราก” (root node) และแบ่งส่วนข้อมูลออกเป็นส่วนย่อย ๆ ด้วยการตัดสินใจที่แต่ละโหนด โดยมีการใช้เงื่อนไข (conditions) เพื่อแบ่งข้อมูลออกเป็นสองกลุ่มและกระจายข้อมูลไปยังโหนดย่อย (child nodes) สองโหนดนั้นจะมีเงื่อนไขต่อไปอีก และกระจายข้อมูลต่อไปในโหนดย่อยของมันเรื่อย ๆ จนกระจายข้อมูลไปถึงโหนดใบ (leaf nodes) ที่ไม่มีการแบ่งข้อมูลต่อแล้ว ซึ่งเป็นการให้ผลลัพธ์ของกระบวนการตัดสินใจ

ต้นไม้ตัดสินใจมักถูกใช้งานที่ต้องการตัดสินใจหรือการจำแนกข้อมูล เช่น ในทางการค้าเราจะจำแนกออกมาว่าลูกค้าจะซื้อสินค้าหรือไม่ซื้อสินค้า รวมถึงในทางการแพทย์เราสามารถประยุกต์ใช้เพื่อวินิจฉัยว่าเป็นโรคหรือไม่เป็นโรค

2. เกรเดียนท์บูตทรี

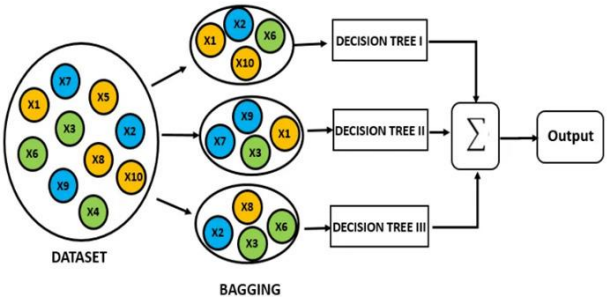
เป็นวิธีที่มีหลักการทำงานคล้ายคลึงกับต้นไม้ตัดสินใจ แต่มีความแตกต่างในขั้นตอนการสร้างและปรับปรุงประสิทธิภาพของตัวแบบ เพื่อเพิ่มความแม่นยำในการทำนายผลลัพธ์ เกรเดียนท์บูตทรีจะใช้เทคนิค Ensemble Learning ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องที่นำต้นไม้ตัดสินใจหลายต้นรวมเข้าด้วยกัน ในขั้นตอนนี้ต้นไม้แต่ละต้นจะถูกสร้างขึ้นมาใหม่โดยมีความสัมพันธ์กับต้นไม้ที่สร้างก่อนหน้า โดยที่ต้นไม้ใหม่จะให้ความสำคัญกับค่าความผิดพลาดของต้นไม้ก่อนหน้ามากกว่าค่าความถูกต้อง และพยายามปรับปรุงค่าผิดพลาดนี้โดยกระบวนการ Gradient Descent จนกว่าจะได้ตัวแบบที่มีความแม่นยำสูงสุด ในขั้นตอนสุดท้ายเกรเดียนท์บูตทรีจะรวมผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจทุกต้นเข้าด้วยกัน เพื่อทำนายผลลัพธ์ได้อย่างมีประสิทธิภาพ



รูปที่ 2 แสดงโครงสร้างเกรเดียนท์บูตทรี [3]

3. ป่าสุ่ม

เป็นอัลกอริทึมที่ใช้ในการทำนายในรูปแบบของการจำแนกข้อมูลและการพยากรณ์ โดยหลักการทำงานของป่าสุ่มจะเป็นการสุ่มเอาข้อมูลไปสร้างเป็นต้นไม้ตัดสินใจหลาย ๆ ต้น ทำให้มีลักษณะคล้ายป่า ซึ่งต้นไม้แต่ละต้นจะได้รับข้อมูล (data set) ที่แตกต่างกัน แต่เป็นส่วนหนึ่งของข้อมูลทั้งหมด ในทางคณิตศาสตร์จะมีความหมายว่าเป็นสับเซต (subset) ของข้อมูลทั้งหมด เมื่อทำการทำนายก็ให้ต้นไม้แต่ละต้นทำการตัดสินใจ ผลลัพธ์ที่ถูกโหวตมากสุดจะถูกเลือกเป็นผลลัพธ์ที่ทำนายออกมา

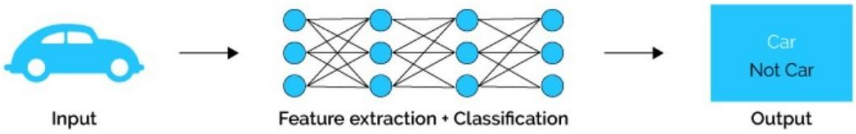


รูปที่ 3 แสดงโครงสร้างป่าสุ่ม [4]

วิธีป่าสุ่ม เป็นอัลกอริทึมที่ที่มีประสิทธิภาพและสามารถใช้งานได้หลากหลาย เช่น การจำแนกออกมาว่าอีเมลเป็นขยะหรือไม่ การพยากรณ์ราคาสินทรัพย์ หรือประยุกต์ใช้กับปัญหาในการจำแนกข้อมูล

4. การเรียนรู้เชิงลึก

เป็นการเรียนรู้ของเครื่องที่เลียนแบบคล้ายโครงข่ายประสาทของมนุษย์ โดยนำระบบโครงข่ายประสาท (neural network) มาซ้อนกันหลายชั้น (layer) และทำการเรียนรู้จากชุดข้อมูลฝึกสอนเพื่อใช้ในการแยกกลุ่มข้อมูล การเรียนรู้เชิงลึกมีลักษณะเด่น คือ การใช้โครงข่ายที่มีจำนวนชั้นและโหนด (neurons) เป็นจำนวนมาก เพื่อให้ระบบสามารถเรียนรู้ได้อัตโนมัติจากข้อมูลเพื่อทำการสร้างรายละเอียดในการจำแนกหรือการทำนายให้มีความแม่นยำมากขึ้น นอกจากนี้ การเรียนรู้เชิงลึกยังสามารถทำงานกับข้อมูลที่มีขนาดใหญ่และจัดการกับข้อมูลที่มีความซับซ้อนได้อย่างมีประสิทธิภาพ



รูปที่ 4 แสดงโครงสร้างของ การเรียนรู้เชิงลึก [5]

5. ตัววัดประสิทธิภาพ (performance)

ในงานวิจัยครั้งนี้เราใช้ตารางเมทริกซ์สับสน (confusion matrix) เพื่อคำนวณค่าความแม่นยำในการทำนายของตัวแบบ ซึ่งตารางเมทริกซ์สับสนเป็นตารางที่ใช้สรุปจำนวนข้อมูลที่ตัวแบบทำนายจำแนกได้ถูกต้องและไม่ถูกต้อง ถือเป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย ที่ทำนายจากตัวแบบที่เราสร้างขึ้นโดยใช้แนวคิดจากคำนวณสัดส่วนของสิ่งที่เกิดจากการทำนายกับสิ่งที่เกิดขึ้นจริง

ตารางที่ 1 แสดงตารางเมทริกซ์สับสน

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

โดยที่ TP คือ ตัวแบบทำนายว่าเป็นมะเร็งปากมดลูกตรงกับข้อมูลที่เกิดขึ้นจริง
FP คือ ตัวแบบทำนายว่าเป็นมะเร็งปากมดลูกแต่ข้อมูลจริงไม่เป็นมะเร็งปากมดลูก
FN คือ ตัวแบบทำนายว่าไม่เป็นมะเร็งปากมดลูกแต่ข้อมูลจริงเป็นมะเร็งปากมดลูก
TN คือ ตัวแบบทำนายว่าไม่เป็นมะเร็งปากมดลูกตรงกับข้อมูลที่เกิดขึ้นจริง

ค่าความแม่นยำ คือ ค่าที่ใช้วัดประสิทธิภาพของตัวแบบทำนาย สามารถคำนวณได้จากสมการ (1)

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

4. วิธีดำเนินการวิจัย

4.1 เก็บรวบรวมข้อมูล Cervical Cancer จากเว็บไซต์ Kaggle.com [7] เป็นเว็บไซต์สาธารณะที่สามารถนำข้อมูลจากเว็บไซต์ดังกล่าวมาศึกษาทางด้านปัญญาประดิษฐ์ หลังจากดาวน์โหลดข้อมูลจากเว็บไซต์ดังกล่าวพบว่ามีจำนวนข้อมูลทั้งหมดเท่ากับ 858 ข้อมูล เมื่อเข้าสู่กระบวนการทำความสะอาดข้อมูล (data cleansing) คือ การกำจัดข้อมูลบางข้อมูลที่ไม่ครบถ้วนหรือมีการสูญหายออกจากข้อมูลชุดฝึก ทำให้เหลือข้อมูลที่สมบูรณ์ที่เหมาะนำไปสร้างตัวแบบจำนวน 660 ข้อมูล แบ่งออกเป็นโรคมะเร็งปากมดลูก 17 ข้อมูล และไม่เป็นมะเร็งปากมดลูกจำนวน 643 ข้อมูล จากข้อมูลดังกล่าวจะพบว่าข้อมูลที่ไม่เป็นมะเร็งปากมดลูกมีจำนวนมากกว่าข้อมูลที่เป็นมะเร็งปากมดลูก หากใช้ข้อมูลทั้งหมดในการฝึกจะทำให้ตัวแบบมีความเอนเอียงที่จะทำนายว่าไม่เป็นมะเร็งปากมดลูกได้ สำหรับการศึกษาครั้งนี้เราใช้ข้อมูลสำหรับการสร้างตัวแบบทั้งสิ้น 110 ข้อมูล แบ่งเป็นข้อมูลที่เป็นมะเร็งปากมดลูกจำนวน 17 ข้อมูล และข้อมูลที่ไม่เป็นมะเร็งปากมดลูกจำนวน 93 ข้อมูล ซึ่งได้มาจากการเลือกข้อมูลแบบสุ่มจากข้อมูล 643 ข้อมูล อัตราส่วนของข้อมูลที่เป็นมะเร็งปากมดลูกต่อข้อมูลที่ไม่เป็นมะเร็งปากมดลูกเท่ากับ 2 : 11 ซึ่งโดยปกติแล้วเรามักนิยมให้ข้อมูลชุดฝึกในแต่ละกลุ่มมีจำนวนเท่า ๆ กัน แต่จากข้อมูลที่เรามีพบว่าหากเราเลือกข้อมูลที่ไม่เป็นมะเร็งปากมดลูกมา 17 ข้อมูล จาก 643 ข้อมูล จะทำให้มีข้อมูลสำหรับการนำไปสร้างตัวแบบทำนายจำนวนทั้งสิ้น 34 ข้อมูล ซึ่งการมีข้อมูลชุดฝึกที่น้อยเกินไปก็อาจทำให้เกิดปัญหา Overfitting ได้

4.2 แบ่งข้อมูลออกเป็นสองส่วน ส่วนแรกนำไปเป็นข้อมูลชุดฝึก (training data) และส่วนที่สองนำไปเป็นข้อมูลชุดฝึก (test data) ในงานวิจัยครั้งนี้เราแบ่งข้อมูลชุดฝึกหัดต่อข้อมูลชุดทดทสอบทั้งหมด 5 อัตราส่วน ได้แก่ 90 : 10, 85 : 15, 80 : 20, 75 : 25 และ 70 : 30

4.3 ตัวแปรต้นทั้งหมด 21 ตัวแปร ได้แก่ อายุ จำนวนคู่นอน อายุที่เริ่มเพศสัมพันธ์ครั้งแรก จำนวนครั้งการตั้งครรภ์ อัตราการสูบบุหรี่ต่อปี การใช้ยาคุมกำเนิด จำนวนปีที่ใช้ยาคุมกำเนิด การใส่ห่วงคุมกำเนิด จำนวนปีที่ใส่ห่วงคุมกำเนิด โรคติดต่อทางเพศสัมพันธ์ หูดที่อวัยวะเพศ หูดบริเวณปากมดลูก หูดบริเวณปากช่องคลอด โรคซิฟิลิซ การติดเชื้อบริเวณอุ้งเชิงกราน การติดเชื้อไวรัสเฮอร์พีส์บริเวณอวัยวะเพศ หูดข้าวสุก โรคเอดส์ การติดเชื้อไวรัสเอชไอวี การติดเชื้อไวรัสเอชพีวี และไวรัสตับอักเสบบี และตัวแปรตาม 1 ตัวแปร จำแนกออกเป็น 2 กลุ่ม คือ เป็นและไม่เป็นมะเร็งปากมดลูก

4.4 สร้างตัวแบบทำนายด้วยวิธีการเรียนรู้ของเครื่อง ทั้งหมด 4 วิธี ได้แก่ ต้นไม้ตัดสินใจ ป่าสุ่ม เกรเดียนท์บูตทรี และการเรียนรู้เชิงลึก โดยใช้โปรแกรม RapidMiner Version 10.1

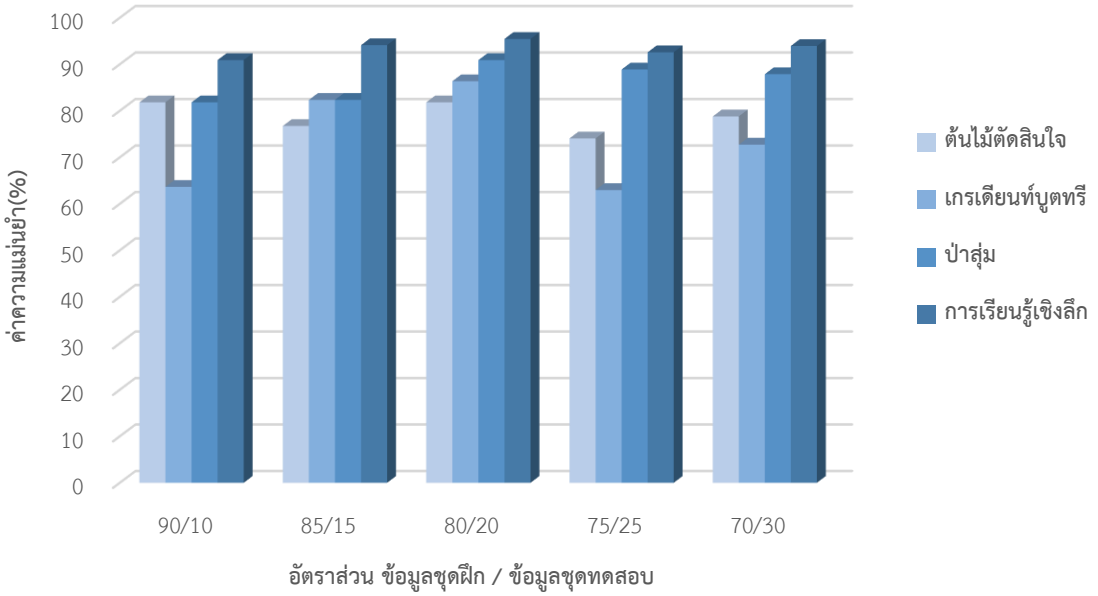
4.5 วัดประสิทธิภาพของตัวแบบทำนายด้วยค่าความแม่นยำ (accuracy) โดยใช้ตารางเมทริกซ์สับสนสำหรับการคำนวณค่าความแม่นยำในการทำนายของตัวแบบ

5. ผลและวิจารณ์

ผลจากการวิเคราะห์ข้อมูลจำนวนทั้งสิ้น 110 ข้อมูล โดยเริ่มการการแบ่งข้อมูลออกเป็นสองส่วน (split validation) คือ ข้อมูลชุดฝึก และข้อมูลชุดทดสอบ ในการวิจัยนี้แบ่งเป็นอัตราส่วนต่าง ๆ แสดงในตารางที่ 2 จากนั้น สร้างตัวแบบทั้งหมด 4 ตัวแบบ เพื่อทำการเปรียบเทียบประสิทธิภาพในการทำนายโรคมะเร็งปากมดลูกด้วยโปรแกรม RapidMiner ซึ่งได้ผลการวิเคราะห์ ดังนี้

ตารางที่ 2 แสดงผลค่าความแม่นยำของตัวแบบเมื่อแบ่งข้อมูลชุดฝึกต่อข้อมูลชุดทดสอบในอัตราส่วนต่าง ๆ

อัตราส่วน	ค่าความแม่นยำในการทำนาย(%)			
ข้อมูลชุดฝึก : ข้อมูลทดสอบ	ต้นไม้ตัดสินใจ	เกรเดียนท์บูตทรี	ป่าสุ่ม	การเรียนรู้เชิงลึก
90 : 10	81.82	63.65	81.82	90.91
85 : 15	76.47	82.35	82.35	94.12
80 : 20	81.82	86.36	90.91	95.45
75 : 25	74.07	62.96	88.89	92.59
70 : 30	78.79	72.73	87.77	93.94



รูปที่ 6 แสดงการเปรียบเทียบค่าความแม่นยำของตัวแบบ

จากตารางที่ 2 แสดงค่าความแม่นยำของประสิทธิภาพการทำนายของตัวแบบทั้งหมด 4 ตัวแบบ โดยแบ่งข้อมูลชุดฝึกต่อข้อมูลชุดทดสอบในอัตราส่วนต่าง ๆ พบว่า ตัวแบบที่มีความแม่นยำในการทำนายโรคมะเร็งปากมดลูก คือ ตัวแบบการเรียนรู้เชิงลึก (deep learning) ซึ่งมีค่าความแม่นยำในการทำนายสูงสุดเท่ากับ 95.45% เมื่อแบ่งข้อมูลชุดฝึกต่อข้อมูลชุดทดสอบในอัตราส่วน 80 : 20 และแสดงผลการทำนายดังตารางที่ 3

ตารางที่ 3 แสดงเมทริกซ์สับสนของตัวแบบการเรียนรู้เชิงลึกที่ทำนายการเป็นมะเร็งปากมดลูกเมื่อแบ่งข้อมูลชุดฝึกต่อข้อมูลชุดทดสอบในอัตราส่วนต่าง ๆ

อัตราส่วน ข้อมูลชุดฝึก : ข้อมูลชุดทดสอบ	ค่าทำนาย	ค่าจริง ไม่เป็นมะเร็ง	ค่าจริง เป็นมะเร็ง	ค่าความแม่นยำ
90 : 10	ไม่เป็นมะเร็ง	8	0	90.91%
	เป็นมะเร็ง	1	2	
85 : 15	ไม่เป็นมะเร็ง	14	1	94.12%
	เป็นมะเร็ง	0	2	
80 : 20	ไม่เป็นมะเร็ง	18	0	95.45%
	เป็นมะเร็ง	1	3	
75 : 25	ไม่เป็นมะเร็ง	23	2	92.59%
	เป็นมะเร็ง	0	2	
70 : 30	ไม่เป็นมะเร็ง	28	2	93.94%
	เป็นมะเร็ง	0	3	

ตัวแบบการเรียนรู้เชิงลึก เป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง แต่มีข้อแตกต่างจากทั้ง 3 ตัวแบบ คือ ตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่มและตัวแบบเกรเดียนท์บูตทรี จะทำการเรียนรู้จากข้อมูลชุดฝึก (training data) ซึ่งข้อมูลที่เรานำเข้าไปฝึกนั้นจะต้องคัดเลือกคุณลักษณะเด่น (feature) ที่มีประโยชน์ต่อการจำแนกข้อมูล ซึ่งการคัดเลือกคุณลักษณะที่กล่าวมานั้นจะกระทำโดยมนุษย์ หลังจากนั้นคอมพิวเตอร์ก็จะทำการเรียนรู้ (train) เพื่อให้ได้ตัวแบบ (model) สำหรับการทำนายออกมา หากพบว่าผลของการทำนายนั้นไม่แม่นยำ การแก้ไขปรับปรุงประสิทธิภาพของตัวแบบก็จะกระทำโดยมนุษย์ แต่สำหรับตัวแบบการเรียนรู้เชิงลึกนั้น จะใช้เทคนิคของเครือข่ายประสามเทียม (artificial neural network) ที่มีความลึก (deep) หลายชั้น ซึ่งมีการทำงานเลียนแบบคล้ายเซลล์สมองของมนุษย์ เมื่อเราป้อนข้อมูลชุดฝึกเข้าไป ตัวแบบการเรียนรู้เชิงลึกจะทำการค้นหาคุณลักษณะเอง (feature extraction) ด้วยเครือข่ายหลายชั้น ดังรูปที่ 4 ส่วนการทำนายนั้นตัวแบบจะสามารถตัดสินใจได้ด้วยตัวเองว่าผลการทำนายนั้นมีความแม่นยำหรือไม่ผ่านเครือข่ายประสาท (neural network) ซึ่งข้อดีของการเรียนรู้เชิงลึก คือ การที่ตัวแบบสามารถเลือกคุณลักษณะเด่นที่จำเป็นต่อการทำนายอาจเป็นที่มาของการทำให้มีประสิทธิภาพในการทำนายที่ดีกว่าตัวแบบต้นไม้ตัดสินใจ ตัวแบบป่าสุ่ม และตัวแบบเกรเดียนท์บูตทรี

6. สรุปผล

งานวิจัยนี้ได้นำข้อมูลของโรคมะเร็งปากมดลูกจากเว็ปไซต์ Kaggle.com มาทำการสร้างตัวแบบในการจำแนกผู้ป่วยโรคมะเร็งปากมดลูก โดยการประยุกต์ใช้ตัวแบบการเรียนรู้ของเครื่องทั้งหมด 4 ตัวแบบ ซึ่งข้อมูลมีความน่าสนใจ เพราะเป็นการเก็บข้อผู้ป่วยที่มีประวัติทั้งเป็นและไม่เป็นมะเร็งปากมดลูก แต่เนื่องจากข้อมูลที่ได้มานั้นมีจำนวนผู้ป่วยที่ไม่เป็นมะเร็งปากมดลูกมากกว่าผู้ที่เป็นมะเร็งปากมดลูกจำนวนมาก ในงานวิจัยครั้งนี้จึงเลือกสุ่มข้อมูลบางส่วนของผู้ที่ไม่เป็นมะเร็งปากมดลูกมาวิเคราะห์ ซึ่งทำให้ข้อมูล

ผู้ที่ถูกวินิจฉัยว่าเป็นมะเร็งปากมดลูกต่อผู้ที่ไม่เป็นมะเร็งปากมดลูก เท่ากับ 2 : 11 จากผลการวิเคราะห์ข้อมูลก็พบว่าตัวแบบการเรียนรู้เชิงลึกมีความแม่นยำในการทำนายสูงสุดเมื่อแบ่งข้อมูลชุดฝึกหัดต่อข้อมูลชุดทดสอบในอัตราส่วน 90 : 10, 85 : 15, 80 : 20, 75 : 25 และ 70 : 30 นอกจากนั้นยังพบว่าค่าความแม่นยำในการทำนายสูงสุดเท่ากับ 95.45% เมื่อแบ่งข้อมูลชุดฝึกหัดต่อข้อมูลชุดทดสอบในอัตราส่วน 80 : 20

7. ข้อเสนอแนะ

ข้อเสนอแนะสำหรับงานวิจัยครั้งนี้ เนื่องจากข้อมูลที่ได้นำมาวิเคราะห์มีจำนวนผู้ป่วยที่ไม่เป็นมะเร็งปากมดลูกมากกว่าจำนวนผู้ป่วยที่เป็นมะเร็งปาดมดลูก ดังนั้น หากนำไปใช้ในการปฏิบัติทางการแพทย์ ผู้วิจัยจึงเสนอให้มีการเก็บรวบรวมข้อมูลของผู้ป่วยที่เป็นมะเร็งและไม่เป็นมะเร็งปากมดลูกในจำนวนที่เท่า ๆ กัน เพราะจะทำให้การเรียนรู้ของเครื่องไม่เกิดความเอนเอียง นอกจากนี้หากนำมาประยุกต์ใช้กับข้อมูลผู้ป่วยของประเทศไทย ต้องนำตัวแบบการเรียนรู้ของเครื่องวิธีต่าง ๆ มาทดลองหาตัวแบบที่เหมาะสมเพื่อให้การทำนายเกิดประสิทธิภาพสูงสุดกับรูปแบบข้อมูลที่มี

8. เอกสารอ้างอิง

[1] พญ.สุขุมาลย์ สว่างวารี. (2566). มะเร็งสตรี รู้ก่อน รักษาไว มีโอกาสหาย ป้องกันได้ คุณภาพชีวิตดี. https://www.chaophya.com/2023/04/มะเร็งสตรี/

[2] Sharma, A. (2023, Sep 18). 4 Simple Ways to Split a Decision Tree in Machine Learning. https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

[3] Silipo, R. (2020, Mar 19). Ensemble Models: Bagging & Boosting. https://medium.com/analytics-vidhya/ensemble-models-bagging-boosting-c33706db0b0b

[4] D'Souza, J. (2018, Mar 20). A trip to Random Forest. https://medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a

[5] ดร.ไพรสันต์ ผดุงเวียง. (2563). สรุปเนื้อหาในหลักสูตร Data Scientist Essentials ตอนที่ 7 Introduction to deep learning. https://rdbi.co.th/2020/01/data-scientist-7/

[6] Sarabun, K. (2020). Learning - Data Science and AI: Machine Learning with Python. Bangkok: Media Network Publisher.

[7] Gokagglers. (2018). Cervical Cancer Risk Classification. https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification

[8] กอบเกียรติ สระอุบล. (2565). เรียนรู้ AI: Deep Learning ด้วย Python (พิมพ์ครั้งที่ 1). สำนักพิมพ์อินเตอร์มีเดีย.