

การพัฒนาแบบจำลองคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุโดยใช้เทคนิคเหมืองข้อมูลเพื่อสนับสนุนการเฝ้าระวังสุขภาพระดับชุมชน

Development of a Data Mining-Based Risk Screening Model for Non-Communicable Diseases among Older Adults to Support Community Health Surveillance

อนุชา เรืองศิริวัฒนกุล, ธิติพร เทียรณนิธิกุล, ธนากร ธนวัฒน์, ชนิตา เรืองศิริวัฒนกุล*

Anucha Ruangsiriwattanakul, Thitiporn Tienkanitikul, Tanakorn Tanawat
and Chanida Ruangsiriwattanakul*

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏอุดรดิตถ์

Faculty of Science and Technology, Uttaradit Rajabhat University

E-mail: anucha@uru.ac.th, thitiporn@uru.ac.th, tanakorn@uru.ac.th and chanida@uru.ac.th*

*Corresponding author

(Received: 20 January 2026, Revised: 3 June 2026, Accepted: 9 June 2026)

<https://doi.org/10.57260/stc.2026.1456>

บทคัดย่อ

ในปี พ.ศ. 2568 ประเทศไทยได้เข้าสู่สังคมสูงวัยโดยสมบูรณ์ (Aged society) โดยมีประชากรอายุ 60 ปีขึ้นไปในสัดส่วนมากกว่าร้อยละ 20 ของประชากรทั้งประเทศ ส่งผลให้โรคไม่ติดต่อเรื้อรัง (Non-communicable diseases: NCDs) เช่น โรคเบาหวานและโรคความดันโลหิตสูง กลายเป็นปัญหาสำคัญด้านสาธารณสุข โดยเฉพาะในพื้นที่ชุมชนชนบทซึ่งมีข้อจำกัดด้านระบบสนับสนุนการตัดสินใจเชิงข้อมูล งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ โดยประยุกต์ใช้เทคนิคเหมืองข้อมูลและการเรียนรู้ของเครื่องจากข้อมูลสุขภาพระดับชุมชน กลุ่มตัวอย่างประกอบด้วยข้อมูลสุขภาพของผู้สูงอายุจำนวน 3,419 ราย ที่ได้รับการคัดกรองสุขภาพในหน่วยบริการสุขภาพระดับปฐมภูมิในพื้นที่อำเภอลับแล จังหวัดอุดรดิตถ์ ระหว่างปี พ.ศ. 2565–2568 ร่วมกับข้อมูลสนับสนุนจากระบบข้อมูลสุขภาพของกระทรวงสาธารณสุข ข้อมูลถูกนำมาผ่านกระบวนการเตรียมข้อมูลและวิเคราะห์ตามกรอบ CRISP-DM โดยใช้แบบจำลอง Random forest และ Extreme gradient boosting (XGBoost) และประเมินประสิทธิภาพด้วย Accuracy, Precision, Recall, F1-score และ Confusion matrix

ผลการศึกษาพบว่าแบบจำลอง XGBoost มีความเหมาะสมสูงกว่าสำหรับงานคัดกรองสาธารณสุข XGBoost โดยสามารถตรวจจับกลุ่มเสี่ยง (Sensitivity) ได้ดีกว่า โดยมีค่าในการตรวจจับกลุ่มเสี่ยงโรคเบาหวานร้อยละ 25.00 และโรคความดันโลหิตสูงร้อยละ 16.00 สูงกว่า Random Forest ที่ได้ร้อยละ

17.30 และ 4.00 ตามลำดับ แม้ Random Forest จะมีค่า Accuracy รวมสูงกว่า (ร้อยละ 91.30 และ 91.70 ตามลำดับ) แต่ XGBoost สามารถลด False negative ได้ดีกว่า ซึ่งมีความสำคัญอย่างยิ่งในบริบทของการคัดกรองสุขภาพเชิงป้องกัน ปัจจัยสุขภาพที่เกี่ยวข้องกับระดับความเสี่ยง ได้แก่ ดัชนีมวลกาย รอบเอว ระดับน้ำตาลในเลือด ความดันโลหิต และอายุ การประยุกต์ใช้เทคนิค Machine Learning ร่วมกับข้อมูลสุขภาพระดับชุมชนสามารถสนับสนุนการคัดกรองความเสี่ยงเบื้องต้น การเฝ้าระวังสุขภาพเชิงรุก และการวางแผนดูแลสุขภาพผู้สูงอายุในระดับพื้นที่ได้อย่างมีประสิทธิภาพ ทั้งนี้แบบจำลองดังกล่าวมีวัตถุประสงค์เพื่อใช้สนับสนุนการตัดสินใจเบื้องต้นเท่านั้น มิใช่เพื่อการวินิจฉัยทางการแพทย์

คำสำคัญ: การคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรัง โรคเบาหวาน โรคความดันโลหิตสูง ผู้สูงอายุ การเรียนรู้ของเครื่อง

Abstract

In 2025, Thailand entered an Aged Society, with adults aged 60 years and older comprising over 20% of the total population, resulting in a growing burden of non-communicable diseases (NCDs), particularly diabetes mellitus and hypertension, which have become major public health concerns, especially in community settings with limited resources and data-driven decision-support tools. This study aimed to develop a risk screening model for non-communicable diseases among older adults using data mining and machine learning techniques applied to community health data. The study analyzed health screening records from 3,419 older adults in Laplae District, Uttaradit Province, Thailand, collected between 2022 and 2025. Data were processed following the CRISP-DM framework. Random Forest and Extreme Gradient Boosting (XGBoost) models were developed and evaluated using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

The results demonstrated that XGBoost was more appropriate for public health screening purposes, achieving a Sensitivity of 25.0% for diabetes risk detection and 16.0% for hypertension, compared to 17.3% and 4.0% for Random Forest, respectively. Although Random Forest yielded higher overall accuracy (91.3% and 91.7%), XGBoost produced fewer false negatives, which is critically important in preventive health screening contexts. Key risk-associated factors included body mass index, waist circumference, blood glucose level, blood pressure, and age. The findings suggest that integrating machine learning with community-level health data can effectively support preliminary risk screening, proactive health surveillance,

and local healthcare planning for older adults. The model is intended as a decision-support tool only, and is not a substitute for clinical diagnosis.

Keywords: Non-communicable disease risk screening, Diabetes mellitus, Hypertension, Older adults, Machine learning

บทนำ

ในปี 2568 ประเทศไทยถูกคาดการณ์ว่าจะมีผู้สูงอายุรวมประมาณ 13.80 - 14.80 ล้านคน คิดเป็นสัดส่วนประมาณ 20.70% - 21.00% ของประชากรทั้งประเทศ ซึ่งถือว่าประเทศไทยได้เข้าสู่สังคมสูงวัยอย่างสมบูรณ์ (สำนักงานสถิติแห่งชาติ, 2567) จากแนวโน้มโครงสร้างประชากรที่มีสัดส่วนผู้สูงอายุเพิ่มขึ้นอย่างรวดเร็วเป็นประเด็นสำคัญที่ส่งผลกระทบต่อระบบสุขภาพของประเทศไทยในระยะยาว ปัจจุบันประชากรอายุ 60 ปีขึ้นไปมีจำนวนเพิ่มขึ้นอย่างต่อเนื่อง การเปลี่ยนแปลงดังกล่าวมีความสัมพันธ์โดยตรงกับการเพิ่มขึ้นของโรคไม่ติดต่อเรื้อรัง (Non-communicable diseases: NCDs) เช่น โรคเบาหวาน โรคความดันโลหิตสูง และโรคหัวใจ ทวีความรุนแรงขึ้นอย่างต่อเนื่อง ซึ่งเป็นสาเหตุสำคัญของการเจ็บป่วยเรื้อรัง ภาวะแทรกซ้อน และการเสียชีวิตก่อนวัยอันควร (World health organization, 2014) ในบริบทของชุมชนชนบทปัญหาโรคไม่ติดต่อเรื้อรังในผู้สูงอายุมีความท้าทายมากยิ่งขึ้น เนื่องจากข้อจำกัดด้านการเข้าถึงบริการสุขภาพเฉพาะทาง ระบบติดตามดูแลสุขภาพอย่างต่อเนื่อง และการประยุกต์ใช้เทคโนโลยีสารสนเทศเพื่อการดูแลสุขภาพเชิงรุก แม้ว่าหน่วยบริการสุขภาพระดับปฐมภูมิ เช่น โรงพยาบาลส่งเสริมสุขภาพตำบลจะมีบทบาทสำคัญในการคัดกรองและติดตามสุขภาพของประชาชน อย่างไรก็ตามการใช้ข้อมูลสุขภาพในระดับพื้นที่ส่วนใหญ่มักอยู่ในรูปแบบของการรายงานเชิงสถิติพื้นฐาน และยังขาดเครื่องมือวิเคราะห์ข้อมูลเพื่อช่วยสนับสนุนการคัดกรองความเสี่ยงและการตัดสินใจเชิงรุกสำหรับบุคลากรสาธารณสุข (กรมควบคุมโรค, 2566)

ในช่วงหลายปีที่ผ่านมา เทคโนโลยีเหมืองข้อมูล (Data mining) และการเรียนรู้ของเครื่อง (Machine learning) ได้รับการนำมาประยุกต์ใช้อย่างแพร่หลายในงานด้านสุขภาพ เนื่องจากสามารถวิเคราะห์ข้อมูลขนาดใหญ่ ค้นหารูปแบบความสัมพันธ์ของตัวแปร และสนับสนุนการประเมินความเสี่ยงด้านสุขภาพได้อย่างมีประสิทธิภาพ (Han et al., 2011; Chen & Guestrin, 2016) จากการศึกษาวิจัยที่เกี่ยวข้องหลายฉบับพบว่า อัลกอริทึม เช่น Logistic regression, Decision tree, Random forest, Support vector machine และ XGBoost มีศักยภาพในการจำแนกกลุ่มเสี่ยงโรคเรื้อรังได้ในระดับที่น่าพอใจ อย่างไรก็ตามงานวิจัยส่วนใหญ่มักใช้ข้อมูลจากโรงพยาบาลระดับทุติยภูมิหรือฐานข้อมูลระดับประเทศ ซึ่งอาจมีบริบทแตกต่างจากข้อมูลสุขภาพระดับชุมชน โดยเฉพาะในพื้นที่ชนบทที่มีข้อจำกัดด้านทรัพยากรและระบบสนับสนุนการตัดสินใจ

งานวิจัยนี้ใช้แนวคิดด้าน Machine learning เพื่อพัฒนาแบบจำลองคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ โดยใช้ข้อมูลสุขภาพระดับชุมชนจากหน่วยบริการสุขภาพปฐมภูมิร่วมกับเทคนิค Machine learning เพื่อสนับสนุนการคัดกรองความเสี่ยงเบื้องต้น การเฝ้าระวังสุขภาพเชิงรุก และการวางแผนดูแล

สุขภาพผู้สูงอายุในระดับพื้นที่ โดยใช้เทคนิคป่าสุ่ม (Random forest) และเทคนิคเอ็กซ์ตรีมเกรเดียนต์บูสติง (eXtreme gradient boosting) และเปรียบเทียบประสิทธิภาพการคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ และประยุกต์ใช้แบบจำลองในการสนับสนุนการคัดกรองสุขภาพเชิงรุกและการเฝ้าระวังสุขภาพในระดับชุมชน ทั้งนี้แบบจำลองดังกล่าวมีวัตถุประสงค์เพื่อเป็นเครื่องมือสนับสนุนการตัดสินใจเบื้องต้น มีใช้เครื่องมือสำหรับการวินิจฉัยทางการแพทย์

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาความสัมพันธ์ของปัจจัยสุขภาพที่เกี่ยวข้องกับความเสี่ยงของโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ
2. เพื่อพัฒนาและเปรียบเทียบแบบจำลองการคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ โดยใช้เทคนิคเหมืองข้อมูล
3. เพื่อประยุกต์ใช้แบบจำลองในการสนับสนุนการคัดกรองสุขภาพเชิงรุกและการเฝ้าระวังสุขภาพในระดับชุมชน

ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงปริมาณ (Quantitative research) โดยประยุกต์ใช้กระบวนการเหมืองข้อมูล (Data mining) ตามกรอบแนวคิด CRISP-DM (Cross industry standard process for data mining) ประกอบด้วย การทำความเข้าใจปัญหา การทำความเข้าใจข้อมูล การเตรียมข้อมูล การสร้างแบบจำลอง การประเมินผล และการประยุกต์ใช้ผลลัพธ์ (Wirth & Hipp, 2000)

ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการวิจัยคือผู้สูงอายุอายุ 60 ปีขึ้นไป ที่เข้ารับบริการคัดกรองสุขภาพในหน่วยบริการสุขภาพระดับปฐมภูมิในพื้นที่อำเภอลับแล จังหวัดอุตรดิตถ์ กลุ่มตัวอย่างเป็นข้อมูลทุติยภูมิจากระบบคัดกรองสุขภาพผู้สูงอายุ จำนวน 3,419 ระเบียน ระหว่างเดือนตุลาคม พ.ศ. 2565 ถึงกุมภาพันธ์ พ.ศ. 2568

การรับรองจริยธรรมการวิจัย

งานวิจัยนี้ได้รับการรับรองจริยธรรมการวิจัยในมนุษย์จากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยราชภัฏอุตรดิตถ์ เลขที่ URU-REC No. 051/2567 (COE No. 008/2024) วันที่ 18 ธันวาคม 2567 ข้อมูลที่นำมาใช้ผ่านกระบวนการผ่านการปกปิดตัวตนแล้ว เพื่อคุ้มครองความเป็นส่วนตัวของผู้ให้ข้อมูล

เครื่องมือที่ใช้ในการวิจัย

การวิจัยครั้งนี้ใช้เครื่องมือสำหรับการประมวลผล วิเคราะห์ข้อมูลเชิงสถิติ และการเรียนรู้ของเครื่อง ได้แก่ โปรแกรมภาษา Python และ Google colab เพื่อสร้างแบบจำลองที่เหมาะสมในการวิเคราะห์ปัจจัยเสี่ยงของโรคไม่ติดต่อเรื้อรังด้วยเทคนิคเหมืองข้อมูล โดยใช้เทคนิคป่าสุ่ม (Random forest) และเทคนิคเอ็กซ์ตรีมเกรเดียนต์บูสติง (eXtreme gradient boosting : XGBoost)

การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้ในการวิจัยนี้ใช้ข้อมูลสุขภาพจากหน่วยงานที่เกี่ยวข้อง ประกอบด้วย 2 ส่วน ได้แก่ 1) ข้อมูลคัดกรองจากโรงพยาบาลส่งเสริมสุขภาพตำบล ในพื้นที่อำเภอลับแล จังหวัดอุตรดิตถ์ โดยใช้ข้อมูลผู้ที่มีอายุ 60 ปีขึ้นไปที่ได้รับการคัดกรองโรคไม่ติดต่อเรื้อรัง ตั้งแต่เดือนตุลาคม พ.ศ. 2565 ถึง เดือนกุมภาพันธ์ พ.ศ. 2568 จำนวน 3,419 ระเบียน ตัวแปรอิสระที่ใช้ในการศึกษา ได้แก่ เพศ อายุ น้ำหนัก ส่วนสูง ดัชนีมวลกาย (BMI) รอบเอว ระดับน้ำตาลในเลือด ความดันโลหิตตัวบน (SBP) ความดันโลหิตตัวล่าง (DBP) และตัวแปรที่สร้างขึ้นจาก Feature engineering โดยตัวแปรตาม คือระดับความเสี่ยงจากการคัดกรองสุขภาพเบื้องต้น ประกอบด้วย ความเสี่ยงโรคเบาหวาน ความเสี่ยงโรคความดันโลหิตสูง (0 = ไม่มีความเสี่ยง, 1 = มีความเสี่ยง) โดยกำหนดตามเกณฑ์คัดกรองของกระทรวงสาธารณสุข และ 2) ข้อมูลจาก MOPH open data โดยดึงข้อมูลผ่าน API ของกระทรวงสาธารณสุข เพื่อวิเคราะห์แนวโน้มอัตราป่วยโรคเบาหวานและความดันโลหิตสูงตามช่วงอายุและเพศในระดับจังหวัด เพื่อสนับสนุนการทำ Feature engineering และการตีความผลลัพธ์ในเชิงบริบท

วิธีดำเนินการวิจัย

การดำเนินการวิจัยเริ่มจากการนำเข้าสู่ข้อมูลสุขภาพผู้สูงอายุจากหน่วยบริการสุขภาพระดับปฐมภูมิ มีขั้นตอนดังนี้

1. การเตรียมข้อมูลโดยแบ่งออกเป็นสองขั้นตอนหลัก ได้แก่

1.1 การทำความสะอาดข้อมูล (Data cleaning) โดยตรวจสอบข้อมูลสูญหาย ลบข้อมูลซ้ำ ตรวจสอบค่าผิดปกติ (Outliers) และการแยกค่าความดันโลหิตตัวบนและตัวล่าง

การแปลงข้อมูล (Data transformation) ได้แก่ การเข้ารหัสตัวแปรเชิงหมวดหมู่ หารูปแบบข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ เช่น เพศ การจัดกลุ่มค่าผลคัดกรองโรค การปรับสเกลข้อมูล และสร้างตัวแปรใหม่จาก Feature engineering เช่น กลุ่มอายุ กลุ่ม BMI และระดับความเสี่ยงจากตัวชี้วัดสุขภาพ

2. การวิเคราะห์ความสัมพันธ์ของตัวแปร เพื่อศึกษาความสัมพันธ์เบื้องต้นระหว่างตัวแปรสุขภาพและกลุ่มความเสี่ยง ผู้วิจัยใช้การวิเคราะห์ความสัมพันธ์ทางสถิติที่เหมาะสมกับชนิดของข้อมูล โดยรายงานค่าสัมประสิทธิ์ความสัมพันธ์และค่า p-value เพื่อสนับสนุนการคัดเลือกตัวแปรสำหรับการสร้างแบบจำลอง นอกจากนี้ได้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระเพื่อตรวจสอบความสัมพันธ์ร่วมกันของข้อมูลก่อนการสร้างแบบจำลอง

3. การสร้างแบบจำลองและการประเมินประสิทธิภาพ

หลังจากได้ชุดข้อมูลผ่านการเตรียมข้อมูลและคัดเลือกคุณลักษณะแล้ว ผู้วิจัยได้เลือกใช้แบบจำลองพยากรณ์โดยใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ Random forest และ Extreme gradient boosting (XGBoost) เนื่องจากเป็นอัลกอริทึมที่มีประสิทธิภาพในการจำแนกข้อมูลสุขภาพที่มีความสัมพันธ์ไม่เชิงเส้น

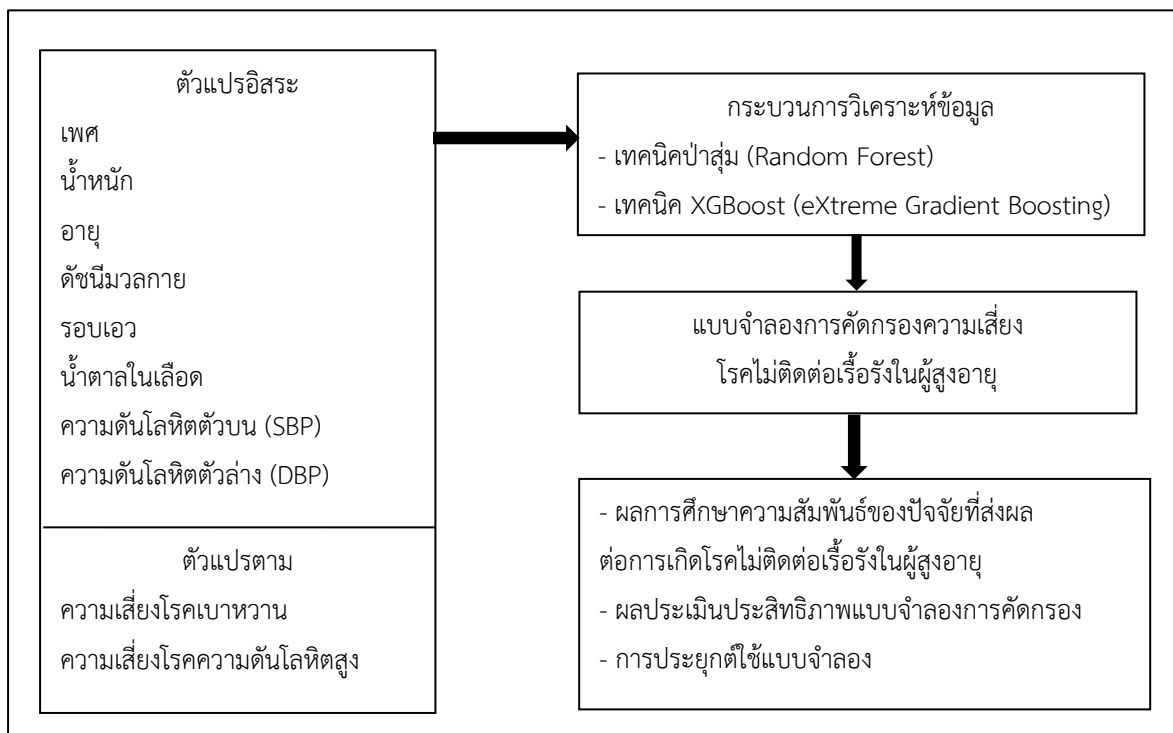
และสามารถจัดการข้อมูลที่มีตัวแปรหลายมิติได้ดี โดยแบ่งข้อมูลเป็นชุดฝึกสอน (Training set) ร้อยละ 80 และชุดทดสอบ (Testing set) ร้อยละ 20

การประเมินประสิทธิภาพของแบบจำลองประเมินด้วย Accuracy precision recall F1-score และ Confusion matrix เพื่อประเมินความสามารถในการคัดกรองกลุ่มเสี่ยงและไม่เสี่ยง (Powers, 2011) เพื่อใช้ในการเปรียบเทียบประสิทธิภาพในการคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ

4. การประยุกต์ใช้และการถ่ายทอดองค์ความรู้สู่ชุมชน

ผลลัพธ์จากแบบจำลองที่มีประสิทธิภาพสูงที่สุดนำไปประยุกต์ใช้ในรูปแบบแอปพลิเคชันบนเว็บ และจัดกิจกรรมคืนข้อมูลในชุมชนเพื่อส่งเสริมพฤติกรรมสุขภาพ เพื่อให้ผู้สูงอายุและบุคลากรสาธารณสุขสามารถประเมินความเสี่ยงโรคเบาหวานและความดันโลหิตสูงได้ด้วยตนเอง โดยแบบจำลองนี้มีวัตถุประสงค์เพื่อใช้สนับสนุนการคัดกรองสุขภาพเบื้องต้นในระดับชุมชนเท่านั้นมิใช่เพื่อการวินิจฉัยทางการแพทย์ ผู้วิจัยได้ตระหนักถึงประเด็นด้านความเป็นส่วนตัวของข้อมูล และข้อจำกัดด้านการนำไปใช้ในพื้นที่อื่น

การออกแบบระบบ Front-end จะให้ผู้กรอกข้อมูลเบื้องต้น ได้แก่ น้ำหนัก ส่วนสูง ความดัน อายุ สำหรับประเมินความเสี่ยง นอกจากนี้ยังจัดเวทีคืนข้อมูลและกิจกรรมส่งเสริมสุขภาพในชุมชน เพื่อถ่ายทอดองค์ความรู้จากงานวิจัยสู่การปรับเปลี่ยนพฤติกรรมสุขภาพอย่างเป็นรูปธรรม โดยกรอบแนวคิดการวิจัย แสดงดังภาพที่ 1



ภาพที่ 1 กรอบแนวคิดการวิจัย

ภาพที่ 1 แสดงกรอบแนวคิดการวิจัยโดยมีกระบวนการพัฒนาแบบจำลอง Machine Learning ตั้งแต่การเตรียมข้อมูลสุขภาพพระระดับชุมชน การคัดเลือกคุณลักษณะเพื่อใช้เป็นตัวแปรในการวิเคราะห์ข้อมูลเพื่อนำไปสร้างและประเมินประสิทธิภาพของแบบจำลองพยากรณ์ ไปจนถึงการนำผลลัพธ์ไปใช้ในการสนับสนุนการดูแลสุขภาพผู้สูงอายุในระดับพื้นที่

ผลการวิจัย

การศึกษานี้ใช้ข้อมูลสุขภาพของกลุ่มตัวอย่าง จำนวน 3,419 ราย จากหน่วยบริการสุขภาพระดับปฐมภูมิในพื้นที่อำเภอลับแล จังหวัดอุตรดิตถ์ ระหว่างปี พ.ศ. 2565–2568 เพื่อพัฒนาแบบจำลองคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรัง โดยนำเสนอผลการศึกษา 4 ส่วน ได้แก่ 1) ลักษณะทั่วไปของกลุ่มตัวอย่าง 2) การวิเคราะห์ความสัมพันธ์ของตัวแปรที่ส่งผลต่อโรคไม่ติดต่อเรื้อรัง 3) ผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง และ 4) การประยุกต์ใช้แบบจำลองเพื่อสนับสนุนการคัดกรองสุขภาพ

1. ลักษณะทั่วไปของกลุ่มตัวอย่าง แสดงดังตารางที่ 1

ตารางที่ 1 ลักษณะทั่วไปของกลุ่มตัวอย่าง (n = 3,419)

ตัวแปร	Mean $\pm$ SD	Min	Max	หมายเหตุ
อายุ (ปี)	62.96 $\pm$ 13.73	60	99	
BMI (kg/m ²)	23.47 $\pm$ 4.80	11.70	96.4	0
รอบเอว (cm)	80.79 $\pm$ 9.09	55	175	ตัดค่า outlier >200 ออก
SBP (mmHg)	118.37 $\pm$ 12.92	83	239	
DBP (mmHg)	75.93 $\pm$ 6.70	45	123	
น้ำตาลในเลือด (mg/dL)	93.07 $\pm$ 10.22	35	202	
มีภาวะเสี่ยงเบาหวาน (n, %)	262 (7.8%)			เสี่ยง + สงสัยป่วย
มีภาวะเสี่ยงความดัน (n, %)	252 (7.4%)			เสี่ยง + สงสัยป่วย

ผลการวิเคราะห์ลักษณะทั่วไปของกลุ่มตัวอย่างพบว่า ผู้มีอายุเฉลี่ย 63 ปี โดยตัวแปรด้านสุขภาพที่ใช้ในการศึกษา ได้แก่ ดัชนีมวลกาย รอบเอว ค่าความดันโลหิต และระดับน้ำตาลในเลือด

2. ผลการศึกษาความสัมพันธ์ตัวแปรเพื่อศึกษาความสัมพันธ์เบื้องต้นของปัจจัยสุขภาพกับระดับความเสี่ยง

ผู้วิจัยได้วิเคราะห์ความสัมพันธ์ทางสถิติของตัวแปรที่เกี่ยวข้องกับระดับความเสี่ยง ได้แก่ อายุ ดัชนีมวลกาย รอบเอว ระดับน้ำตาลในเลือด ความดันโลหิต โดยการวิเคราะห์ค่าสหสัมพันธ์ (Correlation) ระหว่างตัวแปรกับภาวะเสี่ยงโรคเบาหวาน และโรคความดันโลหิตสูง และการวิเคราะห์ตัวแปรที่มีอิทธิพลสูงสุด

(Feature importance) ต่อการสร้างแบบจำลองคัดกรองความเสี่ยงโรคเบาหวาน และความดันโลหิตสูง โดยผลการวิเคราะห์ความสัมพันธ์ของตัวแปร แสดงดังตารางที่ 2

ตารางที่ 2 ค่าสหสัมพันธ์ของตัวแปรที่มีผลต่อความเสี่ยงโรคเบาหวาน (DM)

ลำดับ	Feature	Correlation	p-value	ระดับความสัมพันธ์
1	ระดับน้ำตาลในเลือด (BSL)	0.601	< .001	สูงมาก
2	ความดันตัวบน (Systolic)	0.251	< .001	ต่ำ-ปานกลาง
3	ความดันตัวล่าง (Diastolic)	0.136	< .001	ต่ำ
4	รอบเอว (Waist)	0.064	< .001	ต่ำ
5	ดัชนีมวลกาย (BMI)	0.085	< .001	ต่ำ
6	น้ำหนัก (Weight)	0.085	< .001	ต่ำ
7	ส่วนสูง (Height)	0.000	.995	ไม่มีนัยสำคัญ
8	อายุ (Age)	-0.098	< .001	ต่ำ (ติดลบ)*

จากตารางที่ 2 ผลการวิเคราะห์ค่าสหสัมพันธ์แสดงให้เห็นว่าปัจจัยที่สัมพันธ์กับภาวะเสี่ยงโรคเบาหวานอย่างชัดเจนที่สุดคือระดับน้ำตาลในเลือด ซึ่งมีค่าสหสัมพันธ์สูงสุด รองลงมาคือค่าความดันโลหิต โดยเฉพาะความดันตัวบน ขณะที่ตัวแปร รอบเอว ดัชนีมวลกาย และน้ำหนัก มีความสัมพันธ์ในระดับรอง ส่วนอายุและส่วนสูงไม่พบความสัมพันธ์ที่เด่นชัดกับภาวะเสี่ยงโรคเบาหวานในกลุ่มตัวอย่างนี้

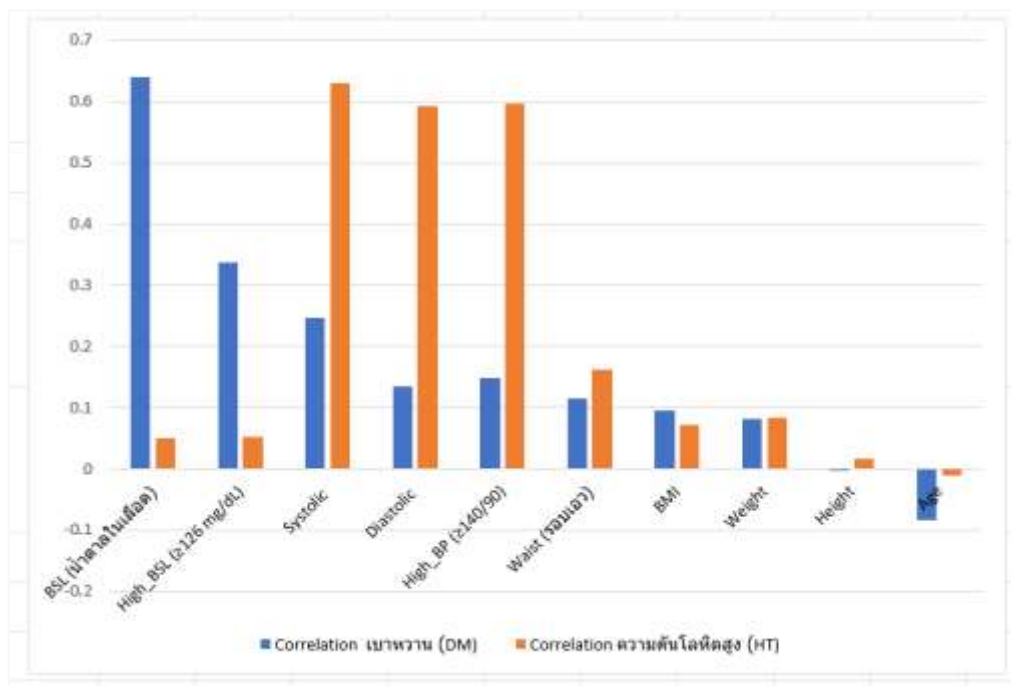
การวิเคราะห์ค่าสหสัมพันธ์ (Correlation) ของตัวแปรกับภาวะเสี่ยงโรคความดันโลหิตสูง เพื่อนำไปใช้ในการสร้างโมเดลคัดกรองความเสี่ยง แสดงดังตารางที่ 3

ตารางที่ 3 ค่าสหสัมพันธ์ของตัวแปรที่มีผลต่อความเสี่ยงโรคความดันโลหิตสูง (HT)

ลำดับ	Feature	Correlation	p-value	ระดับความสัมพันธ์
1	ความดันตัวบน (Systolic)	0.629	< .001	สูงมาก
2	ความดันตัวล่าง (Diastolic)	0.593	< .001	สูงมาก
3	น้ำหนัก (Weight)	0.098	< .001	ต่ำ
4	รอบเอว (Waist)	0.095	< .001	ต่ำ
5	ดัชนีมวลกาย (BMI)	0.067	< .001	ต่ำ
6	ระดับน้ำตาลในเลือด (BSL)	0.068	< .001	ต่ำ
7	ส่วนสูง (Height)	0.019	.268	ไม่มีนัยสำคัญ
8	อายุ (Age)	-0.034	.049	ต่ำ (ติดลบ)*

จากตารางที่ 3 การวิเคราะห์ค่าสหสัมพันธ์ของตัวแปรที่มีผลต่อการพยากรณ์โรคความดันโลหิตสูง ผลการวิเคราะห์พบว่าค่าความดันโลหิตตัวบนและตัวล่างเป็นตัวแปรที่มีความสัมพันธ์สูงที่สุด ขณะที่ตัวแปร น้ำหนัก รอบเอว ดัชนีมวลกาย และระดับน้ำตาลในเลือด มีความสัมพันธ์ในระดับต่ำ ผลลัพธ์ดังกล่าวสะท้อนให้เห็นว่าปัจจัยดังกล่าวมีความสอดคล้องกับหลักเกณฑ์ทางการแพทย์และแนวทางการคัดกรองของกระทรวงสาธารณสุข โดยค่าสหสัมพันธ์แสดงเพื่อวิเคราะห์ความสัมพันธ์เบื้องต้น โดยการสร้างแบบจำลองใช้ตัวแปรที่ไม่ใช่เกณฑ์วินิจฉัยโดยตรง

ภาพที่ 2 แสดงแผนภูมิเปรียบเทียบตัวแปรสำคัญที่วิเคราะห์จากค่าสหสัมพันธ์ (Correlation) กับระดับความเสี่ยงโรคเบาหวานและโรคความดันโลหิตสูง



ภาพที่ 2 แผนภูมิเปรียบเทียบตัวแปรสำคัญที่วิเคราะห์จากค่าสหสัมพันธ์ (Correlation) กับระดับความเสี่ยงโรคเบาหวานและโรคความดันโลหิตสูง

3. ผลการสร้างและประเมินประสิทธิภาพของแบบจำลอง

หลังจากการเตรียมข้อมูลและคัดเลือกคุณลักษณะแล้ว ข้อมูลถูกแบ่งออกเป็นชุดฝึกสอนและชุดทดสอบในอัตราส่วนประมาณ 80:20 จากนั้นดำเนินการสร้างแบบจำลองด้วยเทคนิค Random forest และ XGBoost โดยการประเมินประสิทธิภาพของแบบจำลองคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ โดยใช้ Accuracy precision recall และ F1-score เพื่อวัดประสิทธิภาพของแบบจำลองในการจำแนกกลุ่มผู้มีความเสี่ยงและไม่มีความเสี่ยงได้อย่างถูกต้อง โดยทำการเปรียบเทียบระหว่างแบบจำลอง Random forest และ Extreme gradient boosting (XGBoost) เพื่อเปรียบเทียบประสิทธิภาพการคัดกรองความเสี่ยงโรคเบาหวาน และโรคความดันโลหิตสูง

ผลการเปรียบเทียบประสิทธิภาพการคัดกรองความเสี่ยงโรคเบาหวาน นำเสนอโดยใช้ตาราง Confusion matrix แสดงรายละเอียดของผลการจำแนกกลุ่มเสี่ยงและไม่เสี่ยงของแบบจำลองที่ให้ผลดีที่สุด ดังตารางที่ 4

ตารางที่ 4 Confusion Matrix โรคเบาหวาน (DM) | n_test = 665

	XGBoost		Random Forest	
	Pred ไม่เสี่ยง	Pred เสี่ยง	Pred ไม่เสี่ยง	Pred เสี่ยง
จริง ไม่เสี่ยง (0)	TN = 583	FP = 30	TN = 598	FP = 15
จริง เสี่ยง (1)	FN = 39	TP = 13	FN = 43	TP = 9

โดยผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์โรคเบาหวาน แสดงดังตารางที่ 5

ตารางที่ 5 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการคัดกรองความเสี่ยงโรคเบาหวาน

แบบจำลอง	Accuracy	Precision	Recall	F1-score	Sensitivity (กลุ่มเสี่ยง)
XGBoost	89.6%	0.888	0.896	0.892	25.0%
Random Forest	91.3%	0.889	0.913	0.898	17.3%

จากตารางที่ 4 และตารางที่ 5 พบว่าแบบจำลอง XGBoost สามารถตรวจจับกลุ่มเสี่ยงโรคเบาหวานได้ 25% (TP=13/52) ส่วนแบบจำลอง Random Forest สามารถตรวจจับกลุ่มเสี่ยงโรคเบาหวานได้ 17% (TP=9/52)

ในส่วนของการเปรียบเทียบประสิทธิภาพการคัดกรองความเสี่ยงโรคความดันโลหิตสูง นำเสนอโดยใช้ตาราง Confusion Matrix แสดงรายละเอียดของผลการจำแนกกลุ่มเสี่ยงและไม่เสี่ยงของแบบจำลองที่ให้ผลดีที่สุด ดังตารางที่ 6

ตารางที่ 6 Confusion Matrix โรคความดันโลหิตสูง (HT) | n_test = 684

	XGBoost		Random Forest	
	Pred ไม่เสี่ยง	Pred เสี่ยง	Pred ไม่เสี่ยง	Pred เสี่ยง
จริง ไม่เสี่ยง (0)	TN = 603	FP = 31	TN = 625	FP = 9
จริง เสี่ยง (1)	FN = 42	TP = 8	FN = 48	TP = 2

โดยผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์โรคความดันโลหิตสูง แสดงดังตารางที่ 7

ตารางที่ 7 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการคัดกรองความเสี่ยงโรคความดันโลหิตสูง

แบบจำลอง	Accuracy	Precision	Recall	F1-score	Sensitivity (กลุ่มเสี่ยง)
XGBoost	89.3%	0.882	0.893	0.887	16.0%
Random Forest	91.7%	0.874	0.917	0.891	4.0%

จากตารางที่ 6 และตารางที่ 7 พบว่าแบบจำลอง XGBoost สามารถตรวจจับกลุ่มเสี่ยงโรคความดันโลหิตสูงได้ 16% (TP=8/50) ส่วนแบบจำลอง Random forest สามารถตรวจจับกลุ่มเสี่ยงโรคความดันโลหิตสูงได้ 4% (TP=2/50)

ถึงแม้ว่าแบบจำลอง Random forest จะมีค่า Accuracy รวมสูงกว่า แต่ XGBoost สามารถตรวจจับกลุ่มเสี่ยงได้มากกว่า แต่หากพิจารณาค่า Recall ของกลุ่มเสี่ยง ซึ่งมีความสำคัญในบริบทสาธารณสุข โดยการแสดงค่า False negative สูงซึ่งหมายถึงคนเสี่ยงแต่พยากรณ์ว่าปกติ ย่อมมีผลกระทบต่องานสาธารณสุขสูง ดังนั้นสามารถสรุปได้ว่าแบบจำลอง XGBoost มี Sensitivity สูงกว่า สามารถตรวจจับกลุ่มเสี่ยงได้มากกว่า และลด False negative ได้ดีกว่า จึงมีความเหมาะสมสำหรับงานคัดกรองสาธารณสุขมากกว่า Random forest แม้ Accuracy รวมจะต่ำกว่าเล็กน้อยก็ตาม

ในส่วนของการใช้ข้อมูล MOPH data center เพื่อสนับสนุนการวิเคราะห์ โดยนำข้อมูลแนวโน้มย้อนหลังของโรคเบาหวานและโรคความดันโลหิตสูงจากระบบ MOPH data center API มาใช้สนับสนุนการวิเคราะห์ โดยข้อมูลดังกล่าวช่วยแสดงให้เห็นแนวโน้มการเพิ่มขึ้นของอัตราป่วยในกลุ่มผู้สูงอายุ และถูกนำมาใช้ในกระบวนการ Feature engineering เพื่อสร้างตัวแปรความเสี่ยงเฉลี่ยตามช่วงอายุและเพศ การผนวกข้อมูลระดับจังหวัดเข้ากับข้อมูลระดับตำบลช่วยเพิ่มความแม่นยำและความสอดคล้องเชิงบริบทของแบบจำลอง

4. การประยุกต์ใช้แบบจำลองเพื่อสนับสนุนการคัดกรองสุขภาพ

ผลการดำเนินงานวิจัยถูกนำไปประยุกต์ใช้ในเชิงปฏิบัติ โดยการจัดกิจกรรมคืนข้อมูลและส่งเสริมสุขภาพแก่ผู้สูงอายุในชุมชน ผ่านกิจกรรมด้านโภชนาการและการออกกำลังกาย แสดงดังภาพที่ 3



ภาพที่ 3 การจัดกิจกรรมคืนข้อมูลและส่งเสริมสุขภาพแก่ผู้สูงอายุในชุมชน ผ่านกิจกรรมด้านโภชนาการและการออกกำลังกาย

นอกจากนี้ยังพัฒนาแอปพลิเคชันประเมินความเสี่ยงโรคเบาหวานและความดันโลหิตสูง โดยประยุกต์จากแบบจำลอง XGBoost ที่สามารถใช้งานได้บนเว็บและอุปกรณ์สื่อสารเคลื่อนที่ สามารถใช้งานได้ที่ <https://utdnchs.streamlit.app/> แสดงดังภาพที่ 4-5 โดยผลการประเมินความพึงพอใจของผู้เข้าร่วมกิจกรรมโดยประเมินด้านเนื้อหาและด้านการใช้งาน อยู่ในระดับมาก มีค่าเฉลี่ยเท่ากับ 4.49



ภาพที่ 4 หน้าจอหลักของแอปพลิเคชันเพื่อใช้คัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรัง

The screenshot shows the "Risk Assessment Form" in detail. It includes input fields for:

- อายุ (ปี): 50
- เพศ: ชาย
- ค่าดัชนีมวลกาย (BMI): 25.00
- รอบเอว (cm): 85
- ค่าความดันโลหิตตัวบน (Systolic): 130
- ค่าความดันโลหิตตัวล่าง (Diastolic): 80

 The form also displays the "ระดับไขมันรวม (Cholesterol)" as 180. Below the input fields, there is a "ประเมินความเสี่ยง" (Assess Risk) button. The results section shows:

- ผลการประเมินความเสี่ยง: เสี่ยง
- ความน่าจะเป็นของแต่ละคลาส:
 - ปกติ: 0.26
 - เสี่ยง: 0.50
 - สงสัยป่วย: 0.25

 At the bottom, there is a warning icon and text: "ข้อมูลนี้ใช้เพื่อการประเมินเบื้องต้นเท่านั้น ควรพบแพทย์เพื่อการวินิจฉัย" (This information is for initial assessment only. Please consult a doctor for diagnosis).

ภาพที่ 5 หน้าจอส่วนรับข้อมูลสุขภาพจากผู้ใช้ เพื่อประเมินความเสี่ยงของโรคไม่ติดต่อเรื้อรัง

จากภาพที่ 5 ระบบแสดงผลการประเมินความเสี่ยงในรูปแบบความน่าจะเป็นของแต่ละกลุ่ม ได้แก่ ปกติ เสี่ยง และสงสัยป่วย โดยผลรวมของความน่าจะเป็นทั้งสามกลุ่มเท่ากับ 1.00 และแสดงกลุ่มที่มีค่าความน่าจะเป็นสูงสุดเป็นผลการพยากรณ์หลัก พร้อมคำแนะนำให้ผู้ใช้ปรึกษาบุคลากรทางการแพทย์ ทั้งนี้แอปพลิเคชันมีวัตถุประสงค์เพื่อใช้เป็นเครื่องมือเสริมการคัดกรองเชิงรุก ไม่ใช่การวินิจฉัยโรคแทนแพทย์

การอภิปรายผล

ผลการศึกษาพบว่าแม้ Random forest จะมีค่า Accuracy โดยรวมสูงกว่าแต่ XGBoost มีค่า Sensitivity สูงกว่า และสามารถลดจำนวน False negative ได้ดีกว่าจึงมีความเหมาะสมสำหรับการคัดกรองความเสี่ยงในบริบทสาธารณสุขที่ให้ความสำคัญกับการตรวจพบกลุ่มเสี่ยง โดยเฉพาะโรคเบาหวานและโรคความดันโลหิตสูง ผลการศึกษาแสดงให้เห็นว่า XGBoost สามารถจัดการข้อมูลสุขภาพที่มีความซับซ้อนและมีความสัมพันธ์หลายมิติได้ดีกว่า Random forest ซึ่งผลการวิจัยดังกล่าวสอดคล้องกับงานวิจัยของ Wang et al. (2020) ที่พบว่า XGBoost มีประสิทธิภาพสูงในการประเมินความเสี่ยงโรคเบาหวาน และ Deberneh & Kim (2021) พบว่า Machine Learning สามารถช่วยคัดกรองความเสี่ยงโรคเบาหวานได้อย่างมีประสิทธิภาพเมื่อใช้ข้อมูลสุขภาพเชิงคลินิก และสอดคล้องกับ Chen & Guestrin (2016) และ Hu et al. (2021) ที่กล่าวว่าเทคนิค Gradient boosting มีประสิทธิภาพสูงในการจัดการข้อมูลสุขภาพที่มีลักษณะไม่สมดุลของคลาส

ในส่วน of โรคความดันโลหิตสูง ผลการศึกษานี้มีความสอดคล้องกับ Islam et al. (2022) พบว่าเทคนิค Machine learning สามารถช่วยจำแนกกลุ่มเสี่ยงและระบุปัจจัยที่เกี่ยวข้องกับโรคความดันโลหิตสูงได้อย่างมีประสิทธิภาพ ปัจจัยที่เกี่ยวข้องกับระดับความเสี่ยง ได้แก่ อายุ ดัชนีมวลกาย รอบเอว ระดับน้ำตาลในเลือด และความดันโลหิต ซึ่งสอดคล้องกับองค์ความรู้ทางสาธารณสุขเกี่ยวกับโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ โดยตัวแปรเหล่านี้เป็นตัวชี้วัดสุขภาพพื้นฐานที่สามารถเก็บได้จากกระบวนการคัดกรองสุขภาพในระดับชุมชน ถึงแม้ว่าตัวแปรบางตัวจะมีความสัมพันธ์โดยตรงกับการประเมินความเสี่ยงทางคลินิก แต่งานวิจัยนี้มีวัตถุประสงค์เพื่อสนับสนุนการคัดกรองความเสี่ยงเบื้องต้นในการใช้งานจริงของระบบสุขภาพระดับชุมชน และสนับสนุนการคัดกรองสุขภาพเชิงรุกในระดับพื้นที่ ซึ่งยังมีการศึกษาค่อนข้างจำกัด ทั้งนี้แบบจำลองดังกล่าวมิได้มีวัตถุประสงค์เพื่อใช้เป็นเครื่องมือวินิจฉัยทางการแพทย์ แต่สามารถใช้เป็นเครื่องมือในการสนับสนุนการตัดสินใจสำหรับการเฝ้าระวังสุขภาพเชิงรุก และการจัดลำดับความสำคัญของกลุ่มเสี่ยงที่ควรได้รับการติดตาม

ข้อจำกัดของงานวิจัย

แม้ว่างานวิจัยนี้จะนำ Machine learning มาใช้ในการคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุแต่ยังมีข้อจำกัดบางประการได้แก่ ข้อมูลที่ใช้ในการศึกษามาจากพื้นที่เฉพาะในจังหวัดอุดรธานี ซึ่งอาจมีลักษณะประชากรและบริบทสุขภาพแตกต่างจากพื้นที่อื่น ดังนั้นการนำผลการศึกษาไปประยุกต์ใช้กับพื้นที่อื่นอาจต้องมีการปรับแบบจำลองเพิ่มเติม นอกจากนี้งานวิจัยนี้ใช้ข้อมูลทุติยภูมิจากระบบคัดกรองสุขภาพ ซึ่งอาจมีข้อจำกัดด้านความครบถ้วน ความถูกต้อง และความสม่ำเสมอของข้อมูลบางรายการ และงานวิจัยนี้ยังไม่ได้รวมปัจจัยด้านพฤติกรรมสุขภาพบางด้าน เช่น การบริโภคอาหาร การออกกำลังกาย การสูบบุหรี่ หรือประวัติครอบครัว ซึ่งอาจมีผลต่อความเสี่ยงของโรคไม่ติดต่อเรื้อรัง โดยการประยุกต์ใช้ Machine learning ในงานด้านสุขภาพจำเป็นต้องคำนึงถึงประเด็นด้านจริยธรรม ความเป็นส่วนตัวของข้อมูล ความเป็นส่วนตัวของอัลกอริทึม และความเสี่ยงจากอคติของข้อมูล

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้แสดงให้เห็นว่าการประยุกต์ใช้เทคนิคเหมืองข้อมูลและการเรียนรู้ของเครื่องกับข้อมูลสุขภาพระดับชุมชนสามารถสนับสนุนการคัดกรองความเสี่ยงโรคไม่ติดต่อเรื้อรังในผู้สูงอายุได้อย่างมีประสิทธิภาพ ผลการศึกษาแสดงให้เห็นว่าแม้ Random forest จะมีค่า Accuracy สูงกว่า แต่ XGBoost สามารถตรวจจับกลุ่มเสี่ยงได้ดีกว่าและลด False negative ได้มากกว่า จึงเหมาะสมสำหรับการประยุกต์ใช้ในงานคัดกรองสุขภาพเชิงรุกระดับชุมชน รวมไปถึงการเฝ้าระวังสุขภาพเชิงรุกและสนับสนุนการวางแผนดูแลสุขภาพผู้สูงอายุ อย่างไรก็ตามแบบจำลองดังกล่าวมีวัตถุประสงค์เพื่อใช้สนับสนุนการตัดสินใจเบื้องต้นเท่านั้น ไม่สามารถใช้แทนการวินิจฉัยหรือการพิจารณาทางคลินิกของบุคลากรทางการแพทย์ได้

ข้อเสนอแนะสำหรับการวิจัยในอนาคต ควรขยายพื้นที่ศึกษา เพิ่มข้อมูลด้านพฤติกรรมสุขภาพ เพื่อเพิ่มความสามารถในการประยุกต์ใช้แบบจำลองในบริบทที่หลากหลาย ควรเพิ่มการเปรียบเทียบกับแบบจำลอง Machine learning อื่น เพื่อประเมินประสิทธิภาพอย่างครอบคลุมมากยิ่งขึ้น และพัฒนาเครื่องมือดิจิทัลหรือระบบสนับสนุนการตัดสินใจด้านสุขภาพที่สอดคล้องกับบริบทของชุมชน

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยจากกองทุนส่งเสริมวิทยาศาสตร์ วิจัยและนวัตกรรม (สกว.) ประจำปีงบประมาณ 2568

เอกสารอ้างอิง

- กรมควบคุมโรค. (2566). รายงานสถานการณ์โรคไม่ติดต่อเรื้อรังของประเทศไทย. กระทรวงสาธารณสุข.
- สำนักงานสถิติแห่งชาติ. (2567). รายงานสถานการณ์ผู้สูงอายุในประเทศไทย. สำนักงานสถิติแห่งชาติ.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, 18(6), 3317. <https://doi.org/10.3390/ijerph18063317>
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. (3rd ed.). Morgan Kaufmann.
- Hu, B., Wang, C., Jiang, K., Shen, Z., Yang, X., Yin, M., Liang, B., Xie, Q., Ye, Y., & Gao, Z. (2021). Development and validation of a novel diagnostic model for initially clinical diagnosed gastrointestinal stromal tumors using an extreme gradient-boosting machine. *BMC Gastroenterol*, 21, 481. <https://doi.org/10.1186/s12876-021-02048-1>

- Islam, S. M. S., Talukder, A., Awal, M. A., Siddiqui, M. M. U., Ahamad, M. M., Ahammed, B., Rawal, L. B., Alizadehsani, R., Abawajy, J., Laranjo, L., Chow, C. K., & Maddison, R. (2022). Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian countries. *Frontiers in Cardiovascular Medicine*, 9, 839379. <https://doi.org/10.3389/fcvm.2022.839379>
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. <https://doi.org/10.48550/arXiv.2010.16061>
- Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. *Healthcare*, 8(3), 247. <https://doi.org/10.3390/healthcare8030247>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. (pp. 29–39). <https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>
- World Health Organization. (2014). *Global status report on noncommunicable diseases 2014*. <https://www.who.int/publications/i/item/9789241564854>