

Optimizing Feature Selection for Multi-Task Prediction of Non-Communicable Diseases on Imbalanced Healthcare Data

Yanisa Wongsombut¹ and Nithizethe Mhuadthongon^{2,*}

¹Science in Information and Communication Technology School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand; Email: 2639600218@stou.ac.th

^{2,*}Science in Information and Communication Technology School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand; Email: nithizethe.mhu@stou.ac.th

**Corresponding author: Nithizethe Mhuadthongon (nithizethe.mhu@stou.ac.th)*

Received: 14 November 2025; **Revised:** 13 January 2026;

Accepted: 15 January 2026; **Published:** 16 April 2026

Abstract

This research aims to investigate and compare the performance of feature selection techniques, encompassing Filter and Wrapper methods, for multi-task healthcare data characterized by high class imbalance. The primary objective is to develop a predictive model for non-communicable diseases (NCDs) by integrating the most efficient feature selection strategies with machine learning algorithms, evaluated through statistical metrics appropriate for imbalanced datasets. The methodology consists of six key stages: 1) data collection from Kaggle, comprising 253,680 records and 19 features; 2) data preprocessing and multi-target definition for three diseases; 3) feature selection using Filter methods (Information Gain, Gain Ratio, and Chi-Square) and Wrapper methods (Forward Selection and Backward Elimination); 4) class imbalance mitigation via the SMOTE technique; 5) classification model development using XGBoost, Random Forest, and LightGBM; and 6) model evaluation and hyperparameter optimization to identify the most suitable values. The experimental results reveal that the best-performing model, utilizing the Gain Ratio technique, significantly enhances the performance of the XGBoost model, successfully reducing the feature set from 19 to 7 key features—including BMI, Age, Income, Physical Health, General Health, Education, and Mental Health—while achieving a peak accuracy of 0.8768 and a maximum AUC of 0.9300. These findings indicate that the proposed approach is highly suitable for accurate and efficient multi-task prediction of non-communicable diseases.

Keywords: Multi-Task Learning, Class Imbalance, SMOTE, XGBoost, LightGBM



Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

Citation:

Wongsombut, Y., & Mhuadthongon, N. (2026). Optimizing Feature Selection for Multi-Task Prediction of Non-Communicable Diseases on Imbalanced Healthcare Data. *Journal of Computer and Creative Technology*, 4(1), e3069. <https://doi.org/10.65205/jcct.2026.e3069>

การเพิ่มประสิทธิภาพการคัดเลือกคุณลักษณะสำหรับการพยากรณ์โรคไม่ติดต่อเรื้อรัง แบบหลายงานจากข้อมูลสุขภาพที่ไม่สมดุล

ญาณิศา วงษ์สมบัติ¹ และ นิติเศรษฐ์ หมวดทองอ่อน^{2,*}

¹สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช จังหวัดนนทบุรี 11120 ประเทศไทย;
Email: 2639600218@stou.ac.th

^{2,*}สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช จังหวัดนนทบุรี 11120 ประเทศไทย;
Email: nithizethe.mhu@stou.ac.th

*ผู้ประพันธ์บทความ: นิติเศรษฐ์ หมวดทองอ่อน (nithizethe.mhu@stou.ac.th)

รับบทความ: 14 พฤศจิกายน 2568; แก้ไขบทความ: 13 มกราคม 2569;
ตอบรับบทความ: 15 มกราคม 2569; เผยแพร่บทความ: 16 เมษายน 2569

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบวิธีการ Filter และ Wrapper สำหรับข้อมูลสุขภาพที่เป็นแบบหลายงาน และมีความไม่สมดุลของข้อมูล เพื่อพัฒนาแบบจำลองการพยากรณ์โรคไม่ติดต่อเรื้อรัง (NCDs) โดยการบูรณาการเทคนิคการเลือกคุณลักษณะที่มีประสิทธิภาพสูงสุดร่วมกับแบบจำลองการเรียนรู้ของเครื่อง และเพื่อประเมินผลการคัดเลือกคุณลักษณะสำคัญของแต่ละเทคนิค โดยใช้ตัวชี้วัดทางสถิติที่เหมาะสมกับข้อมูลไม่สมดุล การดำเนินงานวิจัยประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ 1) การเก็บรวบรวมข้อมูลจากเว็บไซต์ Kaggle ซึ่งประกอบด้วย ข้อมูลจำนวน 253,680 แถว และคุณลักษณะ 19 คุณลักษณะ 2) การเตรียมข้อมูลและการกำหนดเป้าหมาย 3) โรค 3) การคัดเลือกคุณลักษณะโดยใช้ 2 วิธี คือ วิธีแบบกรอง (Information Gain, Gain Ratio, Chi-Square) และวิธีแบบควบรวม (Forward Selection, Backward Elimination) 4) การจัดการความไม่สมดุลของข้อมูลด้วยเทคนิค SMOTE 5) การสร้างแบบจำลองการจำแนกโดยใช้เทคนิค XGBoost, Random Forest และ LightGBM และ 6) การประเมินและปรับค่าพารามิเตอร์ของแบบจำลองเพื่อหาค่าที่เหมาะสมที่สุด ผลงานวิจัยพบว่า เทคนิคการคัดเลือกคุณลักษณะด้วยเทคนิค Gain Ratio ช่วยเพิ่มประสิทธิภาพให้เทคนิค XGBoost โดยลดคุณลักษณะจาก 19 เหลือเพียง 7 คุณลักษณะ ได้แก่ ดัชนีมวลกาย อายุ รายได้ สุขภาพภายใน 30 วัน ประเมินสุขภาพตนเอง ระดับการศึกษา และสุขภาพจิตใน 30 วัน ซึ่งให้ค่าความถูกต้องสูงสุดอยู่ที่ 0.8768 และให้ค่าพื้นที่ใต้เส้นโค้ง (AUC) สูงสุดอยู่ที่ 0.9300 แสดงให้เห็นว่าวิธีนี้มีความเหมาะสมในการพยากรณ์โรคไม่ติดต่อเรื้อรังแบบหลายงานได้อย่างแม่นยำและรวดเร็ว

คำสำคัญ: การเรียนรู้แบบหลายงาน, ข้อมูลไม่สมดุล, SMOTE, XGBoost, LightGBM



Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

การอ้างอิง:

ญาณิศา วงษ์สมบัติ, และ นิติเศรษฐ์ หมวดทองอ่อน. (2569). การเพิ่มประสิทธิภาพการคัดเลือกคุณลักษณะสำหรับการพยากรณ์โรคไม่ติดต่อเรื้อรังแบบหลายงานจากข้อมูลสุขภาพที่ไม่สมดุล. วารสารคอมพิวเตอร์และเทคโนโลยีสร้างสรรค์, 4(1), e3069.

<https://doi.org/10.65205/jcct.2026.e3069>

1. บทนำ (Introduction)

ในศตวรรษที่ 21 ปัญหาโรคไม่ติดต่อเรื้อรัง (Non-Communicable Diseases: NCDs) ได้กลายเป็นปัญหาสำคัญระดับโลกที่ส่งผลกระทบต่อคุณภาพชีวิตของประชากร โดยเฉพาะโรคเบาหวาน (Diabetes Mellitus) โรคความดันโลหิตสูง (Hypertension) และโรคหลอดเลือดสมอง (Stroke) ซึ่งเป็นสาเหตุสำคัญของภาวะแทรกซ้อนร้ายแรง เช่น โรคหัวใจ ไตวาย และอัมพาต องค์การอนามัยโลก (World Health Organization, 2023) รายงานว่า โรคความดันโลหิตสูงเพียงโรคเดียวนั้นเป็นสาเหตุการเสียชีวิตมากกว่า 10 ล้านรายต่อปี ขณะที่จำนวนผู้ป่วยเบาหวานเพิ่มขึ้นจาก 200 ล้านรายในปี ค.ศ. 1990 มากกว่า 830 ล้านรายในปี ค.ศ. 2022 จากข้อมูลดังกล่าวสะท้อนให้เห็นถึงความจำเป็นในการพัฒนาระบบวิเคราะห์ข้อมูลเพื่อการพยากรณ์โรคที่มีความแม่นยำ และสามารถประยุกต์ใช้ในระดับนโยบายสาธารณสุขได้

เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning: ML) ถูกนำมาใช้ในการวิเคราะห์ข้อมูลสุขภาพอย่างแพร่หลาย เนื่องจากสามารถจัดการข้อมูลขนาดใหญ่และค้นหารูปแบบเชิงซ้อนได้ดี โดยงานวิจัยก่อนหน้าได้แสดงให้เห็นว่าเทคนิคแบบเอนเซมเบิล เช่น XGBoost และ LightGBM มีประสิทธิภาพสูงในการพยากรณ์โรคไม่ติดต่อเรื้อรัง (Rufo et al., 2021) อย่างไรก็ตาม งานวิจัยส่วนใหญ่ยังมุ่งเน้นการพยากรณ์แบบเป้าหมายเดียว และยังขาดการศึกษาในบริบทของข้อมูลสุขภาพแบบหลายเป้าหมายที่มีความไม่สมดุลสูง

กระบวนการคัดเลือกคุณลักษณะที่สำคัญ (Feature Selection) เป็นแนวทางสำคัญในการลดจำนวนตัวแปรที่ไม่จำเป็นและเพิ่มประสิทธิภาพของแบบจำลอง โดยสามารถแบ่งออกได้เป็น 2 กลุ่มหลัก ได้แก่ (1) Filter Methods เช่น Information Gain, Gain Ratio และ Chi-Square ที่ใช้ตัวชี้วัดทางสถิติในการจัดอันดับของตัวแปร และ (2) Wrapper Methods เช่น Forward Selection และ Backward Elimination ที่ใช้การเรียนรู้ของเครื่องร่วมกับการค้นหาคุณลักษณะสำคัญที่ให้ผลลัพธ์ดีที่สุดต่อแบบจำลอง อย่างไรก็ตาม ข้อจำกัดที่สำคัญของกระบวนการดังกล่าว คือ เทคนิคการเลือกคุณลักษณะที่ใช้กันอย่างแพร่หลายในงานวิจัยหลายแขนงนั้น มักถูกนำไปประยุกต์ใช้ในบริบทของข้อมูลที่มีโครงสร้างง่าย มีเป้าหมายเดียว และไม่คำนึงถึงปัจจัยเชิงระบบที่ซับซ้อน (Noroozi et al., 2023) ดังนั้นจึงส่งผลให้ไม่สามารถสรุปได้ชัดเจนว่า เทคนิคใดเหมาะสมที่สุดต่อการนำไปใช้งานได้จริงในสถานการณ์ที่ข้อมูลมีลักษณะท้าทายหลายด้านในเวลาเดียวกัน ทั้งนี้ การเลือกเทคนิคอย่างไม่เหมาะสมอาจนำไปสู่แบบจำลองที่ไม่แม่นยำ ใช้งานไม่ได้จริง หรือแม้แต่ให้ผลวิเคราะห์ที่บิดเบือน ซึ่งถือเป็นความเสี่ยงสำคัญในบริบทของการพยากรณ์โรค

ดังนั้น งานวิจัยนี้จึงมุ่งศึกษาและเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะภายใต้แบบจำลองการเรียนรู้ของเครื่อง ทั้งในรูปแบบ Filter Methods และ Wrapper Methods และประเมินผลการคัดเลือกคุณลักษณะสำคัญของแต่ละเทคนิค โดยใช้ตัวชี้วัดทางสถิติที่เหมาะสมกับข้อมูลไม่สมดุล เช่น ค่าพื้นที่ใต้กราฟ (AUC) และค่าความถูกต้อง (Accuracy) รวมถึงเพื่อพัฒนาแบบจำลองพยากรณ์โรคไม่ติดต่อเรื้อรังแบบหลายงานจากข้อมูลสุขภาพที่ไม่สมดุล และช่วยคัดกรองผู้ป่วยที่มีความเสี่ยงได้อย่างมีประสิทธิภาพและน่าเชื่อถือในอนาคต

2. วัตถุประสงค์งานวิจัย (Research Objectives)

1. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบ Filter Methods และ Wrapper Methods สำหรับข้อมูลสุขภาพที่เป็นแบบหลายงาน และมีความไม่สมดุลของข้อมูล
2. เพื่อพัฒนาแบบจำลองการพยากรณ์โรคไม่ติดต่อเรื้อรัง โดยการบูรณาการเทคนิคการคัดเลือกคุณลักษณะที่มีประสิทธิภาพสูงสุด ร่วมกับแบบจำลองการเรียนรู้ของเครื่อง
3. เพื่อประเมินผลการคัดเลือกคุณลักษณะสำคัญของแต่ละเทคนิค โดยใช้ตัวชี้วัดทางสถิติที่เหมาะสมกับข้อมูลไม่สมดุล เช่น ค่าพื้นที่ใต้เส้นโค้ง และค่าความถูกต้อง

3. การทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้อง (Literature Review)

3.1 การเรียนรู้ของเครื่อง

จากการทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้อง สรุปได้ว่า การเรียนรู้ของเครื่อง เป็นกระบวนการค้นหารูปแบบหรือความสัมพันธ์ที่มีประโยชน์จากข้อมูลที่เก็บรวบรวมมาในอดีต โดยใช้แบบจำลองต่าง ๆ เพื่อนำมาใช้ในการทำนายการเกิดเหตุการณ์ในอนาคต (Tepdang, 2023) โดยงานวิจัยนี้มุ่งเน้นการประยุกต์ใช้การเรียนรู้ของเครื่อง เพื่อพยากรณ์

ความเสี่ยงของการเกิดโรคไม่ติดต่อเรื้อรัง ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมอง ซึ่งเป็นสาเหตุสำคัญของการเสียชีวิตและความพิการในประชากร ดังนั้น รูปแบบการเรียนรู้ที่ใช้ในงานวิจัยนี้คือ การเรียนรู้แบบมีผู้สอน ซึ่งเป็นการสร้างแบบจำลองจากชุดข้อมูลที่มีทั้งข้อมูลนำเข้าและข้อมูลผลลัพธ์ที่ถูกระบุไว้ล่วงหน้า โดยแบบจำลองที่นำมาใช้คือ แบบจำลองการจำแนกประเภท เนื่องจากลักษณะของข้อมูลผลลัพธ์เป็นแบบข้อมูลเชิงจำแนกประเภท ซึ่งอยู่ในรูปแบบไบนารี คือ การแบ่งผู้ป่วยออกเป็น 2 กลุ่ม ได้แก่ “เป็นโรค” หรือ “ไม่เป็นโรค”

นอกจากนั้นแล้ว ยังมุ่งเน้นไปที่กระบวนการคัดเลือกคุณลักษณะ ซึ่งเป็นส่วนหนึ่งของ ขั้นตอนที่ 2 คือ การเตรียมข้อมูล จากกระบวนการเรียนรู้ของเครื่อง 6 ขั้นตอน (Salhi et al., 2023) เนื่องจากการคัดเลือกคุณลักษณะที่สำคัญ ส่งผลโดยตรงต่อประสิทธิภาพของแบบจำลองที่พัฒนาขึ้น สำหรับการดำเนินการในเบื้องต้นนั้นจะเริ่มจากการจัดเก็บข้อมูลจากแหล่งข้อมูล และดำเนินการตามขั้นตอนย่อย 3 ขั้นตอน ได้แก่ (1) การคัดเลือกข้อมูล (2) การกลั่นกรองข้อมูล และ (3) การแปลงรูปแบบของข้อมูล หลังจากการเตรียมข้อมูลเสร็จสิ้น ข้อมูลจะถูกนำเข้าสู่กระบวนการคัดเลือกคุณลักษณะ โดยใช้เทคนิคที่เหมาะสมเพื่อคัดเลือกเฉพาะคุณลักษณะที่มีความสัมพันธ์กับผลลัพธ์ที่ต้องการวิเคราะห์ พร้อมทั้งตัดคุณลักษณะที่ไม่สำคัญออก

3.2 การคัดเลือกคุณลักษณะสำคัญ

จากการทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้อง สรุปได้ว่า การคัดเลือกคุณลักษณะสำคัญ เป็นกระบวนการที่เน้นการคัดเลือกคุณลักษณะที่ไม่มีความสำคัญหรือไม่เกี่ยวข้องกับเป้าหมายการวิเคราะห์ออก (Jupriyadi et al., 2024) โดยงานวิจัยนี้เลือกใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ได้แก่

1. วิธีแบบกรอง (Filter Methods) คือ การคัดเลือกคุณลักษณะสำคัญ โดยคำนวณค่าน้ำหนักหรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ งานวิจัยนี้เลือกใช้วิธีแบบกรอง 3 วิธี ได้แก่

1) เทคนิค Information Gain เป็นเทคนิคที่ใช้ในการชี้วัดเพื่อคัดเลือกคุณลักษณะข้อมูลสำคัญ โดยมุ่งเน้นคัดเลือกเฉพาะคุณลักษณะที่จำเป็นมากที่สุดต่อการสร้างแบบจำลอง

2) เทคนิค Gain Ratio เป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อย ที่พัฒนามาจาก Information Gain เนื่องจากการใช้ Information Gain ในการแบ่งชุดข้อมูลจะมีโอกาสทำให้เกิดความเอนเอียงขึ้น

3) เทคนิค Chi-Square เป็นการประเมินค่าคุณลักษณะ โดยใช้การคำนวณค่า Chi-Square ทางสถิติเพื่อศึกษาถึงการแจกแจงความถี่ของตัวแปรคุณลักษณะว่าเป็นไปตามรูปแบบที่กำหนดไว้หรือไม่

2. วิธีแบบครอบรวม (Wrapper Methods) คือ การคัดเลือกคุณลักษณะสำคัญ โดยคำนวณค่าน้ำหนักของการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูลเพื่อนำมาสร้างชุดคุณลักษณะใหม่ โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากชุดเดิม (Sunggad & Maneerat, 2023) งานวิจัยนี้เลือกใช้วิธีแบบครอบรวม 2 วิธี ได้แก่

1) เทคนิค Forward Selection ซึ่งดำเนินการตามแนวคิดของ Sequential Forward Selection (SFS) โดยใช้วิธีการคัดเลือกคุณลักษณะของตัวแปรอิสระเข้ามาในสมการทีละตัว และดำเนินการทำซ้ำจนไม่มีตัวแปรที่ผ่านเกณฑ์เพิ่มเข้ามา

2) เทคนิค Backward Elimination ซึ่งดำเนินการตามแนวคิดของ Sequential Backward Elimination (SBE) โดยเริ่มจากการนำตัวแปรพยากรณ์ทั้งหมดในสมการ มาทำการลบตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนต่ำที่สุดออกทีละตัว หากการลบตัวแปรดังกล่าวไม่ส่งผลให้ค่าความแม่นยำของแบบจำลองลดลงอย่างมีนัยสำคัญทางสถิติ กระบวนการจะดำเนินต่อไปโดยลบตัวแปรที่ไม่สำคัญออกอย่างต่อเนื่อง จนกระทั่งเหลือเฉพาะตัวแปรที่มีอิทธิพลอย่างมีนัยสำคัญต่อแบบจำลองเท่านั้น

ทั้งนี้ แม้ว่าจะมีแนวคิด Hybrid Feature Selection ซึ่งเป็นการผสาน Filter Methods และ Wrapper Methods เข้าด้วยกัน แต่ในงานวิจัยนี้มุ่งเน้นการเปรียบเทียบประสิทธิภาพของ Filter Methods และ Wrapper Methods แบบแยกชัดเจน เพื่อให้สามารถวิเคราะห์ผลกระทบของแต่ละแนวทางต่อข้อมูลสุขภาพแบบหลายงานและข้อมูลไม่สมดุลได้อย่างเป็นระบบ

3.3 การจัดการความไม่สมดุลของข้อมูล

การจัดการข้อมูลในกรณีที่จำนวนตัวอย่างในแต่ละกลุ่ม (Class) มีความแตกต่างกันอย่างมาก เช่น ข้อมูลกลุ่มหนึ่งมีสัดส่วนน้อย เมื่อเทียบกับอีกกลุ่มหนึ่งที่มีจำนวนข้อมูลมากกว่า ทำให้ข้อมูลเกิดความไม่สมดุลขึ้นส่งผลให้แบบจำลอง

เอนเอียงไปทำนายข้อมูลกลุ่มที่มีจำนวนมากกว่า และจำแนกข้อมูลกลุ่มส่วนน้อยไม่แม่นยำ งานวิจัยนี้ จึงจัดการความไม่สมดุลของข้อมูล โดยการปรับจำนวนข้อมูลในแต่ละกลุ่มให้มีค่าเท่ากัน เพื่อเพิ่มความแม่นยำและลดอคติของแบบจำลองในการจำแนกกลุ่มเป้าหมาย โดยเฉพาะอย่างยิ่ง กลุ่มผู้ป่วยโรคตามวัตถุประสงค์ของงานวิจัย

3.4 การสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์

การสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ หรือ SMOTE (Synthetic Minority Oversampling Technique) คือ เทคนิคการเพิ่มจำนวนข้อมูล (Oversampling) สำหรับกลุ่มข้อมูลที่มีจำนวนน้อย (Minority Class) โดยสร้างข้อมูลสังเคราะห์ขึ้นมาใหม่จากข้อมูลที่มีอยู่เดิมในบริเวณใกล้เคียง ซึ่งจะใช้วิธีการของอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด k (K-Nearest Neighbors: k-NN) เพื่อคัดเลือกข้อมูลเพื่อนบ้านที่อยู่ใกล้ที่สุด จำนวน k ตัว จากนั้นทำการคำนวณระยะทางระหว่างข้อมูลต้นฉบับกับข้อมูลเพื่อนบ้าน เพื่อหาตำแหน่งจุดใหม่ที่อยู่ระหว่างข้อมูลทั้งสองจุด และนำค่าดังกล่าวมาสร้างเป็นข้อมูลสังเคราะห์เพิ่มเติมในกลุ่มคลาสส่วนน้อยให้มีความสมดุลกับคลาสส่วนใหญ่

การกำหนดค่า k หมายถึง จำนวนข้อมูลที่อยู่ใกล้เคียงที่สุด ซึ่งถูกนำมาใช้ในการคำนวณเพื่อสร้างข้อมูลใหม่ เช่น $k = 1, 2, 3, \dots, n$ โดยกระบวนการนี้จะช่วยลดความไม่สมดุลของข้อมูล และส่งผลให้แบบจำลองการเรียนรู้ของเครื่องสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น (Romsaiyud, 2024) แสดงได้ดังสมการที่ (1)

$$X_{new} = X_i + (\hat{X} - X_i) * \delta \quad (1)$$

โดย X_{new} คือ ตำแหน่งของข้อมูลจากคลาสส่วนน้อยที่ถูกสังเคราะห์ขึ้นมาใหม่ ระหว่างตำแหน่งของ X_i และ $\hat{X}$ ของคลาสส่วนน้อย และ δ เป็นค่าการสุ่มตัวเลขระหว่าง $[0,1]$

3.5 เทคนิคการจำแนกประเภทข้อมูล

เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างแบบจำลองจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างที่เรียกว่า ข้อมูลฝึกสอน (Training Data) ที่แต่ละแถวของข้อมูลประกอบด้วยแอตทริบิวต์ (Attribute) จำนวนมาก ซึ่งอาจจะเป็นค่าต่อเนื่อง ตัวเลข หรือข้อมูลจำแนกประเภท โดยมีแอตทริบิวต์ที่ทำหน้าที่ในการจำแนกประเภทข้อมูลเป็นตัวกำหนดคลาส (Class) งานวิจัยนี้เลือกใช้ 3 เทคนิค ดังนี้

1. เทคนิค XGBoost เป็นเทคนิค Gradient Boosting ที่ได้รับการปรับแต่งให้เหมาะสม โดยใช้ต้นไม้ตัดสินใจที่สร้างขึ้นแบบลำดับขั้น เพื่อแก้ไขข้อผิดพลาดจากต้นไม้ก่อนหน้า รวมถึงการปรับค่าให้เป็นมาตรฐาน (Regularization) เพื่อลดการเกิด Overfitting (Li et al., 2024)

2. เทคนิค Random Forest เป็นเทคนิคการจำแนกประเภทข้อมูลที่อยู่ในกลุ่มอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน โดยใช้การรวมกันของต้นไม้ตัดสินใจหลายต้น ซึ่งทำงานร่วมกันในลักษณะของการเรียนรู้แบบเอนเซมเบิล สามารถทำงานโดยการสร้างต้นไม้ตัดสินใจหลายต้นจากชุดข้อมูลที่แตกต่างกัน และนำผลลัพธ์ของต้นไม้ที่ได้มารวมกันเพื่อหาคำตอบที่ดีที่สุด โดยใช้หลักการของ “เสียงข้างมาก” ในกรณีของการจำแนกประเภท (Classification) หรือใช้ค่าเฉลี่ยในกรณีของ การถดถอย (Regression) (Sai et al., 2023)

3. เทคนิค LightGBM เป็นเทคนิค Gradient Boosting ที่ถูกออกแบบมาร่วมกับเทคนิคการประมวลผลอื่น ๆ เพื่อเพิ่มประสิทธิภาพของการคำนวณ ใช้หน่วยความจำน้อย และสามารถจัดการข้อมูลที่มีขนาดใหญ่และมีมิติสูงได้อย่างมีประสิทธิภาพ เช่น Gradient-based One Side Sampling (GOSS) และ Exclusive Feature Bundling (EFB) (Sai et al., 2023)

3.6 การวัดประสิทธิภาพแบบจำลอง

1. การทดสอบแบบไขว้ทับ (K-Fold Cross Validation) เป็นวิธีที่ใช้ข้อมูลฝึกสอนบางส่วนมาทำการแบ่งชุดข้อมูลออกเป็น k กลุ่ม เพื่อนำมาประเมินประสิทธิภาพของแบบจำลอง โดยจะเริ่มจากการแบ่งกลุ่มข้อมูลเริ่มต้นออกเป็น ส่วน ๆ เท่า ๆ กัน โดยในงานวิจัยนี้เลือกใช้ค่า $k = 10$ ซึ่งหมายถึง กระบวนการทดสอบจะใช้ข้อมูล 9 กลุ่ม ในการพัฒนาแบบจำลอง และข้อมูล 1 กลุ่ม ในการทดสอบแบบจำลอง กระบวนการนี้จะทำซ้ำทั้งหมด 10 ครั้ง โดยในแต่ละครั้งจะสลับกลุ่มข้อมูลที่ใช้ทดสอบ เพื่อให้แน่ใจว่าทุกกลุ่มข้อมูลได้ถูกนำมาใช้ในการทดสอบอย่างน้อย 1 ครั้ง

2. คอนฟิวชันเมตริก (Confusion Matrix) เป็นการคำนวณค่าจากตาราง Confusion Matrix ซึ่งเป็นตารางข้อมูลที่แสดงผลการจำแนกประเภทข้อมูลที่ถูกต้องและไม่ถูกต้อง ตารางนี้จะแสดงความสัมพันธ์ระหว่างค่าจริงและค่าที่แบบจำลองทำนาย ทำให้สามารถวิเคราะห์ข้อผิดพลาดของแบบจำลองได้อย่างละเอียด และนำไปคำนวณค่าต่าง ๆ เช่น ค่าความถูกต้องได้ (Rainio et al., 2024) แสดงดัง Table 1.

Table 1. Confusion Matrix

		Predicted	
		Positive	Negative
Class predict	Positive	TP	FN
	Negative	FP	TN

3. ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องในการจำแนกประเภท เพื่อบ่งบอกระดับความถูกต้องในการจำแนกประเภทข้อมูล แบบจำลองที่ได้จากการประมวลผล ได้แก่ สัดส่วนระหว่างจำนวนข้อมูลทั้งหมดที่จำแนกประเภทความถูกต้อง ทั้งประเภท Positive และ Negative กับจำนวนข้อมูลทั้งหมดที่ถูกนำมาจำแนก โดยพิจารณาทั้งผลลัพธ์ที่จำแนกถูกต้องและผิดพลาด

4. ค่าความแม่นยำ (Precision) เป็นการวัดความสามารถของระบบในการจัดคำตอบหรือข้อมูลที่ไม่เกี่ยวข้องออกไป ถ้าระบบสามารถจัดคำตอบหรือข้อมูลที่ไม่เกี่ยวข้องออกไปได้มาก จะแสดงถึงความแม่นยำของแบบจำลองที่สูงขึ้น

5. ค่าเรียกคืน (Recall) เป็นการวัดประสิทธิภาพของแบบจำลองที่แสดงถึงความสามารถในการระบุข้อมูลที่เป็บวกจริง (True Positive) ได้ครบถ้วน โดยคำนวณจากสัดส่วนของข้อมูลที่แบบจำลองทำนายถูกต้องว่าเป็นบวก เมื่อเทียบกับข้อมูลที่เป็นบวกจริงทั้งหมด

6. ค่าประสิทธิภาพโดยรวม (F1-Score) เป็นการวัดค่า Precision และค่า Recall ใช้สำหรับบอกความมีประสิทธิภาพของแบบจำลองการจำแนกประเภทว่า มีความเหมาะสมในการนำไปใช้กับข้อมูลอื่น ๆ ที่เป็นข้อมูลใหม่ได้ดีเพียงใด

7. ค่าพื้นที่ใต้เส้นโค้ง (Area Under Curve: AUC) เป็นการวิเคราะห์ข้อมูลโดยใช้สถิติ ROC Curve Analysis เพื่อหาความแม่นยำของการทำนายค่าอ้างอิง หากค่าพื้นที่ใต้เส้นโค้งมีค่าเข้าใกล้ 1 แสดงให้เห็นว่า การทำนายมีความแม่นยำสูงและเปรียบเทียบความเหมาะสมของตัวแบบการทำนาย (Khan, 2023)

จากข้อมูลที่กล่าวข้างต้น เมื่อได้ทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง พบว่า งานวิจัยส่วนใหญ่มุ่งเน้นการพยากรณ์โรคเดี่ยว เช่น งานวิจัยของ Tepdang (2023) กล่าวถึงการจำแนกผู้ป่วยโรคเบาหวานด้วยวิธีการเรียนรู้ของเครื่อง โดยการหาปัจจัยของการคัดเลือกคุณลักษณะ หรืองานวิจัยของ Rufo et al. (2021) กล่าวถึงการพัฒนาแบบจำลองวินิจฉัยโรคเบาหวาน โดยใช้เทคนิค LightGBM กับข้อมูลผู้ป่วยจากโรงพยาบาลในเอธิโอเปีย ดังนั้นจึงเห็นได้ว่า การคัดเลือกคุณลักษณะสำหรับข้อมูลสุขภาพแบบหลายงานนั้นยังได้รับการศึกษาอย่างจำกัด ซึ่งงานวิจัยนี้สามารถเติมเต็มช่องว่างขององค์ความรู้ดังกล่าวได้

4. กรอบแนวคิดงานวิจัย (Conceptual Framework)

จากกรอบแนวคิดงานวิจัย (ดัง Figure 1.) ประกอบด้วย การศึกษาวรรณกรรมที่เกี่ยวข้อง ได้แก่ แนวคิดและทฤษฎีที่เกี่ยวข้อง งานวิจัยที่เกี่ยวข้อง และการสรุปองค์ความรู้จากการทบทวนวรรณกรรมที่เกี่ยวข้อง ซึ่งได้อธิบายไว้ในหัวข้อที่ผ่านมา จากนั้นจึงได้ศึกษาวิธีดำเนินงานวิจัย ได้แก่ ขั้นตอนการดำเนินงานวิจัย ประชากรและกลุ่มตัวอย่าง โดยวิธีดำเนินงานวิจัย เริ่มจากการวิเคราะห์โครงสร้างข้อมูลสุขภาพ ความสัมพันธ์ของปัจจัยกำกับ และปัญหาความไม่สมดุลของข้อมูล จากนั้นดำเนินการเตรียมข้อมูลโดยปรับปรุงข้อมูลและใช้เทคนิค SMOTE ร่วมกับการคัดเลือกคุณลักษณะแบบ Filter Methods และ Wrapper Methods ต่อมาใช้แบบจำลองการเรียนรู้ของเครื่อง ได้แก่ XGBoost, Random Forest และ LightGBM และทำการประเมินประสิทธิภาพของแบบจำลองด้วย Accuracy, F1-Score และ AUC ซึ่งกำหนดให้

ตัวแปรต้น คือ เทคนิคการคัดเลือกคุณลักษณะ สำหรับการพยากรณ์โรคไม่ติดต่อเรื้อรังแบบหลายงาน

ตัวแปรตาม คือ ประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องที่ใช้ในการพยากรณ์ เมื่อดำเนินการวิเคราะห์ข้อมูลเสร็จสิ้น จะได้ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคการคัดเลือกคุณลักษณะ เพื่อระบุแนวทางที่เหมาะสมที่สุดสำหรับการพัฒนาแบบจำลองพยากรณ์ที่มีความแม่นยำสูง

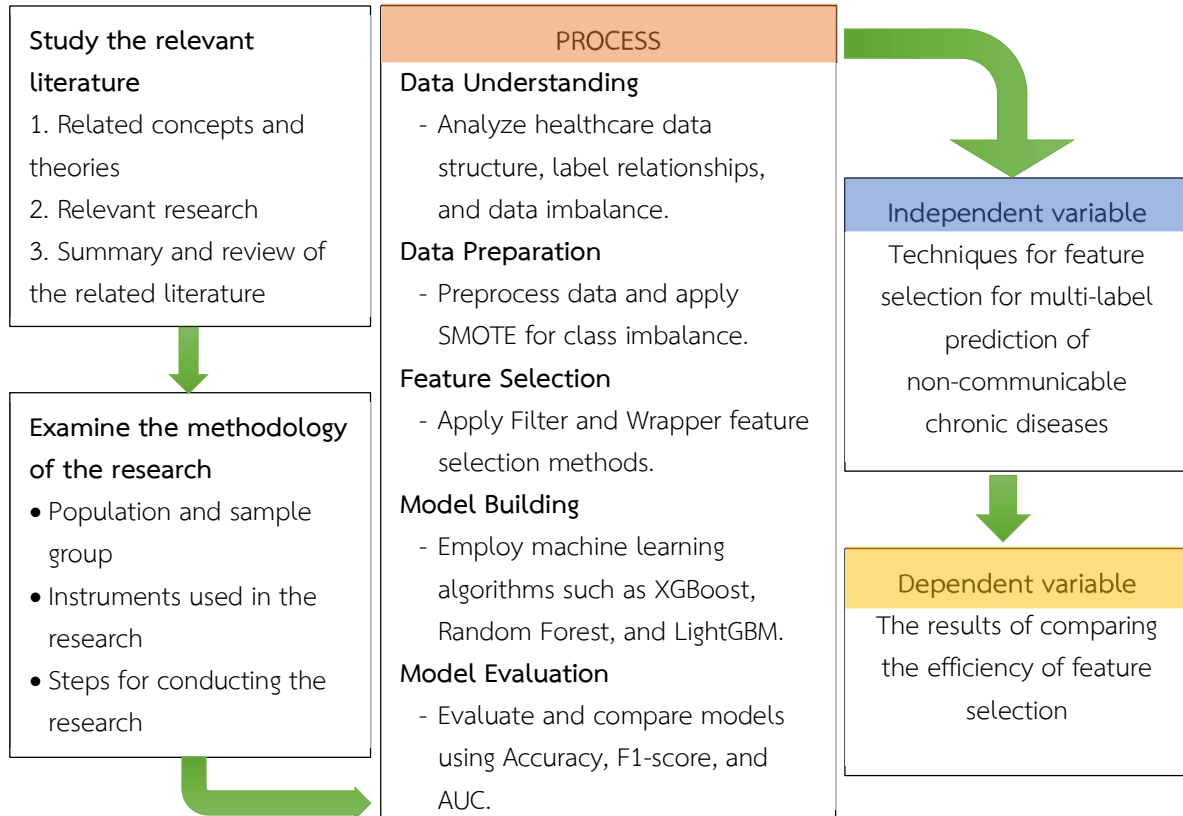


Figure 1. Conceptual Framework

5. วิธีดำเนินงานวิจัย (Research Methodology)

5.1 ชุดข้อมูล

ชุดข้อมูล ได้จากเว็บไซต์ www.kaggle.com ซึ่งเป็นชุดข้อมูลสาธารณะ (Teboul, n.d.) ประกอบด้วย ข้อมูลที่มีคุณลักษณะสำคัญ จำนวน 22 คุณลักษณะ แถวข้อมูลจำนวน 253,680 แถว และกำหนดเงื่อนไขแบบหลายเป้าหมายจำนวน 3 เป้าหมาย ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมอง ทำให้เหลือจำนวนคุณลักษณะที่สำคัญ จำนวน 19 คุณลักษณะ

5.2 เครื่องมือที่ใช้ในงานวิจัย

เครื่องคอมพิวเตอร์ที่มีระบบปฏิบัติการ Windows 11 แบบ 64 bit หน่วยความจำหลัก 8 GB สำหรับรองรับการประมวลผลข้อมูล โปรแกรม Microsoft Excel สำหรับจัดเก็บและเตรียมข้อมูลก่อนการวิเคราะห์ และโปรแกรม Google Collaboratory สำหรับการเขียนและประมวลผลโค้ดโปรแกรมภาษาไพทอน (Python) ผ่านเว็บเบราว์เซอร์

5.3 ขั้นตอนการดำเนินงานวิจัย

ขั้นตอนการดำเนินงานวิจัยประกอบด้วย 6 ขั้นตอน ดังนี้

1. การเก็บรวบรวมข้อมูล

ดำเนินการเก็บรวบรวมข้อมูล ดัง Figure 2. โดยใช้ข้อมูลจากเว็บไซต์ www.kaggle.com และกำหนดเงื่อนไขหลายเป้าหมายจำนวน 3 เป้าหมาย คือ โรคเบาหวาน (มี 3 คลาส ได้แก่ 0,1,2) โรคความดันโลหิตสูง (มี 2 คลาส

ได้แก่ 0, 1) และโรคหลอดเลือดสมอง (มี 2 คลาส ได้แก่ 0,1) โดยรายละเอียดของคุณลักษณะต่าง ๆ ที่สำคัญ สามารถแสดงผลลัพธ์ ดัง Table 2.

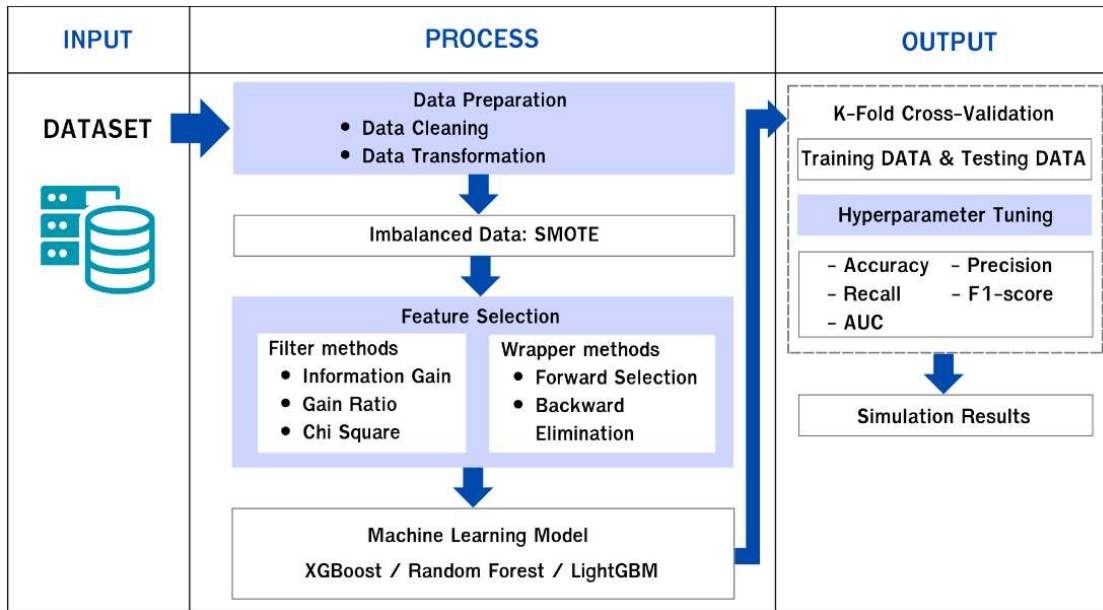


Figure 2. Steps for Feature Selection for Multi-label Prediction of Non-Communicable Chronic Diseases from Imbalanced Health Data

Table 2. Operational Definitions of Variables

No.	Attribute	Value	Description	Encoding
1	Diabetes_012	ordinal	Diabetes Status (3 levels)	0 = No diabetes, 1 = Pre-diabetes, 2 = Diabetes
2	Hypertension	nominal	High Blood Pressure (HighBP)	0 = No, 1 = Yes
3	Stroke	nominal	History of Stroke	0 = No, 1 = Yes
4	HighChol	nominal	High Cholesterol (HighChol)	0 = No, 1 = Yes
5	CholCheck	nominal	Cholesterol Check within 5 years	0 = No, 1 = Yes
6	BMI	scale	Weight balance index	continuous value.
7	Smoker	nominal	Smoking Behavior/ Smoking Status	0 = No, 1 = Yes
8	PhysActivity	nominal	Physical Activity/ Exercise	0 = No, 1 = Yes
9	HeartDiseaseor Attack	nominal	Heart Disease or Myocardial Infarction	0 = No, 1 = Yes
10	Fruits	nominal	Daily Fruit Consumption	0 = No, 1 = Yes
11	Veggies	nominal	Daily Vegetable Consumption	0 = No, 1 = Yes
12	HvyAlcohol Consump	nominal	Heavy Alcohol Consumption	0 = No, 1 = Yes
13	AnyHealthcare	nominal	Health Care Coverage/ Insurance	0 = No, 1 = Yes

Table 2. Operational Definitions of Variables (Cont.)

No.	Attribute	Value	Description	Encoding
14	NoDocbcCost	nominal	No Medical Care due to Cost	0 = No, 1 = Yes
15	GenHlth	ordinal	General Health Self-Perception	1 = Excellent, 2 = Very good, 3 = Good, 4 = Fair, 5 = Poor
16	MentHlth	scale	Number of Poor Mental Health Days	0-30 days
17	PhysHlth	scale	Number of Poor Physical Health Days	0-30 days
18	DiffWalk	nominal	Difficulty Walking/ Physical Disability	0 = No, 1 = Yes
19	Sex	nominal	Biological Sex	0 = Female, 1 = Male
20	Age	ordinal	Age Categories (13 levels)	1 = 18-24 , 2 = 25-29 3 = 30-34 , 4 = 35-39 5 = 40-44 , 6 = 45-49 7 = 50-54 , 8 = 55-59 9 = 60-64 , 10 = 65-69 11 = 70-74 , 12 = 75-79 13 = 80 or older
21	Education	ordinal	Highest Educational Attainment	1=Never attended school or only kindergarten 2=Elementary (Grades 1-8) 3=Some High School (Grades 9-11) 4=High School Graduate (Grade 12 or GED)5=Some 5=College or Technical School (1-3 years) 6=College Graduate (bachelor's degree or higher)
22	Income	ordinal	Income Level/ Household Income	1=Less than \$10,000 2=\$10,000 to less than \$15,000 3=\$15,000 to less than \$20,000 4=\$20,000 to less than \$25,000 5=\$25,000 to less than \$35,000 6=\$35,000 to less than \$50,000 7=\$50,000 to less than \$75,000 8=\$75,000 or more

2. การเตรียมข้อมูล

ดำเนินการเตรียมข้อมูลให้เหมาะสมสำหรับการวิเคราะห์ ดังนี้

2.1 การกลั่นกรองข้อมูล โดยตรวจสอบและคัดกรองข้อมูลที่ผิดปกติ เช่น ค่าดัชนีมวลกาย (BMI) ให้อยู่ช่วง 10-60 เพื่อลดผลกระทบต่อความแม่นยำของแบบจำลอง และไม่จำเป็นต้องเติมเต็มข้อมูลเพิ่มเติมในขั้นตอนนี้ เนื่องจากข้อมูลไม่มีค่าว่างปรากฏในฐานข้อมูล

2.2 การแปลงข้อมูล จากรูปแบบเดิมที่มีค่า 0, 1 และ 2 ให้เป็นข้อมูลประเภทจำนวนเต็มแบบไบนารี โดยกำหนดให้ค่า 0 แทนกลุ่มที่ไม่เป็นโรคเบาหวาน และค่า 1 แทนกลุ่มที่เป็นโรคเบาหวาน เพื่อให้ง่ายต่อการวิเคราะห์ ซึ่งชุดข้อมูลดังกล่าวจะถูกอัปโหลด เพื่อนำเข้าสู่โปรแกรม Google Colaboratory หลังจากคัดกรองข้อมูลที่มีคุณภาพและเหมาะสมต่อการวิเคราะห์ โดยเมื่อดำเนินการแล้วเสร็จ จะได้จำนวนข้อมูลของแต่ละเป้าหมายที่กำหนดไว้ในงานวิจัย ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมอง

2.3 การจัดการความไม่สมดุลของข้อมูล โดยเมื่อทำการเตรียมข้อมูลให้เหมาะสมสำหรับการวิเคราะห์เรียบร้อยแล้ว แต่ยังมีจำนวนข้อมูลที่มีความแตกต่างกันค่อนข้างมากระหว่างคนเป็นโรคกับคนที่ไม่ได้เป็นโรค วิธีการ SMOTE จึงถูกนำมาประยุกต์ใช้เพื่อเพิ่มข้อมูลให้มีจำนวนที่ใกล้เคียงกัน ซึ่งสามารถสังเกตผลลัพธ์ของจำนวนข้อมูลแต่ละกลุ่มเป้าหมายได้ ดัง Table 3.

Table 3. Class Distribution Before and After SMOTE Oversampling

Target Variable	Before			After		
	Class 0	Class 1	Total	Class 0	Class 1	Total
Diabetes Mellitus	213,703	39,977	253,680	213,703	213,703	427,406
Hypertension	144,851	108,829	253,680	144,851	144,851	289,702
Stroke	243,388	10,292	253,680	243,388	243,388	486,776

2.4 การคัดเลือกคุณลักษณะสำคัญ โดยใช้ชุดข้อมูลที่ผ่านการกลั่นกรองข้อมูลและการแปลงรูปข้อมูล มาทำการคัดเลือกคุณลักษณะที่สำคัญ เพื่อเปรียบเทียบ และเลือกเทคนิคที่เหมาะสมบนพื้นฐานเงื่อนไขแบบหลายเป้าหมาย งานวิจัยนี้เลือกใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ได้แก่

1. วิธีแบบกรอง (Information Gain, Gain Ratio, Chi-Square)
2. วิธีแบบควบรวม (Forward Selection, Backward Elimination)
3. การเลือกแบบจำลองการเรียนรู้

โดยนำข้อมูลที่ผ่านการจัดสมดุลและคัดเลือกคุณลักษณะแล้ว มาทำการสร้างแบบจำลองด้วยเทคนิค XGBoost Random Forest และ LightGBM ซึ่งเป็นเทคนิคที่มีความสามารถในการจัดการกับข้อมูลขนาดใหญ่และซับซ้อนได้ดี

4. การฝึกสอนแบบจำลอง

ข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลและคัดเลือกคุณลักษณะแล้ว จะถูกนำเข้าสู่กระบวนการฝึกสอนแบบจำลอง โดยใช้เทคนิค 10-Fold Cross-Validation เพื่อแบ่งข้อมูลออกเป็น 10 ส่วน สำหรับการฝึกสอนและทดสอบแบบหมุนเวียน ซึ่งช่วยลดปัญหาการเกิด Overfitting และเพิ่มความน่าเชื่อถือของประสิทธิภาพแบบจำลอง

5. การวัดประสิทธิภาพแบบจำลอง

งานวิจัยนี้วัดประสิทธิภาพแบบจำลอง โดยพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ซึ่งเป็นการคำนวณค่าจากตาราง Confusion Matrix และค่าพื้นที่ใต้เส้นโค้ง จากนั้นนำมาเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองต่าง ๆ

6. การปรับค่าพารามิเตอร์และการใช้งานการทำนายแบบจำลอง

ในงานวิจัยนี้ ดำเนินการปรับค่าพารามิเตอร์เพื่อเพิ่มประสิทธิภาพของแบบจำลอง โดยเลือกใช้เทคนิค เช่น การตั้งค่า $n_jobs = -1$ เพื่อให้ประมวลผลแบบขนานและลดเวลาในการฝึกสอนแบบจำลอง และการเรียนรู้แบบไล่ระดับ (Gradient Boosting) เพื่อหลีกเลี่ยงการเกิด Overfitting พร้อมทั้งเลือกใช้แบบจำลองที่มีความเหมาะสมกับข้อมูล เช่น เทคนิค LightGBM และเทคนิค Random Forest ซึ่งมีคุณสมบัติในการจัดการกับข้อมูลขนาดใหญ่และไม่สมดุลได้ดี

6. ผลการวิจัย (Results)

6.1 ผลการศึกษาและเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบ Filter Methods และ Wrapper Methods สำหรับข้อมูลสุขภาพที่เป็นแบบหลายงาน และมีความไม่สมดุลของข้อมูล

งานวิจัยนี้เปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบ Filter Methods และ Wrapper Methods ผลลัพธ์ที่ได้แสดงดัง Table 4.

Table 4. Comparison of Feature Selection Techniques

Feature Selection type	Method	Average					
		Accuracy	Precision	Recall	F1-score	Time(s)	AUC
Filter	Information Gain	0.8721	0.8959	0.8461	0.8697	99.56	0.9248
Filter	Gain Ratio	0.8726	0.8976	0.8439	0.8695	83.12	0.9248
Filter	Chi-Square	0.8629	0.8900	0.8323	0.8594	50.28	0.9109
Wrapper	Forward Selection	0.8729	0.9051	0.8377	0.8692	6,119.34	0.8437
Wrapper	Backward Elimination	0.8663	0.8855	0.8449	0.8641	4,755.67	0.9163

จาก Table 4. แสดงผลการศึกษาและเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบ Filter Methods และ Wrapper Methods สำหรับข้อมูลสุขภาพที่เป็นแบบหลายงาน และมีความไม่สมดุลของข้อมูล โดยสามารถสรุปผลได้ดังนี้

1. ประสิทธิภาพสูงสุดของแบบจำลอง จาก Table 4. พบว่า วิธีแบบ Filter Methods สามารถให้ค่าพื้นที่ใต้เส้นโค้งที่สูงกว่ากลุ่ม Wrapper อย่างชัดเจน โดยเทคนิคที่ให้ค่าสูงสุด คือ เทคนิค Gain Ratio ให้ค่าพื้นที่ใต้เส้นโค้ง อยู่ที่ 0.9248 ซึ่งสูงกว่าเทคนิค Forward Selection ที่ทำได้เพียง 0.8437

2. ประสิทธิภาพด้านเวลาและทรัพยากรที่ใช้ จาก Table 4. พบว่า วิธีแบบ Filter Methods มีเวลาในการประมวลผลเร็วกว่าวิธีแบบ Wrapper Methods โดยเทคนิคที่ใช้เวลาน้อยที่สุด อย่างเทคนิค Chi-Square ใช้เวลาเพียง 50.28 วินาที ในขณะที่เทคนิค Forward Selection ใช้เวลาสูงถึง 6,119.34 วินาที

โดยสรุป เทคนิค Forward Selection แม้จะให้ค่าความถูกต้องสูงสุด แต่เมื่อพิจารณาค่าพื้นที่ใต้เส้นโค้ง และประสิทธิภาพด้านเวลา พบว่าเทคนิค Gain Ratio มีความเหมาะสมในการพยากรณ์โรคติดต่อไม่เรื้อรังมากกว่า เนื่องจากให้ค่าพื้นที่ใต้เส้นโค้งที่สูงอยู่ที่ 0.9248 และใช้เวลาประมวลผลน้อยกว่าวิธีแบบ Wrapper Methods ซึ่งแสดงถึงความสามารถในการลดสัญญาณรบกวน และความเสี่ยงต่อการเกิด Overfitting ได้ดีกว่า

6.2 ผลพัฒนาแบบจำลองการพยากรณ์โรคไม่ติดต่อเรื้อรัง โดยการบูรณาการเทคนิคการคัดเลือกคุณลักษณะที่มีประสิทธิภาพสูงสุด ร่วมกับแบบจำลองการเรียนรู้ของเครื่อง

งานวิจัยนี้วัดประสิทธิภาพแบบจำลอง โดยพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าเรียกคืน ค่าประสิทธิภาพโดยรวม เวลา และค่าพื้นที่ใต้เส้นโค้ง ผลลัพธ์ที่ได้แสดงดัง Table 5.

จาก Table 5. แสดงผลการเปรียบเทียบการวัดประสิทธิภาพของแบบจำลองการเรียนรู้ของทั้ง 3 เทคนิค ได้แก่ XGBoost, Random Forest และ LightGBM ร่วมกับการใช้เทคนิคการคัดเลือกคุณลักษณะรูปแบบต่าง ๆ โดยสามารถสรุปผลได้ดังนี้

1. ประสิทธิภาพสูงสุดของแบบจำลอง จาก Table 5. พบว่า เทคนิค LightGBM ที่ใช้คุณลักษณะทั้งหมด โดยไม่มีการคัดเลือกคุณลักษณะ คือ เทคนิคที่ให้ประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง อยู่ที่ 0.8794 และค่าประสิทธิภาพโดยรวม อยู่ที่ 0.8777 ซึ่งสะท้อนให้เห็นว่า เทคนิค LightGBM มีความสามารถในการจัดการกับข้อมูลสุขภาพที่มีความซับซ้อนได้อย่างดีเยี่ยม ทั้งในแง่ของการจำแนกกลุ่มข้อมูลที่ถูกต้องและการรักษาสสมดุลระหว่างค่าความแม่นยำ และค่าเรียกคืน

Table 5. Comparative Analysis of Model Performance

Model	Feature Selection Method	Accuracy	Precision	Recall	F1-score	Time(s)	AUC
XGBoost	None	0.8704	0.8885	0.8526	0.8698	92.29	0.9286
	(Full Features)						
	Information Gain	0.8767	0.9056	0.8464	0.8740	124.92	0.9300
	Gain Ratio	0.8768	0.9058	0.8456	0.8740	126.03	0.9300
	Chi-Square	0.8727	0.9050	0.8373	0.8690	73.18	0.9150
	Forward Selection	0.8760	0.9101	0.8399	0.8725	904.25	0.8374
Random Forest	Backward Elimination	0.8762	0.9103	0.8400	0.8726	2,497.39	0.9279
	None	0.8737	0.8827	0.8651	0.8737	815.56	0.9248
	(Full Features)						
	Information Gain	0.8612	0.8783	0.8394	0.8584	91.02	0.9127
	Gain Ratio	0.8625	0.8828	0.8349	0.8581	40.51	0.9127
	Chi-Square	0.8583	0.8763	0.8347	0.8548	38.04	0.9007
LightGBM	Forward Selection	0.8661	0.8972	0.8294	0.8614	7,625.66	0.7993
	Backward Elimination	0.8626	0.8783	0.8433	0.8603	10,716.93	0.9125
	None	0.8794	0.9036	0.8545	0.8777	97.18	0.9334
	(Full Features)						
	Information Gain	0.8784	0.9037	0.8525	0.8766	82.75	0.9317
	Gain Ratio	0.8785	0.9041	0.8513	0.8763	82.82	0.9317
	Chi-Square	0.8577	0.8888	0.8248	0.8545	39.61	0.9170
	Forward Selection	0.8766	0.9080	0.8438	0.8737	9,828.12	0.8944
	Backward Elimination	0.8600	0.8678	0.8514	0.8595	1,052.69	0.9084

2. การคัดเลือกคุณลักษณะ จาก Table 5. พบว่า เทคนิค XGBoost ร่วมกับการคัดเลือกคุณลักษณะ ด้วยเทคนิค Gain Ratio ให้ค่าความถูกต้องสูงกว่าการไม่คัดเลือกคุณลักษณะ จากค่าความถูกต้องอยู่ที่ 0.8704 เป็น 0.8768 ในขณะที่เทคนิค Random Forest และเทคนิค LightGBM กลับให้ผลลัพธ์ที่ดีที่สุดเมื่อใช้ตัวแปรครบทั้งหมด

3. ประสิทธิภาพด้านเวลาและทรัพยากรที่ใช้ จาก Table 5. พบว่า เมื่อพิจารณาเวลาในการประมวลผลเทคนิค LightGBM ไม่เพียงแต่ให้ค่าความถูกต้องสูงสุดเท่านั้น แต่ยังใช้เวลาในการประมวลผลค่อนข้างน้อย (97.18 วินาที) โดยเมื่อเทียบกับเทคนิค Random Forest แล้วจะใช้เวลาในการประมวลผลนานกว่าถึง 815.56 วินาที นอกจากนี้ วิธีการคัดเลือกคุณลักษณะแบบ Wrapper Methods ได้แก่ Forward Selection และ Backward Elimination แม้จะเป็นการค้นหาที่ละเอียด แต่ก็กลับใช้เวลาในการประมวลผลสูงมากในทุกแบบจำลอง โดยเฉพาะอย่างยิ่งเทคนิค Random Forest ที่ใช้เวลาถึง 10,716 วินาที โดยที่ค่าความถูกต้องไม่ได้เพิ่มขึ้นอย่างมีนัยสำคัญ

โดยสรุป การเลือกใช้เทคนิค LightGBM ร่วมกับการใช้คุณลักษณะทั้งหมด มีประสิทธิภาพสูงสุดสำหรับชุดข้อมูลนี้ เนื่องจากให้ผลลัพธ์ที่แม่นยำมากที่สุดและมีความคุ้มค่าในแง่ของทรัพยากรเวลา ข้อมูลการทดลองนี้ยืนยันว่าสำหรับการทำนายปัจจัยเสี่ยงด้านสุขภาพ การรักษาความครอบคลุมของตัวแปรนำเข้ามีความสำคัญอย่างยิ่งต่อการรักษาประสิทธิภาพของแบบจำลองในการพยากรณ์โรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมอง อย่างไรก็ตาม หากจำเป็นต้องลดจำนวนคุณลักษณะ เพื่อเพิ่มประสิทธิภาพในการประมวลผล เทคนิค XGBoost ร่วมกับเทคนิค Gain Ratio ถือเป็นทางเลือกที่มีประสิทธิภาพสูง โดยให้ค่าความถูกต้องสูงถึง 0.8768 ซึ่งดีกว่าการใช้ตัวแปรทั้งหมดในอัลกอริทึมเดียวกันอย่างชัดเจน

และนอกจากนี้เพื่อให้เห็นถึงรายละเอียดของผลการคัดเลือกคุณลักษณะในแต่ละเป้าหมายการพยากรณ์โรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมอง งานวิจัยนี้ได้นำเสนอจำนวนและรายการคุณลักษณะที่ถูกคัดเลือก เหลือเพียงบางตัวที่มีความสำคัญสูงสุดต่อการพยากรณ์ สามารถแสดงรายละเอียดการคัดเลือกคุณลักษณะดัง Table 6.

Table 6. Feature Importance Results for LightGBM Utilizing the Full Feature Set

Target	Total Features	Selected Features	Accuracy	Top Contributing Features
Diabetes Mellitus	19	3	0.9054	BMI, Age, GenHlth
Hypertension	19	2	0.7565	BMI, Age
Stroke	19	3	0.9762	BMI, Age, Income

จาก Table 6. พบว่า เทคนิค LightGBM ที่ใช้คุณลักษณะทั้งหมดโดยไม่มีการคัดเลือกคุณลักษณะ สามารถลดจำนวนคุณลักษณะสำคัญจาก 19 คุณลักษณะ ของเป้าหมายโรคเบาหวาน เหลือเพียง 3 คุณลักษณะ ได้แก่ ดัชนีมวลกาย อายุ และประเมินสุขภาพตนเอง โรคความดันโลหิตสูง เหลือเพียง 2 คุณลักษณะ ได้แก่ ดัชนีมวลกายและอายุ และโรคหลอดเลือดสมอง เหลือเพียง 3 คุณลักษณะ ได้แก่ ดัชนีมวลกาย อายุ และระดับรายได้

Table 7. Feature Selection Results of XGBoost Integrated with Gain Ratio Technique

Target	Total Features	Selected Features	Accuracy	Top Contributing Features
Diabetes Mellitus	19	7	0.9035	BMI, Age, Income, PhysHlth, GenHlth, MentHlth, Education
Hypertension	19	2	0.7507	BMI, Age
Stroke	19	7	0.9762	BMI, Age, Income, PhysHlth, GenHlth, Education, MentHlth

จาก Table 7. พบว่า เทคนิค XGBoost ร่วมกับการคัดเลือกคุณลักษณะด้วยเทคนิค Gain Ratio สามารถลดจำนวนคุณลักษณะสำคัญจาก 19 คุณลักษณะ ของเป้าหมายโรคเบาหวาน เหลือเพียง 7 คุณลักษณะ ได้แก่ ดัชนีมวลกาย อายุ รายได้ สุขภาพภายใน 30 วัน ประเมินสุขภาพตนเอง สุขภาพจิตใน 30 วัน และระดับการศึกษา โรคความดันโลหิตสูง เหลือเพียง 2 คุณลักษณะ ได้แก่ ดัชนีมวลกายและอายุ และโรคหลอดเลือดสมอง เหลือเพียง 7 คุณลักษณะ ได้แก่ ดัชนีมวลกาย อายุ รายได้ สุขภาพภายใน 30 วัน ประเมินสุขภาพตนเอง ระดับการศึกษา และสุขภาพจิตใน 30 วัน ซึ่งสอดคล้องกับงานวิจัยของ Leite (2025) พบว่า ตัวแปรด้านอายุและดัชนีมวลกาย มีความสัมพันธ์อย่างชัดเจนกับรูปแบบสุขภาพของผู้ป่วยโรคเรื้อรัง โดยงานวิจัยระบุว่าปัจจัยเหล่านี้ส่งผลควบคู่ไปกับการประเมินสุขภาพตนเองและความท้าทายด้านสุขภาพจิต ซึ่งเมื่อนำมารวมกันเป็นปัจจัยด้านพฤติกรรมสุขภาพ จะเป็นตัวแปรสำคัญที่ช่วยเพิ่มความน่าเชื่อถือในการทำนายความเสี่ยง

6.3 ผลการคัดเลือกคุณลักษณะสำคัญของแต่ละเทคนิค โดยใช้ตัวชี้วัดทางสถิติที่เหมาะสมกับข้อมูลไม่สมดุล เช่น ค่าพื้นที่ใต้เส้นโค้ง และค่าความถูกต้อง

เมื่อทำการเปรียบเทียบค่าพื้นที่ใต้เส้นโค้ง และค่าความถูกต้องของแบบจำลองที่พัฒนาด้วยเทคนิคการเรียนรู้ของเครื่องทั้ง 3 เทคนิค ได้แก่ เทคนิค XGBoost, Random Forest และ LightGBM จะได้ผลลัพธ์สูงสุดแสดงดัง Table 8.

Table 8. Comparison of Feature Selection Techniques

Model+Method	Accuracy	Precision	Recall	F1-score	AUC
XGBoost + Gain Ratio	0.8768	0.9058	0.8456	0.8740	0.9300
Random Forest (Full Features)	0.8737	0.8827	0.8651	0.8737	0.9248
LightGBM (Full Features)	0.8794	0.9036	0.8545	0.8777	0.9334

จาก Table 8. พบว่า เทคนิค LightGBM ร่วมกับการใช้คุณลักษณะทั้งหมดให้ผลลัพธ์ที่ดีที่สุด โดยมีค่าความถูกต้อง อยู่ที่ 0.8794 และค่าพื้นที่ใต้เส้นโค้ง (AUC) อยู่ที่ 0.9289 ซึ่งสะท้อนถึงความสามารถในการพยากรณ์โรคได้อย่างแม่นยำ แต่หากพิจารณาการประมวลผลร่วมกับเทคนิคการคัดเลือกคุณลักษณะ พบว่า เทคนิค Gain Ratio ช่วยเพิ่มประสิทธิภาพ ให้กับแบบจำลองสูงสุด โดยให้ค่าความถูกต้อง อยู่ที่ 0.8768 และค่าพื้นที่ใต้เส้นโค้ง (AUC) สูงสุดอยู่ที่ 0.9300

7. สรุปผลการวิจัยและอภิปรายผลการวิจัย (Conclusion and Discussion)

งานวิจัยเรื่อง การเพิ่มประสิทธิภาพการคัดเลือกคุณลักษณะสำหรับการพยากรณ์ โรคไม่ติดต่อเรื้อรังแบบหลายงาน จากข้อมูลสุขภาพที่ไม่สมดุล สามารถสรุปผลและอภิปรายผลการวิจัยได้ ดังนี้

1. ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการคัดเลือกคุณลักษณะ ทั้งในรูปแบบ Filter Methods และ Wrapper Methods สำหรับข้อมูลสุขภาพที่เป็นแบบหลายงานและมีความไม่สมดุลของข้อมูล พบว่า การคัดเลือกคุณลักษณะแบบ Filter Methods ให้ผลลัพธ์ดีกว่าการคัดเลือกคุณลักษณะแบบ Wrapper Methods โดยเทคนิคที่ให้ค่าความถูกต้องสูงที่สุดคือ เทคนิค Gain Ratio ร่วมกับแบบจำลองเทคนิค XGBoost โดยค่าความถูกต้องอยู่ที่ 0.8768 และใช้เวลาในการประมวลผลน้อยกว่า ถึงแม้วิธีการคัดเลือกคุณลักษณะแบบ Wrapper Method ได้แก่ Forward Selection และ Backward Elimination จะเป็นวิธีการค้นหาที่ละเอียด แต่กลับใช้เวลาในการประมวลผลสูงมากในทุกแบบจำลอง โดยเฉพาะอย่างยิ่งในเทคนิค Random Forest ที่ใช้เวลาถึง 10,716 วินาที โดยที่ค่าความถูกต้องไม่ได้เพิ่มขึ้นอย่างมีนัยสำคัญ ซึ่งสอดคล้องกับงานวิจัยของ Tepdang (2023) โดยกล่าวถึงการพยากรณ์ผู้ป่วยโรคเบาหวาน ข้อมูลจากเว็บไซต์ www.kaggle.com ด้วยวิธีการเรียนรู้ของเครื่อง ผลงานวิจัยพบว่า วิธีที่ดีที่สุด คือ Gradient Boosted Trees ให้ประสิทธิภาพสูงสุดที่ร้อยละ 87.14 และจากการคัดเลือกคุณลักษณะด้วยวิธี Filter Approach ตามวิธีโครงสร้างต้นไม้ ใช้เพียง 4 ปัจจัยหลัก ได้แก่ ระดับกลูโคส อายุ จำนวนครั้งที่ตั้งครรภ์ และระดับอินซูลิน ก็สามารถจำแนกผู้ป่วยได้อย่างมีประสิทธิภาพเพื่อการรักษาที่รวดเร็วขึ้น และสอดคล้องกับงานวิจัยของ Pongshaing & Thongkam (2023) ซึ่งพยากรณ์การรักษาโรคความดันโลหิตสูง ด้วยการคัดเลือกคุณลักษณะเทคนิค Gain Ratio เป็นวิธีที่สามารถเพิ่มค่า ความจำเพาะและค่าความถูกต้องของแบบจำลองที่สร้างด้วยเทคนิคเคเนียร์เรสเนเบอร์ได้ดีที่สุด โดยสามารถเพิ่มค่าความจำเพาะ และค่าความถูกต้องได้ร้อยละ 15.35 และ 3.72 ตามลำดับ

2. ผลการคัดเลือกคุณลักษณะสำคัญของแต่ละเทคนิค โดยใช้ตัวชี้วัดทางสถิติที่เหมาะสมกับข้อมูลไม่สมดุล เช่น ค่าพื้นที่ใต้กราฟ (AUC) และค่าความแม่นยำ (Accuracy) พบว่า เมื่อพิจารณาค่าพื้นที่ใต้เส้นโค้ง (AUC) ซึ่งเป็นตัวชี้วัดที่สะท้อนประสิทธิภาพของแบบจำลองได้ดีกว่าค่าความถูกต้องเพียงอย่างเดียว ทำให้เห็นว่าแบบจำลองเทคนิค LightGBM ร่วมกับการคัดเลือกคุณลักษณะด้วยเทคนิค Information Gain และเทคนิค Gain Ratio สามารถเพิ่มประสิทธิภาพให้กับแบบจำลองสูงสุด โดยให้ค่าพื้นที่ใต้เส้นโค้ง (AUC) สูงสุดอยู่ที่ 0.9317 ซึ่งสอดคล้องกับงานวิจัยของ Rufo et al. (2021) ที่พัฒนาแบบจำลองวินิจฉัยโรคเบาหวานโดยใช้เทคนิค LightGBM กับข้อมูลผู้ป่วยจากโรงพยาบาลในเอธิโอเปีย เพื่อช่วยคัดกรองโรคในพื้นที่ที่ขาดแคลนแพทย์ผู้เชี่ยวชาญ พบว่า เทคนิค LightGBM ให้ประสิทธิภาพสูงมาก โดยมีค่าความถูกต้อง และค่าพื้นที่ใต้เส้นโค้ง (AUC) เท่ากับ 98.1% และให้ผลดีกว่าแบบจำลองเทคนิค KNN, SVM, Random Forest และ XGBoost อย่างชัดเจน เหมาะสำหรับการประยุกต์ใช้ในระบบสาธารณสุขที่มีข้อจำกัดด้านทรัพยากร

8. ข้อเสนอแนะงานวิจัย (Recommendation)

8.1 ข้อเสนอแนะในงานวิจัยครั้งต่อไป

1. ควรพิจารณาเลือกใช้ชุดข้อมูลสุขภาพที่มีการจัดเก็บภายในระยะเวลาไม่เกิน 5 ปี เพื่อให้ผลการวิเคราะห์สะท้อนสถานการณ์ด้านสาธารณสุขในปัจจุบันได้อย่างถูกต้อง ทั้งนี้ ชุดข้อมูล Diabetes Health Indicators ที่ใช้ในงานวิจัยครั้งนี้จัดเก็บจากแหล่งข้อมูลสาธารณะ อาจมีข้อจำกัดด้านความทันสมัยของข้อมูลบางส่วน

2. ควรมีการนำแบบจำลองการเรียนรู้ที่พัฒนาขึ้นไปทดสอบกับชุดข้อมูลสุขภาพจริงจากโรงพยาบาลในประเทศไทย เพื่อประเมินความแม่นยำและความสามารถในการนำไปประยุกต์ใช้ในบริบทของประชากรไทย ที่อาจมีปัจจัยเสี่ยงแตกต่างจากข้อมูลใน Kaggle

3. ควรศึกษาและเปรียบเทียบผลการคัดเลือกคุณลักษณะโดยใช้เทคนิคแบบผสมผสาน (Hybrid Feature Selection) ที่รวมจุดเด่นของวิธีแบบกรอง (Filter Methods) และแบบห่อหุ้ม (Wrapper Methods) เข้าด้วยกัน เพื่อเพิ่มความแม่นยำ ลดความซับซ้อนของแบบจำลอง และเพิ่มประสิทธิภาพในการประมวลผลของระบบได้อย่างมีนัยสำคัญ

8.2 การประยุกต์ใช้ในงานวิจัย

ผลงานวิจัยสามารถนำไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจทางการแพทย์ (Clinical Decision Support System: CDSS) ที่เชื่อมโยงกับฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์ (Electronic Health Record: EHR) โดยแบบจำลองการพยากรณ์ที่ผ่านการเลือกคุณลักษณะและการปรับสมดุลข้อมูลแล้ว สามารถนำมาใช้ประเมินความเสี่ยงของผู้ป่วยโรคเบาหวาน โรคความดันโลหิตสูง และโรคหลอดเลือดสมองแบบอัตโนมัติ ซึ่งจะช่วยให้บุคลากรทางการแพทย์สามารถคัดกรองผู้ป่วยกลุ่มเสี่ยงได้อย่างรวดเร็วและแม่นยำ พร้อมทั้งสนับสนุนการวางแผนป้องกันและการรักษาเชิงรุกได้อย่างมีประสิทธิภาพมากขึ้น นอกจากนี้ โรงพยาบาลหรือหน่วยบริการสุขภาพสามารถเชื่อมต่อกับระบบดังกล่าวเข้ากับโครงสร้างสารสนเทศที่มีอยู่เดิม เพื่อพัฒนาเป็นเครื่องมือวิเคราะห์ข้อมูลสุขภาพที่ใช้งานได้จริง และต่อยอดสู่ระบบบริการสุขภาพอัจฉริยะในอนาคต

9. กิตติกรรมประกาศ (Acknowledgement)

งานวิจัยฉบับนี้สำเร็จลงไปได้ด้วยดี ด้วยความกรุณาและความช่วยเหลือเป็นอย่างดียิ่งจาก บุคลากรจากวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนภิเษก ที่ให้ความช่วยเหลือคำแนะนำ และสนับสนุน หวังเป็นอย่างยิ่งว่า การศึกษาค้นคว้าครั้งนี้จะเป็นประโยชน์แก่ผู้นำไปศึกษาค้นคว้าต่อไป

10. การเปิดเผยการใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Declaration of Generative AI)

งานวิจัยนี้ได้ใช้ปัญญาประดิษฐ์ในการแปลบทความภาษาไทยให้เป็นภาษาอังกฤษ โดยใช้โปรแกรมปัญญาประดิษฐ์ เช่น ChatGPT, Perplexity, Claude ซึ่งใช้คำสั่ง ให้แปลบทความข้างต้นเป็นภาษาอังกฤษอย่างถูกต้อง ตามหลักวิชาการและถูกหลักไวยากรณ์

11. การแสดงบทบาทผู้เขียน (CRediT Author Statement)

Yanisa Wongsombut: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Visualization, Writing-Original Draft, Writing-Review & Editing.

Nithizethe Mhuadthongon: Supervision, Validation, Writing - Review & Editing.

All authors read and approved the final manuscript.

12. เอกสารอ้างอิง (References)

Jupriyadi, Budiman, A., Hamidi, E. A. Z., Ahdan, S., & Negara, R. M. (2024, July 4-5). Wrapper-Based Feature Selection to Improve the Accuracy of Intrusion Detection System (IDS). *2024 10th International Conference on Wireless and Telematics*, 1-5. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/icwt62080.2024.10674687>

- Khan, R. A. (2023). Resilience Family of Receiver Operating Characteristic Curves. *IEEE Transactions on Reliability*, 72(2), 716-726. <https://doi.org/10.1109/tr.2022.3194710>
- Leite, Â. (2025). Chronic Illnesses: Varied Health Patterns and Mental Health Challenges. *Healthcare*, 13(12), 1396. <https://doi.org/10.3390/healthcare13121396>
- Li, W., Peng, Y., & Peng, K. (2024). Diabetes Prediction Model Based on GA-XGBoost and Stacking Ensemble Algorithm. *PLOS One*, 19(9), e0311222. <https://doi.org/10.1371/journal.pone.0311222>
- Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the Impact of Feature Selection Methods on Machine Learning Algorithms for Heart Disease Prediction. *Scientific Reports*, 13, 22588. <https://doi.org/10.1038/s41598-023-49962-w>
- Pongshaing, T., & Thongkam, J. (2023). Optimization of Models for Hypertension Treatment Prediction with Factor Selection. *Journal of Science and Technology, Ubon Ratchathani University*, 25(1), 13-20. (In Thai)
- Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports*, 14, 6086. <https://doi.org/10.1038/s41598-024-56706-x>
- Romsaiyud, W. (2024). Fast Synthesis of the Minority Class Using Generative Adversarial Networks for Imbalanced Data Classification Problems. *Journal of Science and Technology Mahasarakham University*, 43(2), 108-121. (In Thai)
- Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics*, 11(9), 1714. <https://doi.org/10.3390/diagnostics11091714>
- Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. (2023). An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. *International Journal of Computational Intelligence Systems*, 16(1), 14. <https://doi.org/10.1007/s44196-023-00184-y>
- Salhi, A., Henslee, A. C., Ross, J., Jabour, J., & Dettwiller, I. (2023). Data Preprocessing Using AutoML: A Survey. *2023 Congress in Computer Science, Computer Engineering, & Applied Computing*, 1619-1623. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/csce60160.2023.00265>
- Sunggad, S., & Maneerat, P. (2023). Comparison of Feature Selection Methods to Improve Diabetes Predictions. *Journal of Science and Technology Thonburi University*, 7(2), 12-24. (In Thai)
- Teboul, A. (n.d.). Diabetes Health Indicators Dataset. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>
- Tepdang, S. (2023). The Classification of Diabetic Patients Using Machine Learning Method by Feature Selection. *RMUTSB Academic Journal*, 11(1), 29-44. (In Thai)
- World Health Organization. (2023). *Global Report on Hypertension: The Race Against a Silent Killer*. <https://www.who.int/teams/noncommunicable-diseases/hypertension-report>