JIT Mahasarakham University, Thailand
http://jit.it.msu.ac.th

Journal of Applied Informatics and Technology, 7(2), 432-447.

Research Article

การคัดเลือกคุณลักษณะด้วยการวิเคราะห์แยกแยะเชิงเส้นเพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลผู้ป่วยโรคหลอดเลือดหัวใจ

Feature Selection with Linear Discriminant Analysis to Improve the Performance of Hearth Disease Classification

รติพร จันทร์กลั่น[1,*], กีระชาติ สุขสุทธิ์[1], เกตุกาญจน์ โพธิจิตติกานต์[1]
Ratiporn Chanklan[1,*], Keerachart Suksut[1], Kedkarn Podhijittikarn[1]

[1] สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน นครราชสีมา 30000 ประเทศไทย
[1] Computer Engineering, Faculty of Engineering and Technology, Rajamangala University of Technology Isan, Nakhon Ratchasima 30000, Thailand
* Corresponding author: Ratiporn Chanklan, ratiporn.ch@rmuti.ac.th

Received:
19 April 2024
Revised:
1 July 2024
Accepted:
19 September 2024

คำสำคัญ:

โรคหลอดเลือดหัวใจ, การคัดเลือกคุณลักษณะ, การวิเคราะห์แยกแยะเชิงเส้น

Keywords:

Heart Disease, Feature Selection, Linear Discriminant Analysis

บทคัดย่อ: เทคโนโลยีปัญญาประดิษฐ์ได้รับความนิยมอย่างแพร่หลายในการประยุกต์ใช้ในหลากหลายสาขา รวมถึงในด้านการแพทย์ ซึ่งถูกนำมาใช้สนับสนุนการวินิจฉัยโรคอย่างมีประสิทธิภาพ โดยเฉพาะโรคหลอดเลือดหัวใจซึ่งเป็นโรคที่สามารถเกิดได้กับทุกเพศ ทุกวัย และทุกเชื้อชาติ อีกทั้งยังเป็นสาเหตุสำคัญของการเสียชีวิตในปัจจุบัน การวินิจฉัยโรคหลอดเลือดหัวใจสามารถดำเนินการร่วมกับเทคโนโลยีปัญญาประดิษฐ์ โดยใช้ข้อมูลจากการตรวจคลื่นไฟฟ้าหัวใจร่วมกับอัลกอริทึมการเรียนรู้ของเครื่อง อย่างไรก็ตาม ข้อมูลที่ใช้ในการวิเคราะห์มักประกอบด้วยคุณลักษณะจำนวนมากเกินความจำเป็น ซึ่งอาจส่งผลต่อประสิทธิภาพของโมเดล งานวิจัยนี้จึงเสนอการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์แยกแยะเชิงเส้นเพื่อเพิ่มความแม่นยำในการจำแนกผู้ป่วยโรคหลอดเลือดหัวใจ พร้อมเปรียบเทียบกับเทคนิคการคัดเลือกคุณลักษณะโดยใช้ค่าสหสัมพันธ์ และค่าเกนความรู้ โดยประเมินผลผ่านการสร้างโมเดลด้วย 3 อัลกอริทึม ได้แก่ การถดถอยโลจิสติก ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ผลการทดลองพบว่า การใช้เทคนิคการวิเคราะห์แยกแยะเชิงเส้นช่วยเพิ่มค่าเฉลี่ยความแม่นยำจากร้อยละ 77.82 เป็นร้อยละ 86.46 (เพิ่มขึ้นร้อยละ 11.10) และเมื่อใช้ร่วมกับโครงข่ายประสาทเทียมสามารถจำแนกข้อมูลได้อย่างแม่นยำสูงสุดที่ร้อยละ 87.39 จากผลการศึกษา ผู้วิจัยจึงพัฒนาโปรแกรมประเมินความเสี่ยงโรคหลอดเลือดหัวใจโดยใช้เทคนิคดังกล่าว ซึ่งสามารถใช้เป็นเครื่องมือช่วยคัดกรองบุคคลที่มีความเสี่ยงสูง พร้อมทั้งให้ข้อมูลเกี่ยวกับโอกาสการเกิดโรคหลอดเลือดหัวใจในแต่ละบุคคลได้อย่างมีประสิทธิภาพ

Abstract: Artificial intelligence (AI) technology has become increasingly popular and is widely applied across various fields. In the medical domain, AI has been employed to support disease diagnosis. Heart disease is a common condition that affects individuals of all genders, ages, and races, and remains a leading cause of mortality worldwide. Currently, the diagnosis of heart disease can be performed using AI by leveraging electrocardiogram (ECG) data in combination with machine learning algorithms. However, in some cases, the number of data features required is excessive, which may reduce model performance. In this research, we propose a feature selection method based on Linear Discriminant Analysis (LDA) to improve the classification accuracy of a heart disease dataset. The proposed method is compared with two other feature selection techniques: correlation-based selection and information gain. We then construct classification models using three algorithms: logistic regression, support vector machines (SVM), and artificial neural networks (ANN). The experimental results show that the proposed technique improves the average classification accuracy from 77.82% to 86.46%, representing an 11.10% increase. The highest classification accuracy of 87.39% is achieved when combining ANN with LDA. The researcher employed this technique to develop a program for assessing the risk of coronary heart disease. The program assists in screening individuals at high risk and provides users with personalized information regarding their likelihood of developing the disease.

1. บทนำ

โรคหลอดเลือดหัวใจเป็นหนึ่งในโรคที่ทำให้ผู้คนเสียชีวิตมากที่สุดทั่วโลก การวินิจฉัยโรคได้อย่างแม่นยำ รวมไปถึงการรู้ล่วงหน้าถึงความเสี่ยงที่มีต่อโรคหลอดเลือดหัวใจสามารถช่วยลดความเสี่ยงในการเสียชีวิตโดยการปรับเปลี่ยนพฤติกรรมที่ส่งผลต่อสุขภาพหัวใจ เช่น การบริโภคอาหาร การออกกำลังกาย และยังช่วยให้ได้รับการรักษาที่เหมาะสมได้ ซึ่งโดยทั่วไปมักจะใช้แพทย์ผู้เชี่ยวชาญในการวินิจฉัยด้วยการตรวจพื้นฐาน เช่น การซักประวัติคนไข้ สอบถามอาการ การเอกซ์เรย์หัวใจ รวมไปถึงใช้วิธีการตรวจคลื่นไฟฟ้าหัวใจ (Electrocardiogram: ECG) เพื่อบันทึกและวิเคราะห์คลื่นไฟฟ้าที่เกิดขึ้นในขณะที่หัวใจทำงาน ซึ่งเป็นวิธีที่ใช้กันอย่างแพร่หลายในการตรวจวินิจฉัยโรคหัวใจ

ในขณะที่เทคโนโลยีทางด้านปัญญาประดิษฐ์ (Artificial Intelligence: AI) สามารถนำมาประยุกต์ใช้ให้คอมพิวเตอร์สามารถประมวลผล วิเคราะห์ผล และจำแนกผลได้ตามแต่ละประเภทที่นำไปประยุกต์ใช้งาน ซึ่งในทางการแพทย์ก็ได้มีการนำเทคโนโลยีด้านปัญหาประดิษฐ์เข้ามาช่วยด้านการวินิจฉัยโรค การบำบัด รวมไปถึงการแนะนำวิธีการรักษาเบื้องต้น การใช้ปัญญาประดิษฐ์เข้ามาประยุกต์ใช้จำแนกโรคหัวใจก็เป็นหนึ่งในแขนงที่นักวิจัยมักต้องการท้าทายเพื่อเพิ่มประสิทธิภาพด้านความแม่นยำในการจำแนกด้วยการใช้เทคโนโลยีคอมพิวเตอร์ด้านอัลกอริทึม หรือกระบวนการในการแก้ปัญหาต่างๆ เข้ามาประยุกต์ใช้ร่วมกับปัญญาประดิษฐ์เพื่อให้การจำแนกโรคหัวใจมีความแม่นยำเพิ่มมากขึ้นซึ่งกำลังเป็นที่นิยมในปัจจุบัน ดังงานวิจัยดังต่อไปนี้

Radhika & George (2021) ใช้การเรียนรู้ของเครื่องเพื่อทำนายแนวโน้มการเกิดโรคหลอดเลือดหัวใจโดยใช้อัลกอริทึม K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes

(NB), Logistic Regression (LR), Decision Tree (DT) และ Random Forest (RF) ในการทดลองใช้ข้อมูลโรคหัวใจจาก UCI (University of California, Irvine C.A) ซึ่งมีจำนวนข้อมูล 303 ข้อมูล มีจำนวน 12 คุณสมบัติ ผลการทดลองพบว่า KNN และ RF ให้ประสิทธิภาพดีที่สุดเท่ากัน โดยมีความแม่นยำที่ร้อยละ 88.52

Chowdhury *et al*. (2021) ใช้ข้อมูลที่รวบรวมข้อมูลโดยใช้แบบสอบถาม ซึ่งมีผู้ให้ข้อมูล 564 คน โดยเป็นผู้ป่วยโรคหลอดเลือดหัวใจทั้งหมด 313 คน และคนที่มีสุขภาพดีอยู่ที่ 251 คน ในงานวิจัยใช้อัลกอริทึม DT, LR, KNN, NB และ SVM โดยอัลกอริทึมที่ให้ประสิทธิภาพสูงสุดด้วยความแม่นยำร้อยละ 91 คือ SVM

Lakshmi & Devi (2023) ใช้ชุดข้อมูลโรคหลอดเลือดหัวใจจากเว็บไซต์ Kaggle เพื่อวิเคราะห์โรคหัวใจ ข้อมูลประกอบด้วย 4238 รายการข้อมูล มีจำนวน 16 คุณสมบัติ มีการเตรียมข้อมูลด้วยการลบข้อมูลที่ไม่เหมาะสมออกจากชุดข้อมูล จากนั้นจึงใช้อัลกอริทึม Whale Optimization Algorithm (EWOA) เพื่อเลือกคุณสมบัติที่เหมาะสมก่อนนำไปสร้างโมเดลที่ชำนายโรคหลอดเลือดหัวใจ ในงานวิจัยใช้อัลกอริทึม SVM, RF, DT, LR และ KNN, SVM-RF (HSVRF) และ SVM-KNN (HSVKN) วิธีการจำแนกประเภทด้วยอัลกอริทึม HSVRF ให้ผลลัพธ์ที่ดีที่สุดด้วยความแม่นยำร้อยละ 85.79

Kavitha *et al*. (2021) ใช้วิธีการสร้างโมเดลไฮบริด ซึ่งเป็นเทคนิคใหม่ที่มีการระบุความน่าจะเป็นที่มาจากโมเดลการเรียนรู้ของเครื่องหนึ่งเป็นอินพุตให้กับโมเดลการเรียนรู้ของเครื่องอื่น ในงานวิจัยนี้ใช้ข้อมูลจาก UCI โดยอัลกอริทึมที่ใช้ประกอบด้วย DT, RF และ Hybrid (DT ร่วมกับ RF) ผลการทดลองพบว่าวิธีการที่นำเสนอให้ผลลัพธ์ที่ดีที่สุดด้วยความแม่นยำร้อยละ 88

Imanbek, Buribayev, & Yerkos (2023) ใช้ข้อมูลโรคหลอดเลือดหัวใจที่มี 12 คุณสมบัติ จำนวน 1,190 ข้อมูล มีการเติมข้อมูลสูญหายด้วยวิธี KNNImputer ซึ่งใช้วิธี One Hot Encoding กับข้อมูล Chest pain type, Resting ecg และ ST slope เพื่อแปลงข้อมูลแยกคุณสมบัติให้เป็น คุณสมบัติย่อย แบบไบนารีตามค่าจริงของข้อมูล ในงานวิจัยใช้อัลกอริทึม RF, XGBoost (XGB) และ Light Gradient Boosting Machine (LGBM) มีการปรับแต่งไฮเปอร์พารามิเตอร์เพื่อผลลัพธ์ที่ดีที่สุดโดยใช้วิธี GridSearchCV และประเมินประสิทธิภาพของโมเดลด้วยวิธี 5-fold cross-validation มีการคัดเลือกคุณสมบัติโดยเลือกใช้ 5 คุณสมบัติจากการพิจารณาของ RF, XGB และ LGBM ผลการทดลอง LGBM ให้ประสิทธิภาพสูงสุดให้ค่าเฉลี่ย ROC score (AUC=0.95)

Modak, Abdel-Raheem, & Rueda (2022) ใช้ข้อมูลโรคหลอดเลือดหัวใจที่มี 14 คุณสมบัติ จำนวน 1,190 ข้อมูล ใช้กระบวนการ Infinite Feature Selection เพื่อคัดเลือกคุณสมบัติ และใช้อัลกอริทึม Deep Neural Networks ในการสร้างโมเดลจำแนก และใช้ 5-fold cross-validation ในการประเมินประสิทธิภาพโมเดล ซึ่งให้ค่าความแม่นยำเฉลี่ยที่ร้อยละ 87.70

Kadhim & Radhi (2023) ใช้ข้อมูลโรคหลอดเลือดหัวใจที่มี 12 คุณสมบัติ จำนวน 1,190 ข้อมูล มีการลบข้อมูลที่มีบางคุณสมบัติไม่มีข้อมูล และลบข้อมูลสัญญาณรบกวนด้วยการพิจารณาด้วย Boxplots แบ่งข้อมูลฝึก (Training Set) 80% และข้อมูลทดสอบ (Test Set) 20% สร้างโมเดลโดยใช้ RF, SVM, KNN และ DT มีการหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละโมเดลด้วย Random Search Optimization ผลการทดลองพบว่า RF ให้ประสิทธิภาพค่าความแม่นยำสูงสุดที่ร้อยละ 94.9

จะเห็นได้ว่าการเพิ่มประสิทธิภาพการจำแนกโรคหลอดเลือดหัวใจมักจะใช้อัลกอริทึมด้านปัญญาประดิษฐ์มาใช้ในการจำแนกร่วมกัน การใช้การคัดเลือกคุณสมบัติที่เหมาะสมของข้อมูลทำให้ได้โมเดลที่มีประสิทธิภาพมากกว่าการใช้ข้อมูลทั้งหมด เนื่องจากคุณลักษณะของข้อมูลบางประการไม่มีความจำเป็นที่ใช้ในการจำแนก ดังนั้นงานวิจัยนี้ใช้เทคนิคการคัดเลือกคุณสมบัติด้วยการวิเคราะห์แยะแยะเชิงเส้นซึ่งเป็นวิธีการคัดเลือกคุณสมบัติโดยการพยายามรักษาข้อมูลที่ใช้ในการแยกคลาส (Class) ให้มากที่สุดและทำให้คลาสต่างๆ ถูกแยกออกจากกันมากที่สุด โดยจะหาทิศทางการวางข้อมูลในมิติใหม่ที่ทำให้ข้อมูลของคลาสต่างๆ มีความแปรปรวนภายในคลาสต่ำสุดและความแปรปรวนระหว่างคลาสสูงสุด เปรียบเทียบกับวิธีการที่มีแนวคิดของการคัดเลือกเลือกคุณสมบัติของข้อมูลที่แตกต่างกัน ได้แก่ ค่าสหสัมพันธ์ (Correlation-based Selection) และค่าเกนความรู้ (Information Gain) โดยค่าสหสัมพันธ์จะตัดข้อมูลที่มีความซ้ำซ้อนกันออก ค่าเกนความรู้จะคัดเลือกคุณสมบัติที่สามารถแบ่งคลาสของข้อมูลได้ดี ซึ่งจะเห็นว่าการคัดเลือกคุณสมบัติด้วยการวิเคราะห์แยะแยะเชิงมีการมีการมองข้อมูลในมิติใหม่และมีเงื่อนไขในการคัดเลือกคุณสมบัติที่ซับซ้อนกว่าค่าสหสัมพันธ์ และค่าเกนความรู้ ในงานวิจัยนี้ใช้เทคโนโลยีปัญญาประดิษฐ์เพื่อสร้างโมเดลที่ใช้สำหรับการจำแนกข้อมูลผู้ป่วยโรคหลอดเลือดหัวใจ ได้แก่ การถดถอยโลจีสติกส์ (Logistic Regression :LR) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

2. ปริทัศน์และวรรณกรรมที่เกี่ยวข้อง

2.1 ค่าสหสัมพันธ์ (Correlation Coefficient)

ค่าสหสัมพันธ์ เป็นกระบวนการสำหรับหาความสัมพันธ์ระหว่างตัวแปรสองตัว โดยพิจารณาว่าตัวแปรที่มาจากแหล่งเดียวกันมีความสัมพันธ์หรือมีความแปรปรวนร่วมกันมากหรือน้อยเพียงใด ซึ่งความแปรปรวมร่วมเป็นตัวชี้วัดความเปลี่ยนแปลงของตัวแปรทั้งสองมีการเปลี่ยนแปลงตามกันมากน้อยเพียงใด โดยพบว่า หากมีความแปรปรวนร่วมกันมาก หมายความว่าตัวแปรทั้งสองตัวนั้นมีความสัมพันธ์กันสูงมาก ในขณะที่หากตัวแปรทั้งสองตัวมีความแปรปรวนร่วมกันน้อย สามารถตีความได้ว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันต่ำ และหากตัวแปรทั้งสองตัวไม่มีความสัมพันธ์กันก็จะส่งผลให้ตัวแปรทั้งสองไม่มีความแปรปรวนร่วมกัน

การหาค่าสหสัมพันธ์แบบเพียร์สัน (Pearson Product Moment Correlation) จะใช้หาความสัมพันธ์ระหว่างตัวแปรสองตัวที่มีความสัมพันธ์เป็นเส้นตรง (Linear Relationship) โดยสามารถหาค่าสหสัมพันธ์แบบเพียร์สันได้ดังสมการที่ 1

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \quad (1)$$

โดยที่

r คือ ค่าสหสัมพันธ์

x_v, y_i คือ ค่าของตัวแปร x และ y ลำดับที่ i ในชุดข้อมูล

$\bar{x}$, $\bar{y}$ คือ ค่าเฉลี่ยของตัวแปร x และ y

ค่าสหสัมพันธ์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยเครื่องหมาย + และ - จะแสดงทิศทางของความสัมพันธ์กัน หากความสัมพันธ์มีค่าเป็นบวก หมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์ในเชิงแปรผันร่วมกัน แต่หากความสัมพันธ์เป็นในทิศทางตรงกันข้าม หรือผกผันกัน จะส่งผลให้ค่าสหสัมพันธ์มีค่าเป็นลบ สำหรับการแปรความหมายเชิงปริมาณของ

ตาราง 1 การแปลความหมายปริมาณของค่าสหสัมพันธ์

ค่าสหสัมพันธ์	การตีความ
0.0 – 0.09	มีความสัมพันธ์เล็กน้อย
0.10 – 0.39	มีความสัมพันธ์ต่ำ
0.40 - 0.69	มีความสัมพันธ์ปานกลาง
0.70 – 0.89	มีความสัมพันธ์สูง
0.90 – 1.00	มีความสัมพันธ์สูงมาก

ค่าสหสัมพันธ์ (Schober, Boer, & Schwarte, 2018) แสดงดังตาราง 1

2.2 การวิเคราะห์แยกแยะเชิงเส้น (Linear Discriminant Analysis: LDA)

การวิเคราะห์แยกแยะเชิงเส้น เป็นอัลกอริทึมการเรียนรู้ของเครื่องที่ใช้ในการจำแนกข้อมูลตั้งแต่ 2 กลุ่มข้อมูลขึ้นไป โดยจะใช้เทคนิคการเรียนรู้แบบมีผู้ฝึกสอน ซึ่งหลักการทำงานจะคล้ายคลึงกับการคัดเลือกคุณลักษณะด้วยการวิเคราะห์องค์ประกอบหลัก (Principle Component Analysis: PCA) เพียงแต่การวิเคราะห์แยกแยะเชิงเส้นจะนำคำอธิบาย (Label) ของข้อมูลมาพิจารณาร่วมด้วย และปรับปรุงเงื่อนไขสำหรับการหาเมทริกซ์ด้วยการหาค่าความแปรปรวนระหว่างกลุ่ม (Covariance between Group: S_B) กับค่าความแปรปรวนร่วมภายในกลุ่ม (Covariance within Group: S_W) สูงสุด แสดงดังสมการที่ 2 และ 3

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \quad (2)$$

$$S_B = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^T \quad (3)$$

โดยที่

c คือ จำนวนกลุ่ม

x_{ij} คือ ข้อมูลตัวที่ i ในกลุ่มที่ j

μ_i คือ ค่าเฉลี่ยของข้อมูลแต่ละกลุ่ม

μ คือ ค่าเฉลี่ยของข้อมูลทั้งหมด

เมื่อหาอัตราส่วนระหว่างค่าความแปรปรวนระหว่างกลุ่มและค่าความแปรปรวนร่วมภายในกลุ่มแล้วจึงนำค่าที่ได้ไปหาค่าไอเกน (Eigenvalues) (Tharwat *et al*., 2017) แสดงดังสมการที่ 4 โดยค่าไอเกนจะใช้บ่งบอกถึงความสำคัญของแต่ละคุณสมบัติในชุดข้อมูล ซึ่งจะนำค่านี้มาพิจารณาเพื่อลดมิติข้อมูล

$$S_w w = \lambda S_B w \quad (4)$$

โดยที่

S_w คือ ค่าความแปรปรวนร่วมภายในกลุ่ม

S_B คือ ค่าความแปรปรวนระหว่างกลุ่ม

w หาค่าได้จาก $w = S_w^{-1} S_B$

λ คือ ค่าไอเกน

2.3 การถดถอยโลจีสติกส์ (Logistic Regression: LR)

การถดถอยโลจีสติกส์เป็นเทคนิควิเคราะห์ทางสถิติที่ใช้สำหรับการทำนายเหตุการณ์ที่สนใจ

ว่าจะเกิดขึ้นหรือไม่ มีหลักการทำงานคล้ายคลึงกับการวิเคราะห์การถดถอยเชิงเส้น การถดถอยโลจีสติกส์จะทำการวิเคราะห์หาความสัมพันธ์ของตัวแปรในรูปแบบความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (Stoltzfus, 2011) โดยการวิเคราะห์การถดถอยโลจีสติกส์จะประกอบไปด้วยตัวแปรทำนาย (ตัวแปรต้น) และตัวแปรผล (ตัวแปรตอบสนอง)

ในการวิเคราะห์การถดถอยโลจีสติกส์ จะใช้ตัวแปรทำนายเพื่อทำนายโอกาสในการเกิดตัวแปรผล โดยอาศัยโอกาสความน่าจะเป็นของตัวแปรทำนายเพื่อหาความน่าจะเป็นที่จะเกิดค่าแต่ละค่าของตัวแปรผล ในกรณีที่มีตัวแปรทำนาย 1 ตัวจะเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นโลจีสติกส์อย่างง่าย (Simple Logistic Regression) แต่หากมีตัวแปรทำนายมากกว่า 1 ตัวขึ้นไปจะเรียกว่า การวิเคราะห์การถดถอยโลจีสติกส์เชิงพหุ (Multiple Logistic Regression) เมื่อพิจารณาที่ตัวแปรผลถ้าตัวแปรผลมีค่าที่เป็นไปได้เพียงสองค่าเท่านั้น เช่น การศึกษาการเป็นโรคหลอดเลือดหัวใจ ตัวแปรผล y มีค่าเป็น 1 คือป่วย และ y มีค่าเป็น 0 คือไม่ป่วย จะเรียกว่า การวิเคราะห์การถดถอยโลจีสติกส์แบบสองกลุ่ม (Binary Logistic Regression) แต่ถ้าตัวแปรผลมีค่ามากกว่า 2 ค่า จะเรียกว่า การวิเคราะห์การถดถอยโลจีสติกส์แบบหลายกลุ่ม (Multinomial Logistic Regression) แสดงสมการถดถอยโลจิสติกส์ดังสมการที่ 5

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i \quad (5)$$

โดยที่

$\hat{Y}$ คือ การประมาณค่าตัวแปรผล หรือตัวแปรตอบสนอง

β_n คือ ค่าสัมประสิทธิ์ความถดถอยของตัวแปรทำนายตัวที่ n

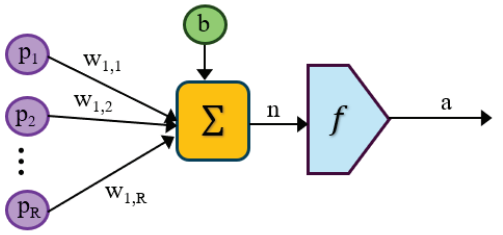
x_i คือ ค่าของตัวแปรทำนายตัวที่ i

2.4 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

โครงข่ายประสาทเทียม เป็นการจำลองการทำงานของเครือข่ายประสาทในสมองของมนุษย์ด้วยแบบจำลองทางคณิตศาสตร์ซึ่งมีการปรับเปลี่ยนตัวเองต่อการตอบสนองของข้อมูลนำเข้าหรือค่าอินพุตตามกฎการเรียนรู้ (Learning Rule) หลังจากที่เครือข่ายได้เรียนรู้สิ่งที่ต้องการ เครือข่ายนั้นจะสามารถทำงานที่กำหนดไว้ได้ เป็นแนวความคิดที่ต้องการให้คอมพิวเตอร์มีความสามารถในการเรียนรู้เหมือนมนุษย์ สมองมนุษย์มีนิวรอนหรือเซลล์ประสาท ซึ่งเป็นหน่วยประมวลผลที่มีการเชื่อมต่อกันมากมายอยู่ในสมองมนุษย์มีประมาณ 10^{11} นิวรอน จึงสามารถกล่าวได้ว่าสมองมนุษย์เป็นคอมพิวเตอร์ที่มีการปรับตัวเอง (Adaptive) ไม่เป็นเชิงเส้น (Nonlinear) และมีการทำงานแบบขนาน (Parallel) ในการจัดการการทำงานร่วมกันของนิวรอน

คุณลักษณะเด่นของโครงข่ายประสาทเทียม คือ โครงข่ายจะประกอบไปด้วยหน่วยประมวลผลย่อยๆ ซึ่งเชื่อมต่อแบบขนานเป็นจำนวนมาก ในหน่วยประมวลผลย่อยแต่ละหน่วยมีโครงสร้างง่ายๆ และไม่ค่อยมีความสามารถ แต่เมื่อหน่วยประมวลผลย่อยๆ เหล่านี้ทำงานร่วมกันแบบกระจายทำให้โครงข่ายประสาทเทียมจะมีการทำงานที่มีประสิทธิภาพ โครงข่ายจะมีการเชื่อมต่อด้วยหน่วยประมวลผลย่อยๆ จำนวนมาก ถ้าเครือข่ายบางส่วนเสียหาย แต่การทำงานของโครงข่ายประสาทเทียมจะยังคงสามารถทำงานได้ คุณสมบัติที่เด่นสุดคือ สามารถเรียนรู้และแก้ไขปัญหาได้อย่างมีประสิทธิภาพ ผลจากการเรียนรู้ด้วยตัวอย่างข้อมูลบางส่วนนำไปสู่การตอบสนองต่อข้อมูลอินพุตที่เข้ามาใหม่

โครงข่ายประสาทเทียมประกอบไปด้วยเซตของโหนดและเส้นเชื่อมระหว่างโหนด โดยที่โหนดจะแบ่งเป็น 3 ระดับ ได้แก่ ชั้นอินพุต (Input layer) ชั้นซ่อน (Hidden layer) และชั้นเอาท์พุต (Output



ภาพประกอบ 1 โครงข่ายประสาทเทียมหนึ่งหน่วยแบบหลายอินพุท

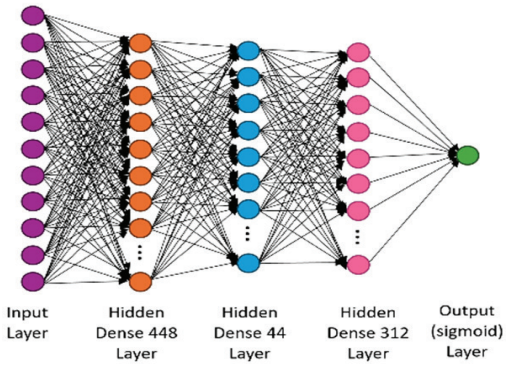
layer) ที่ชั้นซ่อนอาจจะมีได้มากกว่า 1 ชั้นขึ้นอยู่กับการออกแบบโครงข่าย โดยสถาปัตยกรรมของโครงข่ายประสาทเทียมภายในโครงข่ายจะมีเส้นเชื่อมจากทุกโหนดในชั้นอินพุตไปยังทุกโหนดในชั้นซ่อน และมีเส้นเชื่อมจากทุกโหนดในชั้นซ่อนไปยังทุกโหนดในชั้นเอาท์พุต โดยที่เส้นเชื่อมแต่ละเส้นจะมีค่าน้ำหนัก (Weight) โครงข่ายประสาทเทียมหนึ่งหน่วยสามารถมีหลายอินพุทได้ (Chanklan, 2017) แสดงโครงข่ายประสาทเทียมหนึ่งหน่วยดังภาพประกอบ 1

การทำงานของแต่ละโหนดเทียบได้กับเซลล์ประสาทในสมองมนุษย์ 1 เซลล์ อินพุตที่เข้าสู่โหนดจะเป็นเวกเตอร์ของคุณสมบัติของข้อมูลตัวอย่างมีค่า $p = [p_1,p_2,...,p_R]$ ซึ่งเป็นค่าอินพุตที่ถูกป้อนมีจำนวน R องค์ประกอบ และเวกเตอร์น้ำหนัก $W=[w_1,w_2,...,w_R]$ มีค่าเอนเอียงหรือไบเอส b นำอินพุตมาคูณกับน้ำหนักของแต่ละเส้นเชื่อม ผลที่ได้จากอินพุตทุกๆ เส้นเชื่อมของโหนดจะเอามารวมกันและรวมกับค่าไบเอสแล้วส่งต่อไปยังฟังก์ชันถ่ายโอน (Transfer Function) ซึ่งเกิดเป็นค่าเอาท์พุต a ในที่นี้ f เป็นฟังก์ชันถ่ายโอนทำหน้าที่รับค่าอินพุต n เพื่อเปลี่ยนเป็นค่าเอาท์พุต a ค่าเอาท์พุต a สามารถคำนวณได้จากสมการที่ 6

$$a = f(n) = f(Wp+b) \quad (6)$$

โดยที่

f คือ ฟังก์ชันถ่ายโอน



ภาพประกอบ 2 โครงข่ายประสาทเทียมที่ใช้ในงานวิจัย

W คือ เวกเตอร์น้ำหนัก

b คือ ค่าไบแอส

p คือ ค่าอินพุต หรือข้อมูลนำเข้า

สำหรับงานวิจัยนี้ได้ออกแบบโครงสร้างของโครงข่ายประสาทเทียมที่ใช้ในการทดลองโดยหาโครงข่ายประสาทเทียมที่เหมาะสมกับข้อมูลโดยใช้แบบจำลองไฮเปอร์พารามิเตอร์ด้วย Keras Tuner เพื่อกำหนดจำนวนและความกว้างของเลเยอร์ที่ซ่อนอยู่ โดยกำหนดตำนวนโหนดอยู่ในช่วง 8-512 และจำนวน 3 ชั้นเลเยอร์ และกับหนดการโดยได้โครงข่ายประสาทเทียมที่ใช้ในงานวิจัยมีรายละเอียดดังนี้:

ชั้นแรก (Input Layer): มี 11 โหนด และใช้ฟังก์ชันการกระตุ้น relu โดยรับข้อมูลจากจำนวนโหนดขึ้นอยู่กับขนาดของข้อมูลนำเข้า

ชั้นที่สอง: มี 448 โหนด ใช้ฟังก์ชันการกระตุ้น relu

ชั้นที่สาม: มี 44 โหนด ใช้ฟังก์ชันการกระตุ้น relu

ชั้นที่สี่: มี 312 โหนด ใช้ฟังก์ชันการกระตุ้น relu

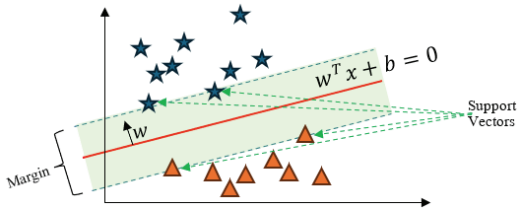
ชั้นสุดท้าย (Output Layer): มี 1 โหนด ใช้ฟังก์ชันการกระตุ้น sigmoid ซึ่งใช้สำหรับการจำแนกประเภทแบบไบนารี

โครงข่ายนี้จะรับข้อมูลนำเข้าผ่านชั้นแรก และผ่านการประมวลผลต่อเนื่องผ่านชั้นต่างๆ จนถึงชั้นสุดท้ายที่จะให้ผลลัพธ์เป็นค่าความน่าจะเป็นระหว่าง 0 ถึง 1 สำหรับการจำแนกประเภทสองกลุ่ม โดยมีค่า learning rate=0.001 แสดงรูปโครงข่ายได้ดังภาพประกอบ 2

2.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

ซัพพอร์ตเวกเตอร์แมชชีน เป็นอัลกอริทึมที่ใช้ในการจำแนกประเภทข้อมูลในแต่ละคลาสที่ได้รับความนิยมมาก (Hearst *et al*.,1998) เนื่องจากมีความสามารถในการจำแนกประเภทข้อมูลแต่ละคลาสด้อย่างมีประสิทธิภาพและมีความแม่นยำสูง โดยหลักการสำคัญของอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนคือการสร้างเส้นแบ่ง (Hyperplane) เพื่อแบ่งแยกประเภทข้อมูลที่ต่างชนิดกันให้แยกออกจากกัน

ในการสร้างเส้นแบ่งของซัพพอร์ตเวกเตอร์แมชชีนเพื่อใช้ในการจำแนกข้อมูล จะสร้างเส้นแบ่งข้อมูลที่มีระยะห่างระหว่างข้อมูลมากที่สุด ซึ่งจะอาศัยเวกเตอร์ถ่วงน้ำหนัก w (Weight Vector) เป็นตัวกำหนดทิศทางและใช้กำหนดความเอียงของระนาบ (Hyperplane) ซึ่งเวกเตอร์ w จะตั้งฉากกับเส้นแบ่ง และข้อมูลจะถูกแปลงให้อยู่ในรูปแบบเวกเตอร์ x สำหรับการตีความว่าข้อมูลจุดนั้นจะถูกกำหนดเป็นแบ่งข้อมูลออกเป็นประเภท (หรือ คลาส) 1 หรือ -1 นั้นจะกำหนดจากตัวแปร y โดยสามารถแสดงสมการในการหาคลาสข้อมูลดังสมการที่ 7



ภาพประกอบ 3 เส้นแบ่งสำหรับแบ่งข้อมูลด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน

$$w^T x + b \geq 1, \text{ เมื่อ } yi = +1 \quad (7)$$

$$w^T x + b \leq 1, \text{ เมื่อ } yi = +1$$

เมื่อ

w คือ เวกเตอร์ถ่วงน้ำหนัก (Weight Vector)

b คือ ค่าไบแอส (Bias)

x คือ ข้อมูลที่จุดใดๆ

โดยกราฟแสดงการสร้างเส้นแบ่งสำหรับแบ่งข้อมูลด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนในภาพประกอบ 3

2.6 เกนความรู้ (Information Gain: IG)

เกนความรู้เป็นวิธีการคำนวณค่าน้ำหนักเพื่อใช้คัดเลือกคุณสมบัติ โดยการคัดเลือกคุณสมบัติจะเลือกคุณสมบัติที่มีค่าเกนความรู้ที่มีค่าสูง การคำนวณค่า Information Gain จะต้องใช้ค่า Entropy ในการคำนวณ ซึ่งค่า Entropy เป็นการวัดความไม่เป็นระเบียบหรือความแตกต่างของข้อมูลมีค่าตั้งแต่ 0 ถึง 1 ถ้าค่า Entropy สูงจะหมายถึงข้อมูลมีความแตกต่างหรือกระจายตัวมาก โดยค่า Entropy ของแต่ละกลุ่มในชุดข้อมูลคำนวณได้ดังสมการที่ 8

$$Entropy\ (Condition) = -\sum_{n=1}^{k} (Prob(C_n) \times log_2(Prob(C_n))) \quad (8)$$

เมื่อ

k คือ จำนวนกลุ่มของข้อมูลทั้งหมด

C_n คือ กลุ่มของข้อมูล n โดยที่ n มีค่าตั้งแต่ 1, 2, 3, ..., k

$Prob(C_n)$ คือ ความน่าจะเป็นกลุ่มของข้อมูลที่สนใจ

เมื่อคำนวณหาค่า Entropy เรียบร้อยแล้ว จึงคำนวณค่าเกนความรู้ของแต่ละคุณสมบัติได้ดังสมการที่ 9

$$IG = Entropy(Parent) - \sum_{i=1}^{k} (Prob(C_n) \times Entropy(C_n)) \quad (9)$$

เมื่อ

$Entropy(Parent)$ คือ ผลรวมของค่า Entropy ของแต่ละกลุ่มข้อมูลในแต่ละคุณสมบัติ

C_n คือ กลุ่มของข้อมูล n โดยที่ n มีค่าตั้งแต่ 1, 2, 3, ..., k

2.7 คอนฟิวชันเมทริกซ์ (Confusion Matrix)

คอนฟิวชันเมทริกซ์ เป็นตารางที่ใช้ในการประเมินประสิทธิภาพของผลลัพธ์ในการทำนายหรือค่าที่คาดการณ์ (Prediction) ซึ่งผลลัพธ์การทำนายจะได้รับจากโมเดลที่สร้างขึ้นจากอัลกอริทึมการเรียนรู้ของเครื่อง (Ting, 2011) โดยจะเป็นตารางที่แสดงค่าสัดส่วนระหว่างค่าจริง (Actual) เปรียบเทียบกับผลลัพธ์การทำนาย ในตาละช่องของคอนฟิวชันเมทริกซ์จะประกอบไปด้วยค่า True Positive (TP), False Negative (FN), True Negative (TN) และ False Positive (FP) แสดงดังภาพประกอบ 4

โดย ค่า True Positive (TP) คือ จำนวนที่โมเดลทำนายว่า Yes ได้ถูกต้องตรงกับค่าจริง

ค่า False Negative(FN) คือ จำนวนที่โมเดลทำนายว่า No ซึ่งทำนายผิดเพราะค่าจริงคือ Yes

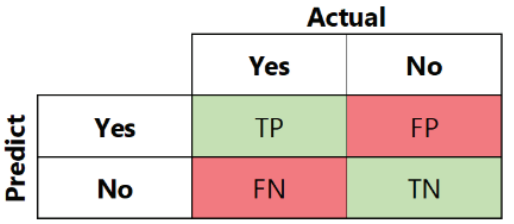
ค่า True Negative (TN) คือ จำนวนที่โมเดลทำนายว่า No ได้ถูกต้องตรงกับค่าจริง

ค่า False Positive (FP) คือ จำนวนที่โมเดลทำนายว่า Yes ซึ่งทำนายผิดเพราะค่าจริงคือ No

สามารถนำค่าในตารางคอนฟิวชันเมทริกซ์มาคำนวณหาค่าความถูกต้อง (Accuracy) เพื่อประเมิน

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope
0	40	1	2	140	289	0	0	172	0	0.00	1
1	49	0	3	160	180	0	0	156	0	1.00	2
2	37	1	2	130	283	0	1	98	0	0.00	1
3	48	0	4	138	214	0	0	108	1	1.50	2
4	54	1	3	150	195	0	0	122	0	0.00	1
5	39	1	3	120	339	0	0	170	0	0.00	1
6	45	0	2	130	237	0	0	170	0	0.00	1
7	54	1	2	110	208	0	0	142	0	0.00	1
8	37	1	4	140	207	0	0	130	1	1.50	2
9	48	0	2	120	284	0	0	120	0	0.00	1

ภาพประกอบ 5 ตัวอย่างข้อมูลโรคหลอดเลือดหัวใจ



ภาพประกอบ 4 ตารางคอนฟิวชันเมทริกซ์

ประสิทธิภาพการทำนายของโมเดลได้ในสมการที่ 10

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \qquad (10)$$

3. ขั้นตอนดำเนินการ

ในงานวิจัยนี้ผู้วิจัยใช้ชุดข้อมูลโรคหลอดเลือดหัวใจจาก https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive ซึ่งเป็นข้อมูลที่ได้รวบรวมข้อมูลจาก 5 ชุดข้อมูล โดยข้อมูลนี้มีทั้งหมด 1,190 รายการโดยไม่มีข้อมูลสูญหาย (Missing Value) ประกอบไปด้วยตัวอย่างข้อมูลคนปกติ (ไม่เป็นโรคหลอดเลือดหัวใจ) จำนวน 561 รายการ และผู้ป่วยโรคหลอดเลือดหัวใจจำนวน 629 รายการ ซึ่งงานวิจัยนี้จะแบ่งข้อมูลออกเป็น 2 ส่วน โดยใช้สัดส่วน 70:30 ซึ่งร้อยละ 70 ของข้อมูลคิดเป็นจำนวน 833 ข้อมูลจะถูกใช้สำหรับฝึกสอนเพื่อหารูปแบบของข้อมูล และข้อมูลทดสอบจำนวน 357 ข้อมูลคิดเป็นร้อยละ 30 ของข้อมูลทั้งหมด ซึ่งข้อมูลที่นำมาใช้มีจำนวนคุณสมบัติทั้งหมด 12 คุณสมบัติ โดยแสดงตัวอย่างของข้อมูลดังภาพประกอบ 5 ชุดข้อมูลโรคหลอดเลือดหัวใจที่มีจำนวนข้อมูลมากที่สุดเพื่อวัตถุประสงค์ในการวิจัยและวินิจฉัยโรคหลอดเลือดหัวใจ โดยชุดข้อมูลทั้ง 5 ชุด มีแหล่งที่มาดังต่อไปนี้

1. Cleveland Dataset: ชุดข้อมูลนี้มาจาก Cleveland Clinic Foundation โดยเป็นข้อมูลได้รับจากผู้ป่วยที่เข้ารับการตรวจที่โรงพยาบาล Cleveland Clinic ในช่วงปี 1988-1991 ชุดข้อมูลนี้มีจำนวนตัวอย่างทั้งหมด 303 รายการ

2. Hungarian Dataset: ชุดข้อมูลนี้รวบรวมจาก Institute of Cardiology, University of Debrecen, Hungary เป็นข้อมูลที่ได้รับจากผู้ป่วยที่เข้ารับการตรวจที่โรงพยาบาลในประเทศฮังการี ชุดข้อมูลนี้มีจำนวนตัวอย่างทั้งหมด 294 รายการ

3. Switzerland Dataset: ชุดข้อมูลนี้มาจาก University Hospital Zurich ในประเทศสวิตเซอร์แลนด์ ซึ่งเป็นข้อมูลจากผู้ป่วยที่เข้ารับการตรวจที่โรงพยาบาล ชุดข้อมูลนี้มีจำนวนตัวอย่างทั้งหมด 123 รายการ

4. Long Beach VA Dataset: ชุดข้อมูลนี้มาจาก Long Beach Veterans Administration Medical Center เป็นข้อมูลจากผู้ป่วยที่เข้ารับการตรวจที่โรงพยาบาลใน เมือง Long Beach ที่รัฐ California ชุดข้อมูลนี้มีจำนวนตัวอย่างทั้งหมด 200 รายการ

5. Statlog (Heart) Data Set: ชุดข้อมูลนี้ได้รับมาจาก Statlog Project ประกอบไปด้วยข้อมูลจากผู้ป่วยที่เข้ารับการตรวจที่โรงพยาบาลในประเทศอังกฤษ ชุดข้อมูลนี้มีจำนวนตัวอย่างทั้งหมด 270 รายการ

เมื่อรวบรวมข้อมูลโรคหลอดเลือดหัวใจทั้ง 5 ชุด ทำให้ได้ชุดข้อมูลที่ประกอบไปด้วยคุณสมบัติทั้งหมด 12 คุณสมบัติซึ่งมีรายละเอียดของแต่ละคุณสมบัติดังต่อไปนี้:

1. age: อายุของผู้ป่วย (ปี)

2. sex: เพศ (1 = ชาย, 0 = หญิง)

3. chest pain type: ประเภทของอาการเจ็บหน้าอก (1 = อาการเจ็บหน้าอกที่เกิดขึ้นจากการขับขี่, 2 = อาการเจ็บหน้าอกที่เกิดขึ้นจากการเดินขึ้นบันไดหรือการเดินขึ้นเนิน, 3 = อาการเจ็บหน้าอกที่ไม่มีการเคลื่อนไหว, 4 = อาการเจ็บหน้าอกที่มีการเคลื่อนไหว)

4. resting bp s: ความดันโลหิตในช่วงการทดสอบ (mm Hg)

5. cholesterol: ระดับคอเลสเตอรอลในเลือด (mg/dl)

6. fasting blood sugar: ระดับน้ำตาลในเลือดที่เกินเกณฑ์ปกติ (1 = เกินเกณฑ์ปกติ, 0 = ไม่เกินเกณฑ์ปกติ)

7. resting ecg: ผลการตรวจ ECG (คลื่นไฟฟ้าหัวใจ (Electrocardiogram)) ในช่วงการทดสอบ (0 = ปกติ, 1 = มีความผิดปกติที่ ST-T, 2 = มีความผิดปกติที่ ST-T และความผิดปกติที่ Q)

8. max heart rate: อัตราการเต้นของหัวใจสูงสุดในช่วงการทดสอบ (ครั้งต่อนาที)

9. exercise angina: อาการเจ็บหน้าอกที่เกิดขึ้นในช่วงการทดสอบ (1 = มี, 0 = ไม่มี)

10. oldpeak: ค่าที่ลดลงของ ST depression ในคลื่นไฟฟ้าของ ECG ที่เกิดขึ้นในช่วงการทดสอบ (เมื่อเทียบกับการพักฟื้น) มีหน่วยเป็น mm (มิลลิเมตร) โดยค่าที่มากขึ้นอาจแสดงถึงความรุนแรงของโรคหัวใจ

11. ST slope: ความลาดของค่า ST segment ในคลื่นไฟฟ้าของ ECG ที่เกิดขึ้นในช่วงการทดสอบ (1 = ลาดขึ้น, 2 = ลาดเรียบ, 3 = ลาดลง)

12. target: เป็นโรคหลอดเลือดหัวใจหรือไม่ (0=ปกติ, 1=เป็นโรคหลอดเลือดหัวใจ)

งานวิจัยนี้ใช้ภาษาไพทอนบน Google Colab ในการทดลองโดยมีขั้นตอนการดำเนินงานดังนี้

1. การเตรียมข้อมูล จะดำเนินโดยการพิจารณาข้อมูลทั้งหมดทั้ง 12 คุณสมบัติเพื่อแยกคุณสมบัติที่ข้อมูลเป็นข้อมูลเชิงตัวเลข (Numeric Data) และข้อมูลเชิงหมวดหมู่ (Categorical Data) จากนั้นทำการแปลงข้อมูลคุณสมบัติเชิงหมวดหมู่แปลงข้อมูลให้เป็นรูปแบบที่สามารถใช้กับอัลกอริทึมการเรียนรู้ของเครื่องได้ dummy variables โดยการสร้างคอลัมน์ใหม่สำหรับแต่ละค่าหมวดหมู่ และกำหนดค่าในคอลัมน์นั้นเป็น 1 หากข้อมูลในแถวนั้นตรงกับค่าหมวดหมู่ที่คอลัมน์นั้นแทน และเป็น 0 หากไม่ตรง

2. ลดมิติข้อมูลด้วยการคัดเลือกคุณสมบัติที่เหมาะสมโดยใช้การหาค่าสหสัมพันธ์ ค่าเกนความรู้ และการวิเคราะห์แยกแยะเชิงเส้น โดยเลือกคุณสมบัติจำนวน 5 คุณสมบัตินำมาใช้เป็นตัวแทนของข้อมูล

3. การแบ่งข้อมูล ดำเนินการโดยสุ่มแบ่งชุดข้อมูลออกเป็นชุดสำหรับการฝึก (Training set) 70% และชุดสำหรับการทดสอบ (Test set) 30% การแบ่งข้อมูลลักษณะนี้ช่วยให้สามารถประเมินประสิทธิภาพของโมเดลได้อย่างเป็นกลาง เพราะแบ่งข้อมูลสำหรับการฝึกเพื่อสร้างโมเดลในการการทำนาย และใช้ข้อมูลสำหรับการทดสอบเพื่อประเมินประสิทธิภาพซึ่งข้อมูลชุดนี้จะทำหน้าที่เป็นตัวแทนของข้อมูลใหม่ที่โมเดลไม่เคยเห็นมาก่อน ทำให้การประเมินประสิทธิภาพของโมเดลมีความน่าเชื่อถือ

4. การสร้างโมเดล เมื่อได้คุณสมบัติที่เป็นตัวแทนของข้อมูลจากวิธีการคัดเลือกคุณสมบัติในแต่ละวิธีแล้ว จะนำข้อมูลเข้าสู่กระบวนการจำแนกด้วยอัลกอริทึมด้านการเรียนรู้ของเครื่อง ได้แก่ LR, ANN และ SVM โดยใช้ 4 เคอร์เนล (Kernel) ได้แก่ Radial Basis Function (RBF), Linear, Polynomial และ Sigmoid

5. การเปรียบเทียบประสิทธิภาพ ในการทำนาย ในการทดลองจะเปรียบเทียบประสิทธิภาพการทำนายของโมเดลที่เกิดจากอัลกอริทึมที่กล่าวข้างต้น โดยสร้างโมเดลจากข้อมูลนำเข้า 4 แบบ ได้แก่ ข้อมูลดั้งเดิม ข้อมูลที่ถูกคัดเลือกคุณสมบัติด้วยวิธีการค่าสหสัมพันธ์ ค่าเกนความรู้ และวิเคราะห์แยกแยะเชิงเส้นที่นำเสนอ ในการพิจารณาเปรียบเทียบประสิทธิภาพจะใช้ค่าความถูกต้องในการทำนาย โดยแสดงวิธีการดำเนินการวิจัยดังภาพประกอบ 6

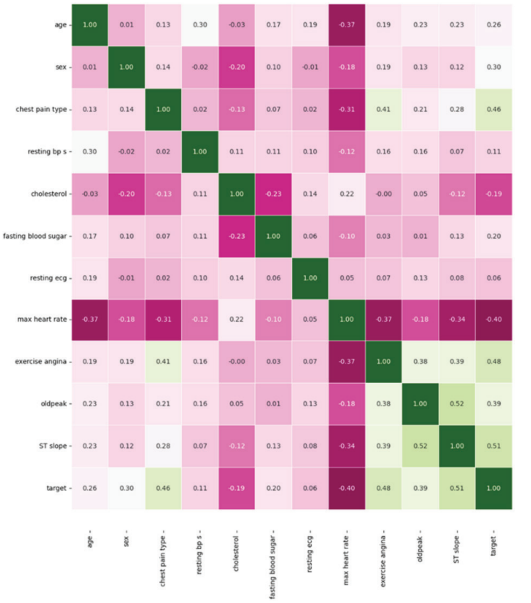
4. ผลการศึกษาและอภิปรายผล

ในการคัดเลือกคุณลักษณะด้วยการใช้อัลกอริทึมในการคัดเลือกคุณลักษณะจำนวน 2 อัลกอริทึม ดังต่อไปนี้

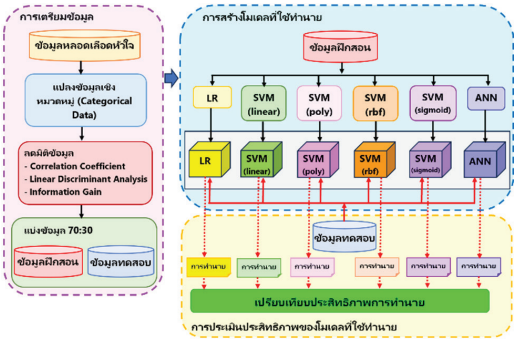
การคัดเลือกคุณลักษณะด้วยค่าสหสัมพันธ์ โดยใช้ความสัมพันธ์ระหว่างข้อมูลแต่ละคุณลักษณะเพื่อพิจารณาความสัมพันธ์กัน ซึ่งค่าที่ได้จะมีค่าอยู่ระหว่าง -1 ถึง 1 ซึ่งหากค่าสัมบูรณ์ของค่าสหสัมพันธ์ที่ได้มีค่าสูง จะสามารถตีความได้ว่าคุณสมบัติทั้งสองมีความสัมพันธ์กันสูงเช่นกัน โดยเครื่องหมายจะบ่งบอกถึงทิศทางของความสัมพันธ์ว่ามีความสัมพันธ์กันในทิศทางใด โดยผลการวิเคราะห์ค่าสหสัมพันธ์ของข้อมูลโรคหลอดเลือดหัวใจแสดงในภาพประกอบ 7 โดยคัดเลือกคุณลักษณะที่มีค่าสัมบูรณ์ของค่าสหสัมพันธ์ที่มีค่าสูงกว่า 0.4 นั้นหมายความว่าคุณสมบัติที่เลือกเหล่านี้จะมีความสัมพันธ์กันระดับปานกลางขึ้นไป โดยจะดูความสัมพันธ์ทุกคู่คุณสมบัติที่เป็นไปได้ ทำให้ได้คุณสมบัติที่สามารถนำไปใช้งานทั้งสิ้น 4 คุณลักษณะได้แก่ chest pain type, exercise angina, oldpeak และ ST slope

การคัดเลือกคุณลักษณะด้วยการวิเคราะห์แยกแยะเชิงเส้น โดยแสดงค่าไอเกนที่ใช้ในการบ่งบอกถึงความสำคัญของแต่ละคุณสมบัติในชุดข้อมูลทั้ง 12 คุณสมบัติ โดยผลการวิเคราะห์ความสัมพันธ์แสดงดังภาพประกอบ 8 โดยค่าไอเกนแสดงถึงปริมาณความแปรปรวนที่แต่ละฟีเจอร์สามารถอธิบายในการแยกคลาสไม่มีเกณฑ์ที่แน่นอนสำหรับค่าที่เหมาะสมในการควรเลือกคุณสมบัติ ดังนั้นในงานวิจัยจึงคัดเลือกคุณสมบัติที่มีค่าไอเกนสูงที่สุดจำนวน 4 คุณสมบัติ เพื่อให้เท่ากับจำนวนที่ใช้ในค่าสหสัมพันธ์ ได้แก่ ST slope, exercise angina, sex และ chest pain type

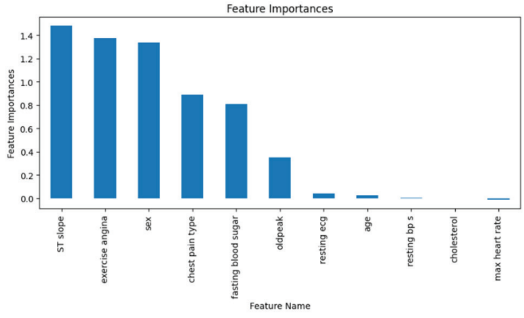
การคัดเลือกคุณสมบัติด้วยการคำนวณค่าเกนความรู้ โดยแสดงค่าเกนความรู้เพื่อใช้เป็นค่าน้ำหนักในการคัดเลือกคุณสมบัติในชุดข้อมูลดังภาพประกอบ 9 ค่าเกนความรู้ไม่มีค่าเกณฑ์ที่แน่นอน



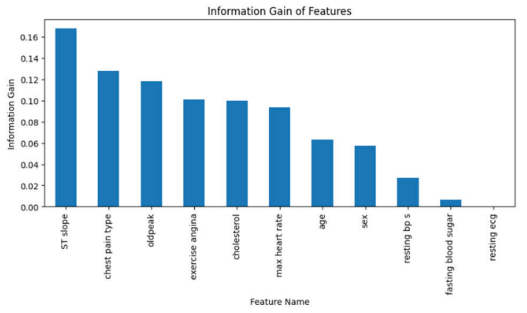
ภาพประกอบ 7 ค่าสหสัมพันธ์ของข้อมูลโรคหลอดเลือดหัวใจ



ภาพประกอบ 6 กรอบแนวคิดงานวิจัย



ภาพประกอบ 8 ค่าไอเกนของแต่ละคุณสมบัติของข้อมูลโรคหลอดเลือดหัวใจด้วย



ภาพประกอบ 9 ค่าเกนความรู้ของแต่ละคุณสมบัติของข้อมูลโรคหลอดเลือดหัวใจด้วย

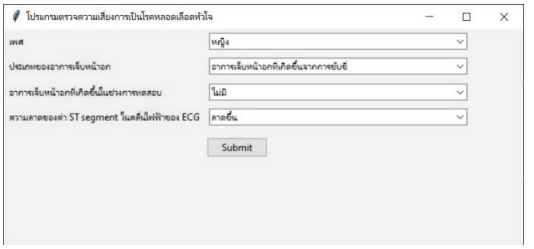
ในการคัดเลือกว่าควรใช้ค่าค่าเกนความรู้เท่าไหร่ถึงควรเลือกคุณสมบัติ ดังนั้นในงานวิจัยจึงคัดเลือกคุณสมบัติจำนวน 4 คุณสมบัติเพื่อให้เท่ากับจำนวนที่ใช้ในค่าสหสัมพันธ์ ได้แก่ ST slope, chest pain type, max heart rate และ exercise angina

เมื่อได้ผลการคัดเลือกคุณสมบัติจากค่าสหสัมพันธ์และการวิเคราะห์แยกแยะเชิงเส้นแล้วจะนำไปสร้างโมเดลเพื่อใช้ในการทำนายผู้ป่วยโรคหลอดเลือดหัวใจโดยใช้อัลกอริทึมด้านการเรียนรู้ของเครื่องที่ ได้แก่ อัลกอริทึมการถดถอยโลจิสติกส์ อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (โดยใช้ 4 เคอร์เนล ได้แก่ RBF, Linear, Polynomial, Sigmoid) มีการกำหนดค่าพารามิเตอร์ gamma=0.1, C=1.0 เท่ากันทุกโมเดลเพื่อไม่ให้เกิดความเอนเอียง และอัลกอริทึมโครงข่ายประสาทเทียมใช้ Keras Tuner เพื่อให้ได้โครงข่ายที่เหมาะสมกับข้อมูลและใช้โครงข่ายดังกล่าวในทุกโมเดลที่ใช้โครงข่ายประสาทเทียม โดยจะเปรียบเทียบผลประสิทธิภาพด้านความถูกต้องในการทำนายผู้ป่วยโรคหลอดเลือดหัวใจ โดยแสดงผลการทำนายของโมเดลที่ข้อมูลจากเทคนิคการคัดเลือกคุณลักษณะด้วยการวิเคราะห์แยกแยะเชิงเส้นเปรียบเทียบกับการใช้ข้อมูลคุณลักษณะดั้งเดิม (ไม่มีการคัดเลือก) และการคัดเลือกคุณลักษณะด้วยค่าสหสัมพันธ์และค่าเกนความรู้ ผลการทดลองแสดงดังตาราง 2

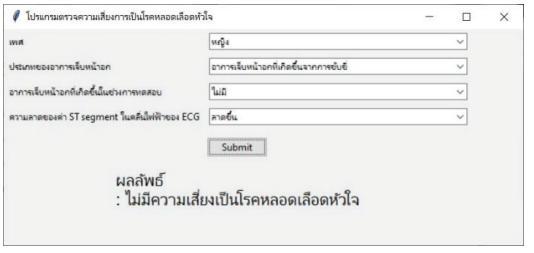
จากตาราง 2 จะเห็นว่าที่ข้อมูลดั้งเดิมโมเดลที่มีค่าความถูกต้อง (Accuracy) สูงสุดคือที่ซัพพอร์ทเวกเตอร์แมชชีนด้วยเคอเนล Linear (SVM (linear)) โดยมีค่าความถูกต้องร้อยละ 85.99 เมื่อนำข้อมูลที่คัดเลือกคุณสมบัติด้วยค่าสหสัมพันธ์ (Correlation) ไปสร้างโมเดลด้วย SVM (RBF) ทำให้ประสิทธิภาพสูงสุดที่ ร้อยละ 84.87 และเมื่อใช้เทคนิคการวิเคราะห์แยกแยะเชิงเส้น (LDA) ในการคัดเลือกคุณสมบัติมีค่าความถูกต้องสูงสุดที่ร้อยละ 87.39 และเมื่อนำข้อมูลที่ถูกคัดเลือกคุณสมบัติโดยค่าเกนความรู้ (IG) โมเดลมีประสิทธิภาพสูงสุดที่ SVM (Linear) โดยมีค่า

ตาราง 2 ผลการทดลอง

Algorithm	Accuracy			
	ข้อมูลดั้งเดิม	Correlation	LDA	IG
LR	84.87%	84.31%	87.11%	84.03%
SVM (linear)	85.99%	80.11%	84.03%	84.31%
SVM (poly)	82.35%	83.47%	86.83%	81.23%
SVM (rbf)	73.11%	84.87%	86.55%	82.63%
SVM (sigmoid)	56.86%	72.83%	86.83%	56.86%
ANN	83.75%	83.47%	87.39%	84.03%
เฉลี่ย	77.82%	81.51%	86.46%	78.85%

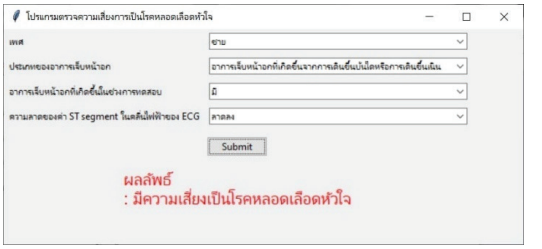


ภาพประกอบ 10 โปรแกรมที่ใช้สำหรับประเมินความเสี่ยงการเป็นโรคหลอดเลือดหัวใจ



ภาพประกอบ 11 แสดงตัวอย่างการทำนายของโปรแกรมเมื่อไม่มีความเสี่ยงเป็นโรคหัวใจ

ความถูกต้องที่ร้อยละ 84.31 การใช้ LDA เพื่อคัดเลือกคุณสมบัติที่เหมาะสมเมื่อเปรียบเทียบกับการใช้ข้อมูลดังเดิมในการสร้างโมเดลจะเห็นว่ามีประสิทธิภาพดีขึ้นแทบทุกโมเดลยกเว้น SVM (Linear) แต่ถ้าพิจารณาที่ค่าความถูกต้องเฉลี่ยเป็นวิธีการที่มีค่าเฉลี่ยมากที่สุดเมื่อเทียบกับโมเดลที่สร้างจากข้อมูลดังเดิมและข้อมูลจากการคัดเลือกคุณสมบัติจากวิธีการอื่นๆ ซึ่งแสดงให้เห็นว่าเป็นวิธีการที่ทำให้ช่วยเพิ่มประสิทธิภาพการทำนาย เนื่องจาก LDA เป็นวิธีการคัดเลือกคุณสมบัติที่ซับซ้อนโดยพิจารณาข้อมูลในมิติใหม่โดยพิจารณาสองเงื่อนไขได้แก่ ข้อมูลจะต้องมีความแปรปรวนภายในคลาสต่ำสุดและความแปรปรวนระหว่างคลาสสูงสุดทำให้ข้อมูลที่ถูกคัดเลือกสามารถใช้ในการแยกคลาสต่างๆ ออกจากกันมากที่สุด ซึ่งในขณะที่ Correlation ตัดคุณลักษณะที่ซ้ำซ้อนกันออกโดยพิจารณาเพียงแค่ค่าความแปรปรวน และ IG ที่คัดเลือกคุณลักษณะที่สามารถแบ่งคลาสของข้อมูลได้ดี โดยพิจารณาจาก Entropy ที่แสดงถึงความความแตกต่างหรือกระจายตัวมากในการแยกคลาส



ภาพประกอบ 12 แสดงตัวอย่างการทำนายของโปรแกรมเมื่อมีความเสี่ยงเป็นโรคหัวใจ

เมื่อได้โมเดลที่ให้ประสิทธิภาพที่ดีที่สุดในการทำนายโรคหลอดเลือดหัวใจและจึงได้นำไปพัฒนาเป็นโปรแกรมที่ใช้สำหรับประเมินความเสี่ยงการเป็นโรคหลอดเลือดหัวใจ แสดงหน้าต่างโปรแกรมดังภาพประกอบ 10 ซึ่งการใช้งานโปรแกรมผู้ใช้งานจะต้องมีข้อมูล เพศ ประเภทของการเจ็บหน้าอก อาการเจ็บหน้าอกที่เกิดขึ้นในช่วงทดสอบ และค่าความลาดของค่า ST segment ในคลื่นไฟฟ้าของ ECG เมื่อใส่ข้อมูลครบถ้วนแล้วกดปุ่ม Submit โปรแกรมจะแสดงผลการทำนายความเสี่ยงการเป็นโรคหลอดเลือดหัวใจดังภาพประกอบ 11 ในกรณีที่ไม่มีความเสี่ยง และภาพประกอบ 12 ในกรณีที่มีความเสี่ยง

5. สรุปผล

งานวิจัยนี้นำเสนอเทคนิคการคัดเลือกคุณสมบัติของข้อมูลด้วยการวิเคราะห์แยกแยะเชิงเส้น เพื่อลดขนาดมิติของข้อมูลเพื่อเพิ่มประสิทธิภาพด้านความถูกต้องในการจำแนกข้อมูลผู้ป่วยโรคหลอดเลือดหัวใจ เมื่อทำการคัดเลือกคุณลักษณะของข้อมูลแล้ว จะใช้เฉพาะข้อมูลของคุณสมบัติที่ได้รับเลือกไปใช้ในการสร้างโมเดลจำแนกประเภทข้อมูลทางด้านการเรียนรู้ของเครื่อง ได้แก่ อัลกอริทึมการถอดถอยโลจิสติกส์ อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน และอัลกอริทึมโครงข่ายประสาทเทียม ซึ่งเปรียบเทียบประสิทธิภาพความถูกต้องในการจำแนกกับการใช้ชุดข้อมูลดั้งเดิมร่วมกับอัลกอริทึมดังกล่าว รวมไปถึงเปรียบเทียบ

กับการคัดเลือกคุณลักษณะด้วยการใช้ค่าสหสัมพันธ์ ผลการทดลองแสดงให้เห็นว่าการคัดเลือกคุณลักษณะด้วยวิธีการที่นำเสนอสามารถช่วยเพิ่มประสิทธิภาพความถูกต้องโดยดูจากค่าเฉลี่ยรวมของทุกโมเดลที่ใช้ในการจำแนกประเภทข้อมูลผู้ป่วยโรคหลอดเลือดหัวใจ นอกจากนี้ยังนำโมเดลที่ได้ไปประยุกต์ใช้พัฒนาเป็นโปรแกรมที่ใช้สำหรับประเมินความเสี่ยงการเป็นโรคหลอดเลือดหัวใจเพื่อให้ผู้ใช้งานได้รับทราบโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดหัวใจของตนเองและสามารถหาทางรักษาได้อย่างทันท่วงที แต่เนื่องด้วยข้อมูลที่ต้องใช้ในการประเมินความเสี่ยงเป็นข้อมูลที่ต้องมีการตรวจจากผู้เชี่ยวชาญจึงเป็นโปรแกรมที่เหมาะสมนำไปใช้ในการช่วงคัดกรองในสถานพยาบาล

สำหรับแนวทางการวิจัยในอนาคตจะมุ่งเน้นการใช้แหล่งข้อมูลโรคหัวใจที่มาจากหลากหลายแหล่งที่มาของข้อมูลเพื่อเพิ่มประสิทธิภาพความแม่นยำและความน่าเชื่อถือของการวิจัย

เอกสารอ้างอิง

Chowdhury, M. N. R., Ahmed, E., Siddik, Md. A. D., & Zaman, A. U. (2021). Heart disease prognosis using machine learning classification techniques. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/i2ct51068.2021.9418181

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications, 13*(4), 18–28. https://doi.org/10.1109/5254.708428

Imanbek, R., Buribayev, Z., & Yerkos, A. (2023). Processing of ischemic heart disease data using ensemble classification methods of machine learning. *Journal of Problems in Computer Science and Information Technologies, 1*(2). https://doi.org/10.26577/jpcsit.2023.v1.i2.06

Kadhim, M. A., & Radhi, A. M. (2023). Heart disease classification using optimized machine learning algorithms. *Iraqi Journal for Computer Science and Mathematics*, 31–42. https://doi.org/10.52866/ijcsm.2023.02.02.004

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart disease prediction using hybrid machine learning model. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1329–1333. https://doi.org/10.1109/icict50816.2021.9358597

Lakshmi, A., & Devi, R. (2023). Heart disease prediction using enhanced whale optimization algorithm based feature selection with machine learning techniques. *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 644–648. https://doi.org/10.1109/smart59791.2023.10428617

Modak, S., Abdel-Raheem, E., & Rueda, L. (2022). Heart disease prediction using adaptive infinite feature selection and deep neural networks. *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 235–240. https://doi.org/10.1109/icaiic54071.2022.9722652

Radhika, R., & George, S. T. (2021). Heart disease classification using machine learning techniques. *Journal of Physics: Conference Series, 1937*(1), 012047. https://doi.org/10.1088/1742-6596/1937/1/012047

Chanklan, R. (2017). *Modeling with machine learning techniques to predict runoff [Doctor dissertation, Suranaree University of Technology]*. Retrieved from http://sutir.sut.ac.th:8080/jspui/handle/123456789/7683 [In Thai]

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine, 18*(10), 1099–1104. Portico. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications, 30*(2), 169–190. https://doi.org/10.3233/aic-170729

Ting, K. M. (2011). Confusion matrix. *Encyclopedia of Machine Learning*, 209–209. https://doi.org/10.1007/978-0-387-30164-8_157