# การพยากรณ์การเกิดโรคหลอดเลือดหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้โครงข่ายประสาทเทียม

# PREDICTION OF CORONARY HEART DISEASE BASED ON EXTERNAL MEASURABLE FACTORS USING NEURAL NETWORK


กนกภัณฑ์ โชติชะวารานนท์

KANOKPHAN CHOTICHAVARANON


สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

มหาวิทยาลัยศรีปทุม

ปีการศึกษา 2563

# การพยากรณ์การเกิดโรคหลอดเลือดหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้โครงข่ายประสาทเทียม

กนกภัณฑ์ โชติชะวารานนท์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม
ปีการศึกษา 2563

# PREDICTION OF CORONARY HEART DISEASE BASED ON MEASURABLE EXTERNAL FACTORS USING NEURAL NETWORK

**KANOKPHAN CHOTICHAVARANON**

**A THEMATIC SUBMITTED IN PARTIAL FULFILLMENT**

**OF THE REQUIREMENTS FOR THE DEGREE OF MASTER**

**OF SCIENCE IN INFORMATION TECHNOLOGY**

**SCHOOL OF INFORMATION TECHNOLOGY**

**SRIPATUM UNIVERSITY**

**ACADEMIC YEAR 2020**

| | |
|---|---|
| **หัวข้อสารนิพนธ์** | การพยากรณ์การเกิดโรคหลอดเลือดหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้โครงข่ายประสาทเทียม<br>PREDICTION OF CORONARY HEART DISEASE BASED ON EXTERNAL MEASURABLE FACTORS USING NEURAL NETWORK |
| **ชื่อนักศึกษา** | กนกภัณฑ์ โชติชะวารานนท์ รหัสประจำตัว 63502665 |
| **หลักสูตร** | วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ |
| **คณะ** | เทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม |
| **อาจารย์ที่ปรึกษาสารนิพนธ์** | ผู้ช่วยศาสตราจารย์ ดร.ปรีชา ตั้งเกรียงกิจ |

**คณะกรรมการการสอบสารนิพนธ์**

……..…….ประธานกรรมการ
(รองศาสตราจารย์ ดร.ทศนัย ชุ่มวัฒนะ)

…….กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ปรีชา ตั้งเกรียงกิจ)

…….กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุรศักดิ์ มังสิงห์)

…….กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ธนา สุขวารี)

**คณะเทคโนโลยีสารสนเทศมหาวิทยาลัยศรีปทุม**อนุมัติให้นับสารนิพนธ์ฉบับนี้เป็นส่วนหนึ่ง ของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณบดีเทคโนโลยีสารสนเทศ

……..…….
(ผู้ช่วยศาสตราจารย์ ดร.ธนา สุขวารี)
วันที่ 13 เดือน……∴…….. พ.ศ. 2564

| | |
|---|---|
| **วิทยานิพนธ์เรื่อง** | การพยากรณ์การเกิดโรคหลอดเลือดหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้ โครงข่ายประสาทเทียม |
| **คำสำคัญ** | โครงข่ายประสาทเทียม/ โรคหลอดเลือดหัวใจ/ โรคหัวใจและหลอดเลือด/ การคัดเลือกคุณลักษณะ/ การพยากรณ์ |
| **นักศึกษา** | กนกภัณฑ์ โชติชะวารานนท์ |
| **อาจารย์ที่ปรึกษาสารนิพนธ์** | ผู้ช่วยศาสตราจารย์ ดร.ปรีชา ตั้งเกรียงกิจ |
| **หลักสูตร** | วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ |
| **คณะ** | เทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม |
| **ปีการศึกษา** | 2563 |

## บทคัดย่อ

ในทุกปีจะมีประชากรที่เป็นโรคหลอดเลือดหัวใจเพิ่มขึ้นแต่กลับมีการเข้าถึงการตรวจโรคได้น้อยและจะมีการตรวจพบเมื่อเป็นโรคแล้วเท่านั้นการป้องกันการเกิดโรคจึงทำได้ยาก งานวิจัยนี้มีวัตถุประสงค์ในการศึกษาปัจจัยที่มีผลต่อการเกิดโรคหลอดเลือดหัวใจ และพัฒนาแบบจำลองการพยากรณ์การเกิดโรคหลอดเลือดหัวใจโดยใช้เพียงปัจจัยที่วัดได้จากภายนอก ข้อมูลจะผ่านการเตรียมข้อมูลและการหาคุณสมบัติที่เหมาะสมโดยวิธี Forward Selection จากนั้นทำการสร้างแบบจำลองโดยใช้โครงข่ายประสาทเทียม ผลการทดลองพบว่าสามารถใช้เพียงปัจจัยที่วัดได้จากภายนอกในการสร้างแบบจำลองการพยากรณ์โดยมีค่าความถูกต้องมากกว่า 90% AUC มีค่ามากกว่า 0.9 ค่า recall และ ค่า specificity มากกว่า 90% โดยใช้ 8 คุณลักษณะดังนี้ อายุ, เพศ, ความดันช่วงบน, ความดันช่วงล่าง,ความสูง, น้ำหนัก, การสูบบุหรี่ และค่า BMI ทั้งยังพบว่าปัจจัยที่วัดได้จากภายนอกที่ถูกใช้มีความขัดแย้งกับแบบคาดการณ์ ATP III hard โรคหลอดเลือดหัวใจ risk score (FRS 2002) ซึ่งไม่ได้ใช้ความดันช่วงล่างในการทำนาย แต่กลับสอดคล้องกับแบบคาดการณ์ของ W.F. Wilson (FRS 1998) ที่มีความสามาถในการจำแนกต่ำกว่า

| | |
|---|---|
| **TITLE** | PREDICTION OF CORONARY HEART DISEASE BASED ON MEASURABLE EXTERNAL FACTORS USING NEURAL NETWORK |
| **KEYWORDS** | NEURAL NETWORK/ CHD/ CVD/ FEATURE/ SELECTION/ PREDICTION |
| **STUDENT** | KANOKPHAN CHOTICHAVARANON |
| **ADVISOR** | ASST.PROF PREECHA TANGKRAINGKIJ, PH.D. |
| **LEVEL OF STUDY** | MASTER OF SCIENCE INFORMATION TECHNOLOGY |
| **FACULTY** | INFORMATION TECHNOLOGY SRIPATUM UNIVERSITY |
| **ACADEMIC YEAR** | 2020 |

## ABTRACT

EVERY YEAR, THERE IS A GROWING POPULATION OF CORONARY HEART DISEASE (CHD). STILL, PEOPLE HAVE LITTLE ACCESS TO EXAMINATIONS, AND IT IS ONLY DETECTED WHEN THE DISEASE IS ALREADY DEVELOPED, SO IT IS DIFFICULT TO PREVENT IT. THIS RESEARCH AIMED TO STUDY CHD FACTORS AND DEVELOPED A CHD PREDICTION MODEL USING ONLY EXTERNALLY MEASURABLE FACTORS. THE DATA WENT THROUGH DATA PREPARATION AND FEATURE SELECTION BY A FORWARD SELECTION METHOD. THE RESULTS THAT HIGH EFFICIENCY WAS SELECTED TO DEVELOP AND TEST THE MODEL WITH A NEURAL NETWORK. THE RESULTS SHOWED THAT ONLY EXTERNALLY MEASURABLE FACTORS COULD USE TO CREATE A PREDICTION MODEL WITH A TOTAL ACCURACY GREATER THAN 90%, AUC MORE

SIGNIFICANT THAN 0.9, RECALL AND SPECIFICITY GREATER THAN 90%, USING THE EIGHT FACTORS WERE AGE, GENDER, SYSTOLIC BLOOD PRESSURE, DIASTOLIC BLOOD PRESSURE, HEIGHT, WEIGHT, SMOKING, AND BMI. THE EXTERNALLY MEASURABLE FACTORS THAT WERE USED CONTRADICTED THE ATP III HARD CHD RISK SCORE (FRS 2002) PREDICTION MODEL IN WHICH THE DIASTOLIC BLOOD PRESSURE WAS NOT APPLIED TO THE PREDICTION. INSTEAD, IT WAS CONSISTENT WITH PETER W.F. WILSON'S PREDICTION MODEL (FRS 1998) IDENTIFIED AS HAD A LOWER DISCRIMINATIVE PERFORMANCE.

# สารบัญ

# สารบัญตาราง

# สารบัญภาพ

# บทที่ 1

# บทนำ

## 1.1 ความเป็นมาและความสำคัญของปัญหา

โรคหลอดเลือดหัวใจ (ชื่อทั่วไป: Coronary Heart Disease: CHD, ชื่อทางการแพทย์: Coronary Artery Disease: CAD) เป็นโรคหัวใจและหลอดเลือด (Cardiovascular Disease: CVD) ชนิดหนึ่งที่เกิดจากการอุดตันของไขมันในหลอดเลือดโคโรนารี่ซึ่งจะทำให้เกิดโรคหัวใจขาดเลือด (Ischemic Heart Disease: IHD) อันเป็นเหตุถึงแก่ชีวิตได้ โรคหลอดเลือดหัวใจมักใช้เวลาในการพัฒนาของโรคในหลายปีจึงทำให้สังเกตุเห็นได้ยากจนกระทั่งพบว่ามีการอุดตันอย่างสมบูรณ์หรือมีอาการหัวใจวายแล้ว

รายงานขององค์กรอนามัยโลก (World Health Organization: WHO) พบว่ามีผู้เสียชีวิตทั่วโลกจากโรคหลอดเลือดหัวใจ ใน พ.ศ. 2553 คือ 122 คนต่อประชากร 100,000 คน และ พ.ศ. 2559 เพิ่มขึ้นเป็น 126 คนต่อประชากร 100,000 คน (WHO, 2561) ในประเทศไทย พ.ศ. 2561 อัตราความชุก (Prevalence rate) ของผู้ป่วยโรคหลอดเลือดหัวใจที่มีอายุ 15 ปี ขึ้นไป มีจำนวนประมาณ 1,397 คนต่อประชากรแสนคน (กองระบาดวิทยา กรมควบคุมโรค, 2562)

จากงานวิจัยของ Helen B. Hubert ในปี พ.ศ. 2526, Peter W.F. Wilsonในปี พ.ศ. 2530, Stanley S. Franklin ในปี พ.ศ. 2544, Christopher J.O'Donnell ในปี พ.ศ. 2551, Gary F. Mitchell ในปี พ.ศ. 2553 และMaruf Ahmed Tamal ในปี พ.ศ. 2562 ทำให้สามารถอนุมานได้ว่าปัจจัยในการก่อโรคหลอดเลือดหัวใจ นั้นเกี่ยวข้องกับ เพศ อายุ ความดัน ความอ้วน การเจ็บหน้าอก และการสูบบุหรี่ แต่จากงานวิจัยของ R P Fleet ในปี พ.ศ. 2537 พบว่ามีผู้ป่วยถึง 30% ในกลุ่มตัวอย่างที่มีอาการเจ็บหน้าอกแต่ไม่ได้เป็นโรคหลอดเลือดหัวใจ ดังนั้นการเจ็บหน้าอกไม่สามารถใช้ในการยืนยันโรคหลอดเลือดหัวใจที่แน่นอนได้ ในส่วนของปัจจัยอื่นนั้นเกือบทั้งหมดสามารถวัดค่าได้จากภายนอกc และการตรวจพบโรคนั้นจะพบเมื่อเป็นโรคแล้วเท่านั้น ขั้นตอนการตรวจค่อนข้างมีความซับซ้อนและมีราคาสูงทำให้มีคนจำนวนไม่น้อยไม่ได้รับการตรวจโรคส่งผลให้ไม่มีความตระหนักถึงการดูแลรักษาสุขภาพมิให้เกิดโรค

ผู้วิจัยจึงได้ทำการศึกษาการวิเคราะห์ข้อมูล The Original Cohort จาก Framingham Heart Study โดยใช้เทคนิคโครงข่ายประสาทเทียม และการหาคุณลักษณะที่เหมาะสมมาพยากรณ์ความเสี่ยงที่จะเป็นโรคจากปัจจัยที่วัดได้จากภายนอกเพื่อลดความซับซ้อนในการตรวจโรคและเพื่อให้ผู้คนได้ตระหนักถึงพฤติกรรมเสี่ยงที่จะก่อให้เกิดโรคหลอดเลือดหัวใจ

**1.2 คำถามการวิจัย**

1. จะสามารถใช้เพียงคุณลักษณะที่วัดได้จากภายนอกได้ในการพยากรณ์ได้หรือไม่
2. การพยากรณ์มีความแม่นยำเพียงใด
3. ใช้คุณลักษณะใดบ้างในการพยากรณ์
4. จำนวนชั้น และเซลล์ประสาทที่ใช้ในการสร้างโครงข่ายประสาทเทียม

**1.3 วัตถุประสงค์ของการวิจัย**

1. เพื่อศึกษาคุณลักษณะต่างๆที่มีผลต่อการเกิดโรคหลอดเลือดหัวใจ
2. เพื่อพัฒนาแบบจำลองการพยากรณ์การเกิดโรคหลอดเลือดหัวใจ
3. เพื่อพิสูจน์ทราบว่าสามารถใช้เฉพาะคุณลักษณะที่วัดได้จากภายนอกนั้นเพียงพอต่อการการพยากรณ์อย่างแม่นยำ

**1.4 สมมติฐานการวิจัย**

1. การใช้คุณลักษณะที่แตกต่างกันมีผลต่อความแม่นยำในการพยากรณ์
2. จำนวนชั้น และเซลล์ประสาทมีผลต่อความแม่นยำในการพยากรณ์
3. เฉพาะคุณลักษณะที่วัดได้จากภายนอกนั้นเพียงพอต่อการการพยากรณ์อย่างแม่นยำ

**1.5 ขอบเขตของการวิจัย**

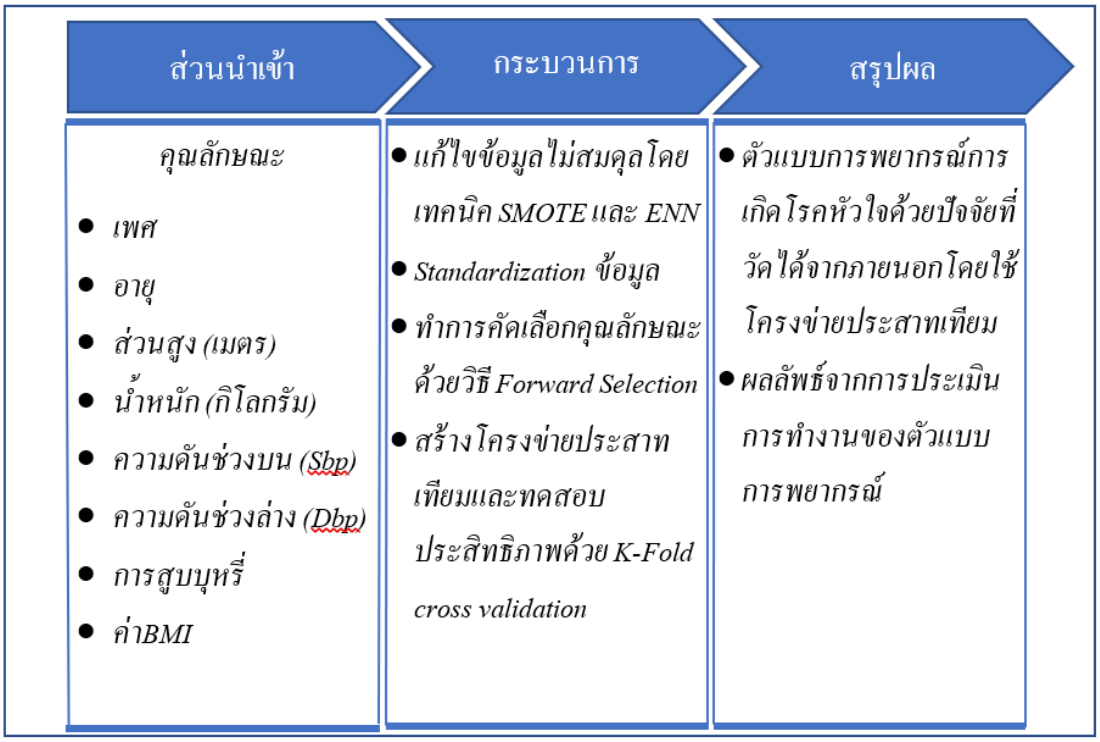
1. ข้อมูลที่นำมาใช้วิเคราะห์ข้อมูลคุณลักษณะปัจจัยที่วัดได้จากภายนอกของอาสาสมัครที่มาจาก The Original Cohortจาก Framingham Heart Study ตั้งแต่ปี พ.ศ. 2491-2501

2. ตัวแปรต้นของข้อมูลคุณลักษณะของอาสาสมัครที่สามารถวัดได้จากภายนอก ประกอบด้วย (1) เพศ (2) อายุ (3) ส่วนสูง(เมตร) (4) น้ำหนัก(กิโลกรัม) (5) ความดันช่วงบน (Sbp) (6) ความดันช่วงล่าง (Dbp) (7) การสูบบุหรี่ (8) ค่า BMI

3. ระยะเวลาในการดำเนินงานวิจัย 12 เดือน ตั้งแต่เดือน สิงหาคม พ.ศ. 2563 ถึง กรกฎาคม พ.ศ. 2564

**1.6 กรอบแนวคิดการวิจัย**

กระบวนการสังเคราะห์งานวิจัยที่มีกระบวนการที่ชัดเจนจะสามารถสรุปผลการวิเคราะห์ได้อย่างมีประสิทธิภาพผู้วิจัยจึงได้ดำเนินการสร้างกรอบแนวคิด ดังนี้




**ภาพประกอบที่ 1.1** กรอบแนวคิดการพัฒนาแบบจำลองการพยากรณ์การเกิดโรคหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้โครงข่ายประสาทเทียม

**1.7 ข้อจำกัดของการวิจัย**

1. กลุ่มข้อมูล Original Cohort ของ Framingham Heart Study นั้นไม่ได้วัดปัจจัยอย่างครอบคลุม เช่น อัตราการเต้นของหัวใจ
2. กลุ่มข้อมูล Original Cohort ของ Framingham Heart Study นั้นใช้มาตราอิมพีเรียลในการเก็บข้อมูลซึ่งผู้วิจัยได้ทำการแปลงเป็นมาตราเมตริกแล้ว
3. การพยากรณ์ในงานวิจัยนี้อยู่ในช่วง 10 ปีหลังวัดค่าปัจจัยเท่านั้น
4. สามารถพยากรณ์ว่าจะเป็นโรค (True) หรือไม่เป็น (False) เท่านั้น

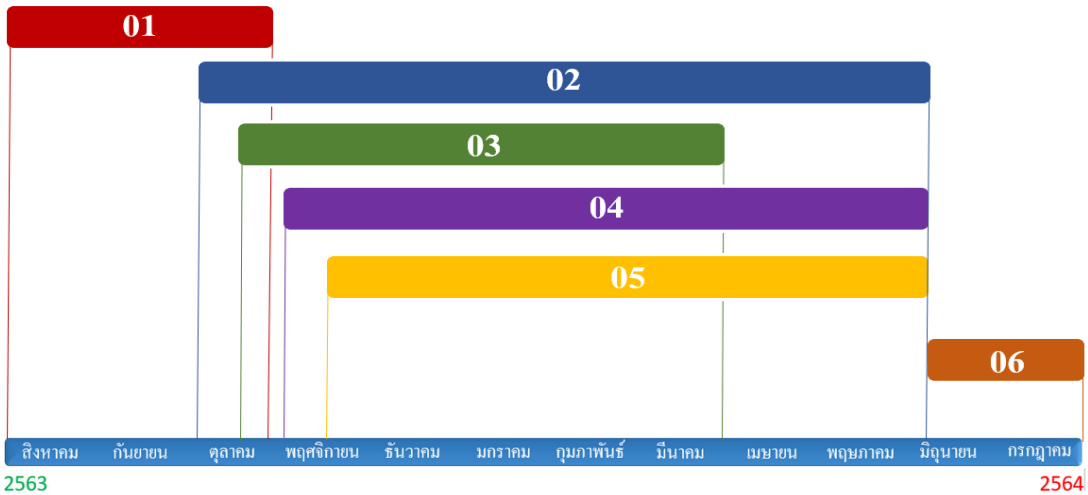
**1.8 ประโยชน์ที่คาดว่าจะได้รับ**

1. ได้องค์ความรู้ในการพัฒนาแบบจำลองการพยากรณ์ความเสี่ยงในการเกิดโรคหลอดเลือดหัวใจ

2. ช่วยวิเคราะห์ข้อมูลคุณลักษณะต่างๆที่มีผลต่อการเกิดโรคหลอดเลือดหัวใจ

3. ได้เผยแพร่ผลงานทางวิชาการ และ เป็นแนวทางในการขยายผลองค์ความรู้ด้านการวิเคราะห์ข้อมูลด้วยเทคนิคการเรียนรู้เชิงลึกในรูปแบบโครงข่ายประสาทเทียมต่อไปในอนาคต

**1.9 แผนการในการทำวิจัย**


**ภาพประกอบที่ 1.2**. แผนภูมิแสดงแผนการในการทำวิจัย

**1.10 คำนิยามศัพท์**

**1.10.1 โครงข่ายประสาทเทียม (Artificial Neural Networks)** คือ การสร้างคอมพิวเตอร์ที่จำลองวิธีการทำงานของสมองมนุษย์ที่ทำให้คอมพิวเตอร์คิดและจดจำในรูปแบบเดียวกับโครงข่ายประสาทของมนุษย์ เป็นหนึ่งในเทคนิคของการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นโมเดลทางคณิตศาสตร์ สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการจำแนก

**1.10.2 เซลล์ประสาท (Neuron)** เป็นส่วนประกอบของโครงข่ายประสาทเทียม โดยถ้าเป็น Input Layer ข้างในจะมีข้อมูลที่รับมา แต่ถ้าเป็น Hidden Layer จะมีสมการที่ช่วยในการคำนวณเพื่อการพยากรณ์ชนิดในการจำแนก (Classification) หรือคำนวณแบบถดถอย (Regression)และถ้าเป็น Output ก็จะคือคำตอบของการพยากรณ์

**1.10.3 การเรียนรู้แบบการแพร่กระจายย้อนกลับ (Back Propagation)** เป็นสถาปัตยกรรมที่กำหนดให้การส่งข้อมูลจากข้อมูลใน Input Layer เข้ามาภายใน Hidden Layer และส่งไปยัง Output Layer จะมีทิศทางในการไหลของข้อมูลไปในทิศทางเดียวกัน ข้อมูลที่ประมวลผลในวงจรข่ายจะถูกส่งไปในทิศทางเดียวจาก Input ส่งต่อมาเรื่อยๆ จนถึง Output และย้อนกลับจาก Output กลับมาที่ Input อีกครั้งโดยที่ Backpropagation หน้าที่ปรับค่า Weights ในแต่ละเส้นอีกครั้งโดยดูจาก Error/Cost ที่เกิดขึ้นในแต่ละเซลล์ประสาท

**1.10.4 ข้อมูลไม่สมดุล (Imbalanced Data)** คือข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของอีกกลุ่ม หนึ่งเป็นจำนวนมาก ตัวอย่างเช่น การจัดประเภท 2 ประเภท (ไบนารี) โดยมี 100 อินสแตนซ์ (แถว) อินสแตนซ์ทั้งหมด 90 รายการเป็นประเภทที่1 และ10 อินสแตนซ์ที่เหลือเป็นประเภทที่ 2 ชุดข้อมูลนี้จะเป็นชุดข้อมูลที่ไม่สมดุลและอัตราส่วนของอินสแตนซ์ ประเภทที่1ต่อ ประเภทที่ 2 คือ 9: 1

**1.10.5 Standardization (z-Score Normalization)** เป็นการปรับสเกลของข้อมูลเพื่อให้ง่ายในการเรียนรู้ของเครื่องโดยจะปรับให้ ค่าเฉลี่ย (Mean) = 0 และ ส่วนเบี่ยงเบนมาตราฐาน (Standard Deviation) = 1 (Unit Variance)

**1.10.6 วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE)** เป็นเทคนิคการสุ่มตัวอย่างของการสุ่มเพิ่ม โดยใช้ข้อมูลเดิมในการทำการสังเคราะห์ข้อมูลโดยใช้หลักการของ K-NN ซึ่งขั้นตอนในการสังเคราะห์ข้อมูลใหม่มีขั้นตอนคือระบุเพื่อนบ้านที่ใกล้เคียงที่สุด k จำนวน แล้วสุ่มเลือกจุด ระหว่างจุด k จุดและสร้างกรณีใหม่

**1.10.7 Edited Nearest Neighbor Rule (ENN)** เป็นการลบข้อมูลออกโดยจะลบข้อมูลที่เป็นคลาสส่วนใหญ่แต่เมื่อจำแนกด้วย K-NN แล้วกลายเป็นคลาสส่วนน้อย (โดยทั่วไป k = 3) ซึ่งเป็นการลบข้อมูลรบกวนทำให้ข้อมูลมีความนุ่มนวลขึ้นและช่วยลดความต้องการในพื้นที่การจัดเก็บข้อมูล

**1.10.8 Forward Selection** เป็นวิธีการเลือกคุณลักษณะโดยจะเพิ่มคุณลักษณะที่ช่วยปรับปรุงโมเดลไปเรื่อยๆจนกว่าจะไม่มีคุณลักษะที่ช่วยในการปรับปรุงโมเดลซึ่งวิธีนี้จะเสียทรัพยากรในการคำนวณสูงมาก

**1.10.9 K- Fold cross validation** เป็นหนึ่งในเทคนิคการทำResamplingโดยจะทำการแบ่งข้อมูลเป็น k ส่วนเท่าๆกันอย่าง random เพื่อสร้างและทดสอบโมเดลเช่น k=10 จะทำการแบ่งข้อมูลเป็น10ส่วนและใช้ชุดข้อมูลที่1-9ในการฝึกและชุดข้อมูลที่10ในการทดสอบซึ่งจะทำเช่นนี้เรื่อยๆโดยสลับชุดข้อมูลที่ใช้ทดสอบกับหนึ่งในชุดข้อมูลที่ใช้ฝึก ในแต่ละรอบจะบันทึกค่า validation error ไว้ด้วยเพื่อนำไปสรุปผลหลังจบกระบวนการ cross validation ทั้งหมด

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

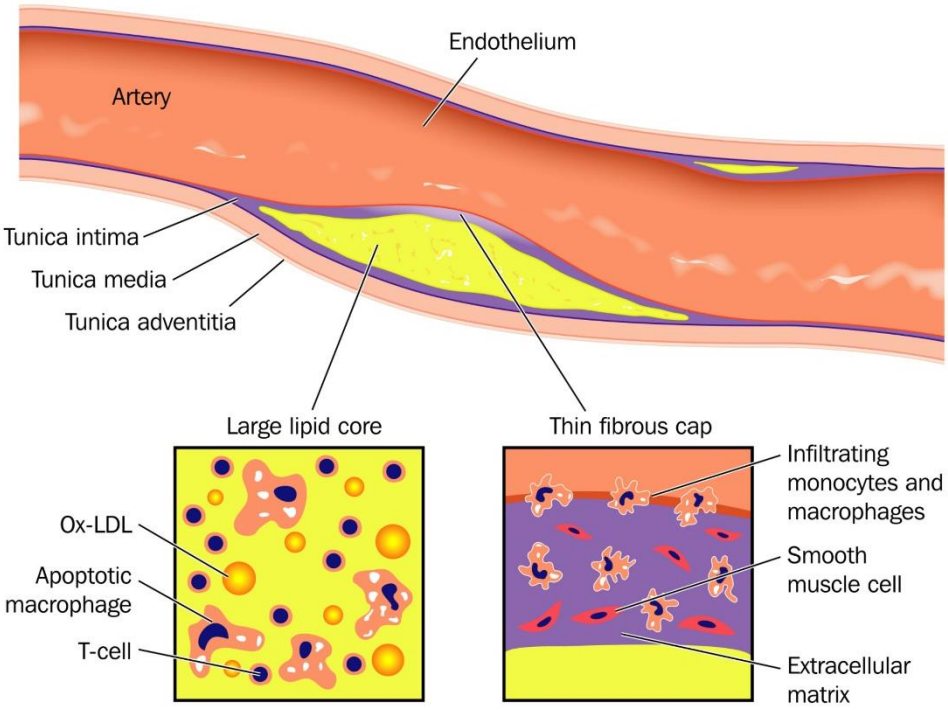
## 2.1 โรคหลอดเลือดหัวใจ (Coronary Heart Disease: CHD)

มีชื่อทางการแพทย์คือ Coronary Artery Disease (CAD) คือโรคหัวใจและหลอดเลือด (Cardiovascular Disease: CVD) ชนิดหนึ่งที่เกิดจากการมีคราบไขมันอุดตันในหลอดเลือดส่งผลให้หลอดเลือดตีบตันและไม่สามารถส่งเลือดไปยังกล้ามเนื้อหัวใจได้จนนำไปสู่โรคกล้ามเนื้อหัวใจขาดเลือด (Ischemic Heart Disease: IHD) จนเกิดอาการหัวใจวายอันเป็นเหตุให้ถึงแก่ชีวิตได้ รวมทั้งคราบไขมันอาจแตกกลายเป็นลิ่มเลือดส่งผลให้เกิดภาวะกล้ามเนื้อหัวใจตายเฉียบพลันอันเป็นเหตุถึงแก่ชีวิตเช่นกัน

โรคหลอดเลือดหัวใจจะใช้เวลาหลายปีในการพัฒนาจนกระทั่งแสดงอาการเมื่อมีการอุดตันอย่างสมบูรณ์หรือมีอาการหัวใจวายอันเป็นเหตุให้สังเกตอาการได้ยาก จากรายงานขององค์การอนามัยโลก (World Health Organization: WHO) พบว่ามีผู้เสียชีวิตจากโรคหลอดเลือดหัวใจ ในปี พ.ศ. 2553 จำนวน 122 คนต่อ 100,000 คน และเพิ่มขึ้นเป็น 126 คนต่อ 100,000 คนในปี พ.ศ. 2559 (WHO, 2561) โดยในประเทศไทยมีอัตราความชุกของผู้ป่วยโรคหลอดเลือดหัวใจ ประมาณ 1397 คนต่อ 100,000 คน (กองระบาดวิทยา กรมควบคุมโรค, 2562)

2.1.1 อาการของผู้ป่วยโรคหลอดเลือดหัวใจ

- เจ็บแน่นหน้าอก คือ รู้สึกแน่นหน้าอกเหมือนถูกกดทับโดยจะเกิดจากความเครียดและยังพบอาการเดียวกันได้ที่ไหล่ แขน คอ กราม หรือหลัง
- หายใจหอบถี่ คือ เมื่อร่างกายไม่ได้รับออกซิเจนจากเลือดที่เพียงพอ จะทำให้หอบหายใจถี่ขึ้นเมื่อทำกิจกรรม
- หัวใจวาย คือ เมื่อหลอดเลือดถูกอุดตันโดยสมบูรณ์อาจทำให้เกิดอาการหัวใจวาย
- หัวใจล้มเหลว คือ หัวใจจะอ่อนแรงจนไม่สามารถสูบฉีดเลือดไปเลี้ยงอวัยวะส่วนต่างๆ ของร่างกายได้ ผู้ป่วยจะหายใจติดขัดจากภาวะน้ำท่วมปอดเป็นได้ทั้งแบบเฉียบพลันและเรื้อรัง

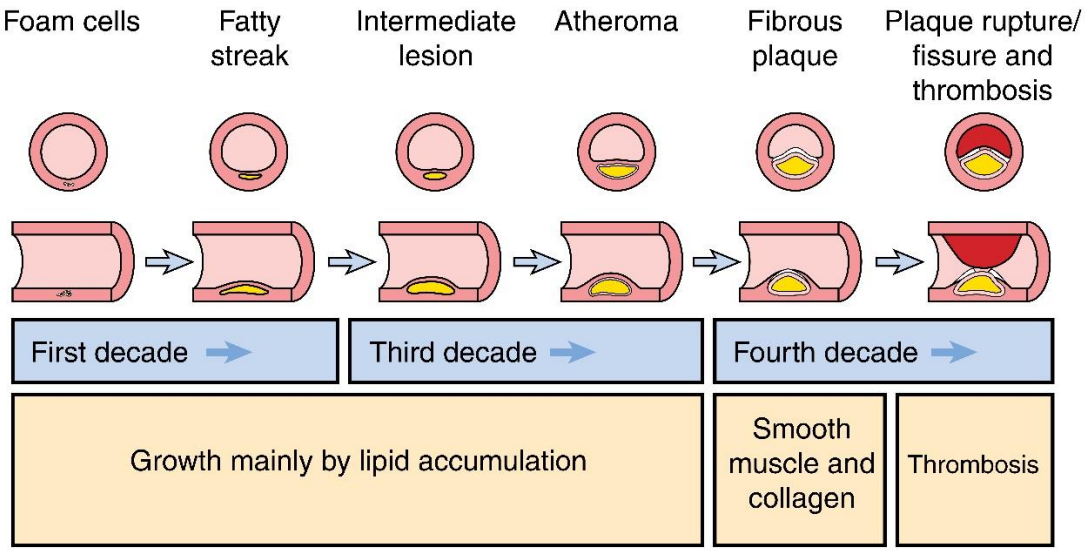



**ภาพประกอบที่ 2.1** คราบไขมันบริเวณผนังหลอดเลือดเมื่อพอกปริมาณมากจะลดการไหลเวียนของเลือด

2.1.2 สาเหตุของโรคหลอดเลือดหัวใจ

เกิดจากคราบไขมันกับของเสียอื่นๆ เกิดเป็นก้อนแข็ง เรียกว่า Atheromatous plaque (ภาพประกอบที่ 2.1) ซึ่งจะส่งผลให้เกิดภาวะหลอดเลือดแข็งตัว (Atherosclerosis) ซึ่งปัจจัยที่ส่งผลต่อการเกิดกระบวนการเหล่านี้ ได้แก่

- คอเลสตอรอล คือ ไขมันมี 2 ประเภทสำคัญ คือ
  - LDL (Low Density Lipoprotein) ที่จะปิดกันหลอดเลือด
  - HDL (High Density Lipoprotein) ที่ต่อต้านการสะสมของ LDL
- ความดันโลหิต ความดันโลหิตที่สูงจะทำให้เกิดภาวะหลอดเลือดแข็งตัวโดยจะแบ่งเป็น
  - ความดันช่วงบน (Systolic Pressure) ที่วัดขณะหัวใจบีบตัว
  - ความดันช่วงล่าง (Diastolic Pressure) วัดขณะหัวใจคลายตัว




**ภาพประกอบที่ 2.2** การก่อตัวของ Atheromatous plaque ในหลอดเลือดและในที่สุดจะแข็งตัวเป็นลิ่มเลือดที่อุดตันหลอดเลือดอย่างสมบูรณ์

- ภาวะหลอดเลือดอุดตัน (Thrombosis) เป็นผลจากการเกิดภาวะหลอดเลือดแข็งตัว โดยจะเป็นก้อนลิ่มเลือดขนาดใหญ่ที่ขัดขวางการสูบฉีดของเลือด (ภาพประกอบที่ 2.2)
- การสูบบุหรี่ สารนิโคตินและแก๊สคาร์บอนมอนออกไซด์จะเพิ่มความเครียดให้กับหัวใจและเพิ่มโอกาสการเกิดของลิ่มเลือด
- โรคเบาหวาน การมีระดับน้ำตาลในเลือดสูงเป็นเวลานานจะเพิ่มโอกาสในการเป็นโรคหลอดเลือดหัวใจ มากขึ้น
- เพศ โดยทั่วไปเพศชายจะมีความเสี่ยงมากกว่าเพศหญิง

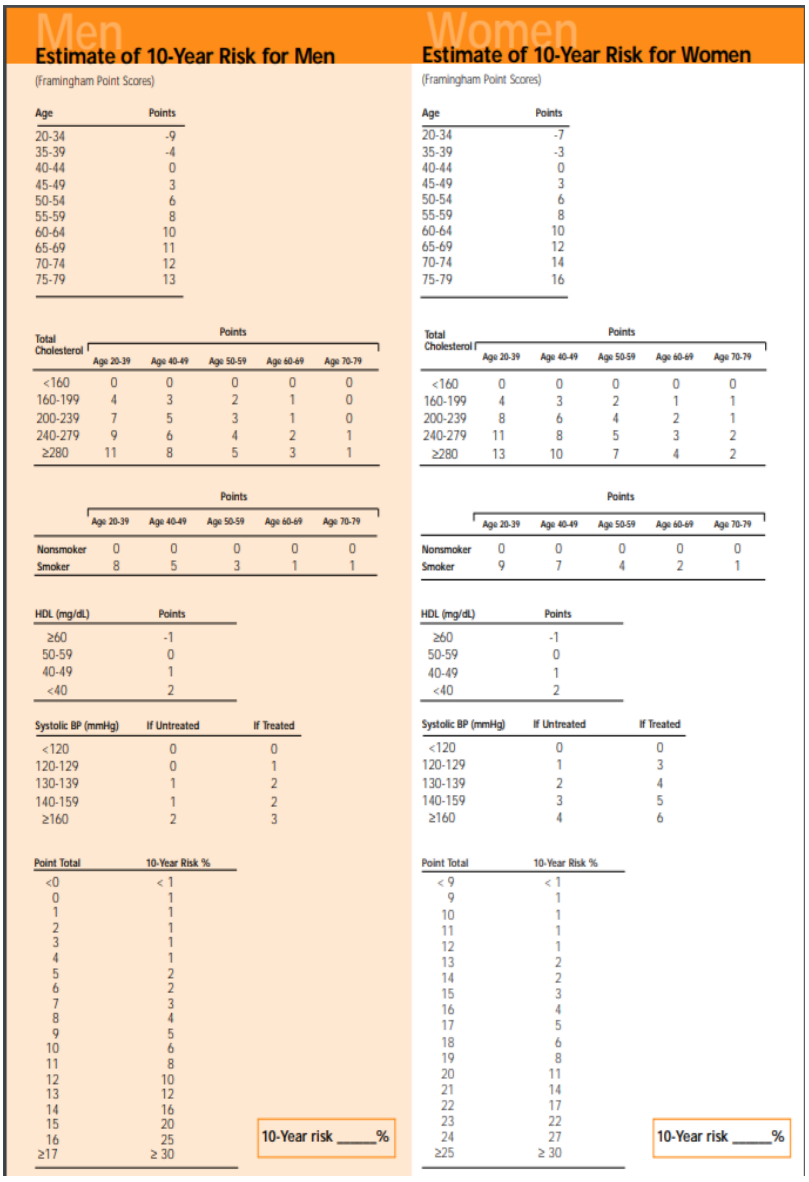
**2.2 Framingham Heart Study**

เป็นการศึกษาหลอดเลือดและหัวใจแบบต่อเนื่องในเมือง Framingham รัฐ Massachusetts ประเทศสหรัฐอเมริกา เริ่มในปีพ.ศ. 2491 โดยเป็นโครงการของสถาบันหัวใจ ปอดและหลอดเลือดแห่งชาติ (National Heart, Lung and Blood Institute) ร่วมมือกับมหาวิทยาลัยบอสตัน

2.2.1 Framingham Risk Score (FRS)

เป็นแบบประเมินความเสี่ยงที่จะเป็นโรคหลอดเลือดหัวใจแบบจำเพาะต่อเพศโดยในปัจจุบันมีมาแล้ว 3 ฉบับคือ Original FRS ในปี 1998, ATP III Hard CHD Rick Score ในปี 2002

และ Framingham General Cardiovascular Risk Score ในปี 2008 โดย 2 ฉบับแรกจะมีความจำเพาะต่อ CHD มากกว่าฉบับ 3 ที่จะทำนายความเสี่ยงโรคกลุ่ม CVD แบบองค์รวม โดยฉบับ 2002 ที่มีความแม่นยำมากกว่า (J. A. Damen, 2019) จะใช้ เพศ อายุ คลอเรสเตอรอล HDL การสูบบุหรี่ และความดันช่วงบนในการทำนายโดยได้สร้างแบบทดสอบได้ตามภาพประกอบที่ 2.3



**Men**

**Estimate of 10-Year Risk for Men**

(Framingham Point Scores)

| Age | Points |
|---|---|
| 20-34 | -9 |
| 35-39 | -4 |
| 40-44 | 0 |
| 45-49 | 3 |
| 50-54 | 6 |
| 55-59 | 8 |
| 60-64 | 10 |
| 65-69 | 11 |
| 70-74 | 12 |
| 75-79 | 13 |

| Total Cholesterol | Points | | | | |
|---|---|---|---|---|---|
| | Age 20-39 | Age 40-49 | Age 50-59 | Age 60-69 | Age 70-79 |
| <160 | 0 | 0 | 0 | 0 | 0 |
| 160-199 | 4 | 3 | 2 | 1 | 0 |
| 200-239 | 7 | 5 | 3 | 1 | 0 |
| 240-279 | 9 | 6 | 4 | 2 | 1 |
| ≥280 | 11 | 8 | 5 | 3 | 1 |

| | Points | | | | |
|---|---|---|---|---|---|
| | Age 20-39 | Age 40-49 | Age 50-59 | Age 60-69 | Age 70-79 |
| Nonsmoker | 0 | 0 | 0 | 0 | 0 |
| Smoker | 8 | 5 | 3 | 1 | 1 |

| HDL (mg/dL) | Points |
|---|---|
| ≥60 | -1 |
| 50-59 | 0 |
| 40-49 | 1 |
| <40 | 2 |

| Systolic BP (mmHg) | If Untreated | If Treated |
|---|---|---|
| <120 | 0 | 0 |
| 120-129 | 0 | 1 |
| 130-139 | 1 | 2 |
| 140-159 | 1 | 2 |
| ≥160 | 2 | 3 |

| Point Total | 10-Year Risk % |
|---|---|
| <0 | < 1 |
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |
| 6 | 2 |
| 7 | 3 |
| 8 | 4 |
| 9 | 5 |
| 10 | 6 |
| 11 | 8 |
| 12 | 10 |
| 13 | 12 |
| 14 | 16 |
| 15 | 20 |
| 16 | 25 |
| ≥17 | ≥ 30 |

10-Year risk _____%

**Women**

**Estimate of 10-Year Risk for Women**

(Framingham Point Scores)

| Age | Points |
|---|---|
| 20-34 | -7 |
| 35-39 | -3 |
| 40-44 | 0 |
| 45-49 | 3 |
| 50-54 | 6 |
| 55-59 | 8 |
| 60-64 | 10 |
| 65-69 | 12 |
| 70-74 | 14 |
| 75-79 | 16 |

| Total Cholesterol | Points | | | | |
|---|---|---|---|---|---|
| | Age 20-39 | Age 40-49 | Age 50-59 | Age 60-69 | Age 70-79 |
| <160 | 0 | 0 | 0 | 0 | 0 |
| 160-199 | 4 | 3 | 2 | 1 | 1 |
| 200-239 | 8 | 6 | 4 | 2 | 1 |
| 240-279 | 11 | 8 | 5 | 3 | 2 |
| ≥280 | 13 | 10 | 7 | 4 | 2 |

| | Points | | | | |
|---|---|---|---|---|---|
| | Age 20-39 | Age 40-49 | Age 50-59 | Age 60-69 | Age 70-79 |
| Nonsmoker | 0 | 0 | 0 | 0 | 0 |
| Smoker | 9 | 7 | 4 | 2 | 1 |

| HDL (mg/dL) | Points |
|---|---|
| ≥60 | -1 |
| 50-59 | 0 |
| 40-49 | 1 |
| <40 | 2 |

| Systolic BP (mmHg) | If Untreated | If Treated |
|---|---|---|
| <120 | 0 | 0 |
| 120-129 | 1 | 3 |
| 130-139 | 2 | 4 |
| 140-159 | 3 | 5 |
| ≥160 | 4 | 6 |

| Point Total | 10-Year Risk % |
|---|---|
| < 9 | < 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 1 |
| 12 | 1 |
| 13 | 2 |
| 14 | 2 |
| 15 | 3 |
| 16 | 4 |
| 17 | 5 |
| 18 | 6 |
| 19 | 8 |
| 20 | 11 |
| 21 | 14 |
| 22 | 17 |
| 23 | 22 |
| 24 | 27 |
| ≥25 | ≥ 30 |

10-Year risk _____%

**ภาพประกอบที่ 2.3** แบบคำนวนความเสี่ยงของ Framingham Heart Study 2002, Reprinted from The National Heart, Lung, and Blood Institute, NIH Publication No. 01-3305, No. 01-3670

2.2.2 ความถูกต้องในการทำนาย

Framingham Risk Score ได้รับการตรวจยืนยันแล้วในสหรัฐอเมริกาโดยสามารถใช้งานได้ดีในหมู่คนผิวขาวและผิวดำโดยจากงานวิจัยของ D'Agostino RB, Grundy S, Sullivan LM และWilson P ในปี พ.ศ. 2554 แสดงว่ามีความแม่นยำของ Framingham Risk Score ในเพศชายมีค่า AUC ที่ 0.79 และเพศหญิงที่ 0.83 (ตารางที่ 2.1) และมีความแม่นยำ AUC ที่ 0.63 ในการทดสอบกับชาวเอเชียในงานวิจัยของ Chia YC, Gray SYW, Ching SM และคณะในปี พ.ศ. 2558 (ตารางที่ 2.2)

**ตารางที่ 2.1** ผลการตรวจสอบความถูกต้องของ The Framingham Risk Score ที่ถูกใช้กับ The Framingham Heart Study Offspring Cohort และชุดข้อมูลอื่นๆ

| **Dataset** | **FHS** | **ARIC** | | **PHS** | **HHP** | **PR** | **SHS** | **CHS** |
|---|---|---|---|---|---|---|---|---|
| **Ethnic Groups** | **White** | **White** | **Black** | **White** | **Japanese American** | **Hispanic** | **Native American** | **White** |
| Men | 0.79 | 0.75 | 0.67 | 0.63 | 0.72 | 0.69 | 0.69 | 0.63 |
| Women | 0.83 | 0.83 | 0.79 | - | - | - | 0.75 | 0.66 |

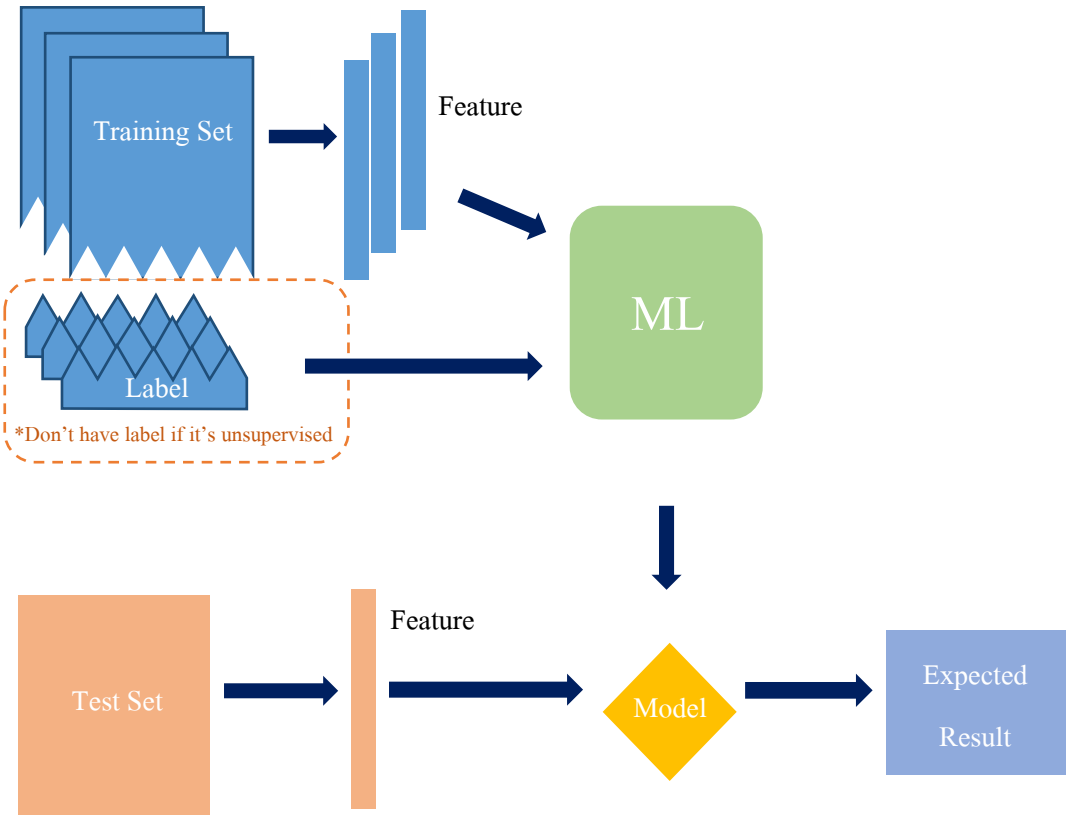
**ตารางที่ 2.2** ผลการตรวจสอบความถูกต้องของ The Framingham Risk Score ในกลุ่มตัวอย่างชาวเอเชีย

| **Title** | **CVD event** | | **Result** | |
|---|---|---|---|---|
| | **Observed** | **Predicted** | | |
| | **N (%)** | **N (%)** | **AUC (95% CI)** | **p Value** |
| Overall (N=967) | 127 (13.1) | 204 (21.1) | 0. 63 (0.58 to 0.68) | <0.001 |
| Malay (N=225) | 26 (11.6) | 42 (18.5) | 0. 65 (0.53 to 0.77) | 0.014 |
| Chinese (N=438) | 45 (10.3) | 102 (23.2) | 0. 60 (0.52 to 0.68) | 0.027 |
| Indian (N=291) | 54 (18.6) | 63 (21.5) | 0. 65 (0.57 to 0.73) | 0.001 |

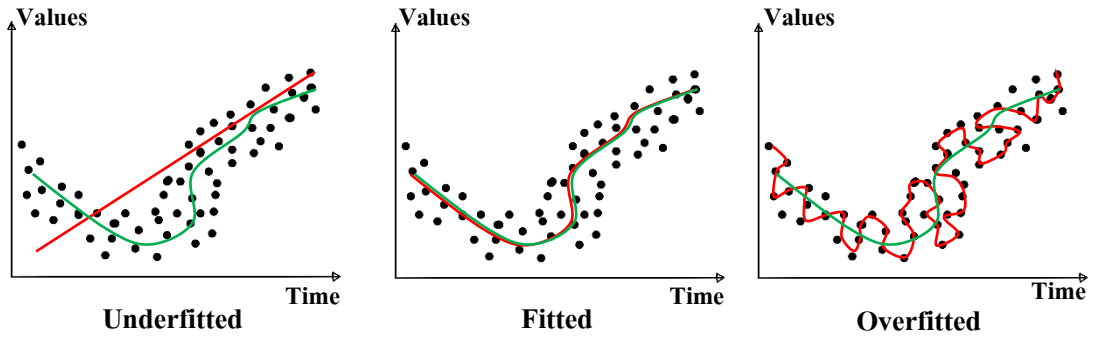
**2.3 การเรียนรู้ของเครื่อง (Machine Learning: ML)**

คือ ระบบที่สามารถเรียนรู้ได้ด้วยตนเอง เพื่อที่จะหาอัลกอริธึมของแบบจำลอง (Model) ที่เหมาะสมในการเชื่อมโยงข้อมูลและผลลัพธ์โดยจะเป็นการใส่ข้อมูลและผลลัพธ์ลงไปเพื่อสร้างแบบจำลอง ในการสร้างแบบจำลองจะใช้ข้อมูลที่แบ่งเป็น 2 ชุด คือ

- ชุดข้อมูลฝึก (Training Set) ใช้สำหรับการฝึก ML
- ชุดข้อมูลทดสอบ (Test Set) เป็นข้อมูลใช้ทดสอบแบบจำลองที่สร้างออกมาโดยต้องแยกออกมาจาก Training Set เพื่อป้องกันการเกิดความลำเอียงจากการรู้ข้อมูลก่อน




**ภาพประกอบที่ 2.4** ผังงานการเรียนรู้ของเครื่อง




**ภาพประกอบที่ 2.5** กราฟแสดง Under Fitting / Good Fitting / Over Fitting. สีเขียวคือฟังชั่นท์จริง และ สีแดงคือแบบจำลอง

เมื่อทดสอบด้วย Test Set แล้ว ถ้าผลลัพธ์ออกมาไม่แม่นยำหรือแม่นยำน้อยมากจะเรียกว่าเกิดปัญหา Under Fitting หรือ Over Fitting (ภาพประกอบที่ 2.5)

- Under Fitting คือ ฝึกออกมาไม่เข้ากับ Training Set ทำให้ผลการทำนายไม่แม่นยำ
- Over Fitting คือ ฝึกออกมาได้เข้ากับ Training Set มากเกินไป จนไม่สามารถทำนายข้อมูลของ Test Set ได้แม่นยำ

**2.4 การเรียนรู้แบบมีผู้สอน (Supervised Learning)**

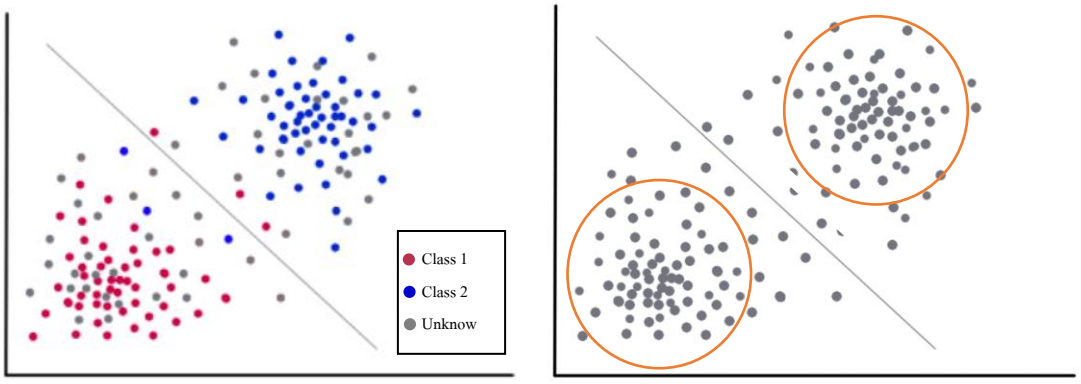
จะทำให้คอมพิวเตอร์หาคำตอบของปัญหาได้ หลังจากการเรียนรู้ด้วยข้อมูลและผลลัพธ์ไประยะหนึ่ง มี 2 รูปแบบ ดังนี้

2.4.1 Classification เป็นการจำแนกชนิด โดยจะให้เครื่องเรียนรู้การแยกกลุ่ม เปรียบเครื่องเป็นเด็กไม่มีความรู้ ผู้สอนจะสอนว่าสัตว์ตัวไหนเป็นแมว ตัวไหนเป็นสุนัขไปเรื่อยๆ ระยะหนึ่ง (Training Set) จากนั้นนำแมวและสุนัขที่ไม่เคยสอนมารวมกัน (Test Set) แล้วให้เด็กจำแนกว่าตัวไหนคือแมวหรือสุนัข

2.4.2 Regression เป็นการคาดคะเนจากแนวโน้มของข้อมูล เช่น สอนเครื่องว่าเพชรขนาด 4 กะรัต สีเหลือง ระดับความสะอาด VS2 มีราคา 8 ล้านบาท และ 4 กะรัต สีแดง ระดับความสะอาด VS1 ราคา 12 ล้านบาท (Training Set) ทำเช่นนี้ไปเรื่อยๆ จนเครื่องสามารถคาดเดาราคาเพชรที่เครื่องไม่เคยเห็นได้ (Test Set)

**2.5 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)**

จะทำให้คอมพิวเตอร์สามารถหาคำตอบของปัญหาได้ด้วยตนเองโดยไม่มีตัวอย่างจึงสามารถทำได้เพียงการแบ่งกลุ่ม (Clustering) เท่านั้น




Supervised Unsupervised

**ภาพประกอบที่ 2.6** การเรียนรู้แบบมีผู้สอน (ซ้าย) และการเรียนรู้แบบไม่มีผู้สอน (ขวา)

**2.6 การเรียนรู้เชิงลึก (Deep Learning: DL)**

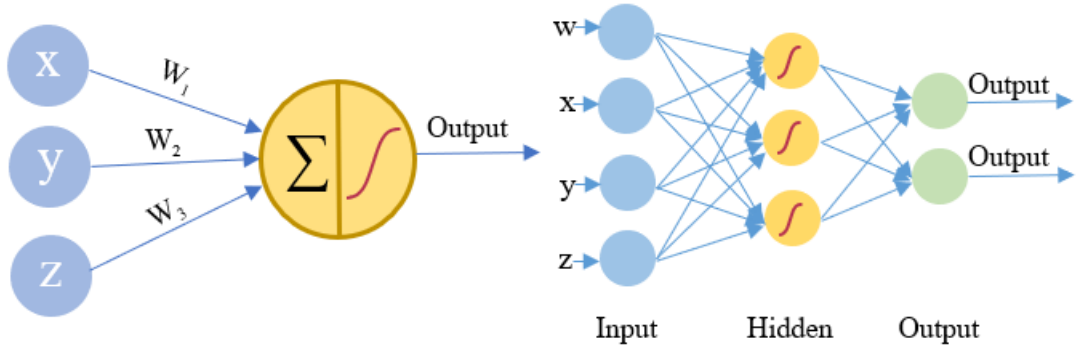
เป็น ML ที่จำลองการทำงานของโครงข่ายประสารทของมนุษย์ โดยจะนำไปใช้ในการตรวจจับรูปแบบหรือการจำแนกข้อมูล

2.6.1 โครงข่ายประสาทเทียม (Neural Networks: NNs) เป็นโมเดลทางคณิตศาสตร์สำหรับประมวลผลแบบคอนเนคชันนิสต์ (Connectionist) แนวคิดเริ่มต้นมาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ ซึ่งมีเซลล์ประสาท (Neuron) และจุดประสาน (Synapses) โดยจะทำงานเป็นเครือข่ายร่วมกัน

โดยในยุคแรกถูกคิดค้นโดย Rosenblatt เรียกว่า 1-Layer Perceptron โดยที่ 1 นิวรอนจะทำนายได้ 1 คลาส ถ้ามี Output มากกว่า จำเป็นต้องมีหลาย 1-Layer Perceptron โดยในนิวรอนจะมี Weight Sum คือ การรวมน้ำหนักที่ต่อเข้านิวรอน

แต่เนื่องจาก 1-Layer Perceptron ไม่สามารถแก้ปัญหา XOR Problem ได้ จึงเกิดMulti-Layer Perceptron โดยการเพิ่ม Hidden Layer เข้าไป

NNs จะทำการเพิ่มประสิทธิภาพโดยการฝึกซ้ำๆ Input ที่เข้ามาสามารถเป็นคุณลักษณะจาก Training Set หรือ Output จาก Layer ก่อน โดยจะมีการนำ Weight มาปรับค่า Input ในแต่ละนิวรอน และนำผลรวมของ Weight มาปรับค่าด้วย Activation Function แล้วส่งไปยังนิวรอน ถัดไปโดยจะทำการเรียนรู้ในการหาชุดของWeight และBiases โดยสิ้นเปลืองน้อยที่สุด เรียกว่าGradient Descent.




**ภาพประกอบที่ 2.7** 1-Layer Perceptron (ซ้าย) และ Multi-Layers Perceptron (ขวา)

2.6.1.1 NNs จะมีลักษณะ มี 3 Layers คือ

1) Input Layer ที่จะรับข้อมูลที่เข้ามา โดยจำนวนของนิวรอน จะขึ้นกับจำนวน Input และจะมีเพียงชั้นเดียว

2) Hidden Layers สามารถมีมากกว่า 1 ชั้นได้ โดยพื้นฐานถ้าต้องการความแม่นยำที่เพิ่มขึ้น จะเพิ่มจำนวนชั้นและจำนวน นิวรอน ภายใน Hidden Layer จะมี Activation Function อยู่ภายใน

3) Output Layer เป็นชั้นสุดท้ายที่จะมี Weight ของคลาสที่ต้องการ

2.6.1.2 NNs มีส่วนประกอบดังต่อไปนี้

- นิวรอน (Neuron) โดยถ้าเป็น Input Layer จะมีข้อมูลที่รับมา ถ้าเป็น Hidden Layer จะมีฟังก์ชันที่ใช้คำนวณ เพื่อการจำแนกหรือการคำนวณแบบถดถอย (Regression) และถ้าเป็น Output ก็คือผลลัพธ์ในการทำนาย

- Weight (Coefficients) เป็นค่าสัมประสิทธิ์ในแต่ละ นิวรอน ซึ่งจะถูกปรับเมื่อเทียบกับค่า Error ซึ่งจะทำซ้ำจนกว่า Error จะลดลงจนยอมรับได้

- Bias เป็นค่าคงที่ที่มากับ นิวรอน และจะนำไปบวกค่า Weight ก่อนที่จะเข้า Activation Function

- Synapse เป็นถนนใน Neural Network ที่เชื่อมระหว่าง นิวรอน ซึ่งในแต่ละการเชื่อมต่อ จะมีค่า Synapses และ Weight ที่ไม่ซ้ำกัน

- Activation Function เป็นฟังก์ชันใน Hidden Layer มีความซับซ้อนแบบ Non-Linear Relationship โดยจะแก้ไขค่าที่เข้ามาให้ง่ายต่อการคำนวณในชั้นถัดไป

**2.7 ReLU Function (Rectified Linear Unit Function)**

เป็น Activation Function เส้นตรงที่ถูกปรับแก้ไม่ให้เป็นรูป Sโดยถ้า Input ที่เข้ามาเป็นบวกจะมี slope = 1 ทำให้ไม่เกิดปัญหาที่ Weight ไม่ถูกอัพเดตและอนุพันธ์ที่เป็นไปได้จะเป็น 0 หรือ 1 เท่านั้น ส่งผลให้ฝึก โมเดลได้รวดเร็วโดยจะมีสมการ ดังสมการ (2.1) และมีอนุพันธ์ดังสมการที่ (2.2)

$$f(x) = \max(0, x) = \begin{cases} 0 \; for \; x \leq 0 \\ x \; for \; x > 0 \end{cases} \tag{2.1}$$

$$f'(x) = \begin{cases} 0 \; for \; x \leq 0 \\ 1 \; for \; x > 0 \end{cases} \tag{2.2}$$

**2.8 Optimizer**

จะปรับปรุงค่า Error และ Loss เพื่อให้ได้ค่าที่ดี โดยจะเปลี่ยนแปลงค่า Weight และ Bias เพื่อให้มีค่า Loss ออกมาน้อยที่สุด

2.8.1 SGD (Stochastic gradient descent) เป็น Optimizer ที่แบ่งGradientเป็นส่วนๆก่อนจะทำการหาค่า Weight และ Bias โดยจะอัพเดตค่าพารามิเตอร์ 1 ครั้ง ต่อการฝึก 1 รอบโดยจะอัพเดททีละตัวอย่าง ทุกครั้งที่อัพเดตค่าพารามิเตอร์ จะมีค่าความแปรปรวนสูงและส่งผลกับค่า Loss ดังสมการ (2.3) เมื่อ $x$ คือ ตัวอย่าง, $y$ คือ ผลลัพธ์, $\theta$คือ พารามิเตอร์ และ $\eta$ คือ อัตราการเรียนรู้

$$\theta = \theta - \eta \cdot \nabla_{\theta} J\left(\theta; x^{(i)}; y^{(i)}\right) \tag{2.3}$$

2.8.2 ADAM (Adaptive Moment Estimation) เป็น Optimizer ที่สามารถแก้ปัญหาที่ต้องหา Learning rate ให้เหมาะสม โดยจะปรับค่า Learning rate ตามพารามิเตอร์แต่ละตัว ทำให้มีความเร็วในการฝึก และลดการแกว่งของพารามิเตอร์ที่มากเกินไป โดยใช้ Exponential Weighted Moving Average (EMA) ของ Gradient และของ Gradient$^2$ มาอัพเดตพารามิเตอร์ใช้เป็นUpdater ร่วมกับ SGD

**2.9 การเรียนรู้แบบการแพร่กระจายย้อนกลับ (Back Propagation)**

เป็นสถาปัตยกรรมที่ส่งข้อมูลจาก Input Layer ไป Output Layer จากนั้นจะทำการปรับค่า Weight โดยใช้ค่า Error แพร่ย้อนกลับไปแก้ไข Weight ให้สอดคล้องและใกล้เคียงกับผลลัพธ์ที่คาดไว้ โดยนำ Output จริงเทียบกับ Output ที่กำหนดไว้เพื่อหาค่า Error ในแต่ละเถว ดังสมการ (2.4)

$$e^q = \frac{1}{2}\sum_{j=1}^{J}\left(t_j - z_j\right)^2 \tag{2.4}$$

จากนั้นปรับค่า weight ระหว่างนิวรอนใน Hidden Layer และ Output Layer ดังสมการ (2.5)

$$W_{mj}^{(r+1)} = W_{mj}^{(r)} + \eta\{\left(t_j^{(q)} - z_j^{(q)}\right) * \left[z_j^{(q)}\left(1 - z_j^{(q)}\right)\right] * y_m^{(q)} \tag{2.5}$$

และปรับค่า weight ระหว่างนิวรอนใน Input Layer และ Hidden Layer ดังสมการ (2.6)

$$W_{nm}^{(r+1)} = W_{nm}^{(r)} + \eta\left\{\sum_{j=1}^{J}\left(t_j^{(q)} - z_j^{(q)}\right)\left[z_j^{(q)}\left(1 - z_j^{(q)}\right)\right]W_{mj}^{(r)}\right\} * [y_m^{(q)}\,(1 - y_m^{(q)})][x_n^{(q)}] \tag{2.6}$$

จากนั้นคำนวณค่าผิดพลาดรวมเฉลี่ย (Mean Squared Error: MSE)ด้วยสมการ (2.7)

$$E = \frac{1}{Q}\sum_{q=1}^{Q} e^{(q)} \tag{2.7}$$

แล้วตรวจErrorจริงว่าน้อยกว่า Error ที่กำหนดไว้หรือไม่ และครบรอบการฝึกแล้วหรือไม่ถ้าครบข้อใดข้อหนึ่งให้จบการฝึก ถ้าไม่ให้เริ่มการฝึกรอบใหม่ โดยตัวแปรทั้งหมดมีค่าดังนี้

$y_m$ = Output ของ Hidden Layerหลังทำการปรับค่าของนิวรอนที่ $m$ จากทั้งหมด $M$ นิวรอน

$z_j$ = Output ของ Output Layerหลังทำการปรับค่าของนิวรอนที่ $j$ จากทั้งหมด $J$ นิวรอน

$t_j$ = Output ที่ต้องการจาก Output Layer ของนิวรอนที่ $j$ จากทั้งหมด $J$ นิวรอน

$W_{nm}$ = Weight ของเส้นเชื่อมระหว่าง Input Layer และ Hidden Layer

$W_{mj}$ = Weight ของเส้นเชื่อมระหว่าง Hidden Layer และ Output Layer

$\eta$ = อัตราการเรียนรู้

$R$ = จำนวนรอบที่จะทำการฝึกมี $R$ เป็นจำนวนรอบที่กำหนด

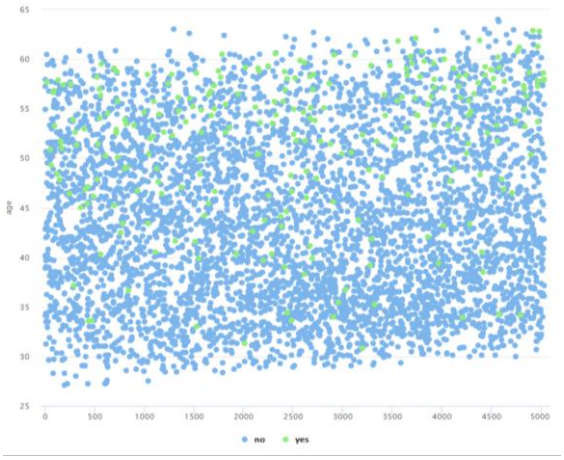
$Q$ = จำนวนชุดข้อมูลตัวอย่าง มี $Q$ เป็นตัวกำหนด

$e^{(q)}$ = ค่าผิดพลาดของข้อมูลตัวอย่าง

$E$ = ค่าผิดพลาดรวมเฉลี่ยของข้อมูลตัวอย่าง

**2.10 ข้อมูลที่ไม่สมดุล (Imbalanced Data)**

คือ ข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งมาก เช่น การจำแนก 2 ประเภท โดยมี 1,000 อินสแตนซ์ (แถว) 990 อินสแตนซ์ เป็นประเภทที่ 1 และ 10 อินสแตนซ์ที่เหลือเป็นประเภทที่ 2




**ภาพประกอบที่ 2.8** ตัวอย่างข้อมูลไม่สมดุลย์

**2.11 วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE)**

เป็นหนึ่งในวิธีการสุ่มเกินเพื่อแก้ปัญหาข้อมูลที่ไม่สมดุลย์โดยจะทำการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อยให้มีจำนวนใกล้เคียงคลาสส่วนมากโดยจะสร้างข้อมูลสังเคราะห์จากข้อมูลตัวอย่างด้วยการวัดระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดใกล้เคียงที่สุดตามจำนวน k (K-NN) แล้วสุ่มสร้างข้อมูล โดยข้อมูลที่สร้างขึ้นจะอยู่ภายในระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียงที่จะถูกสุ่มเลือกอีกครั้ง ซึ่งจุดใหม่จะแสดงดังสมการ (2.8)

$$N_{point} = O_{point} + \left(Random[0,1] * distance(x, y, \dots, z)\right) \qquad (2.8)$$

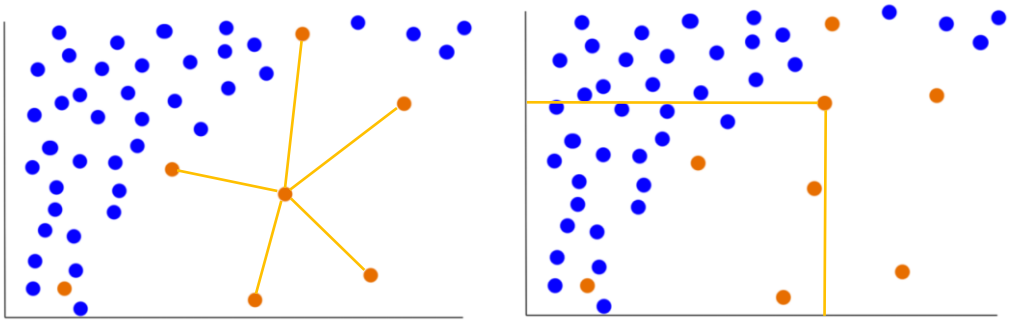
โดยที่

$N_{point}$ คือ จุดข้อมูลของคลาสส่วนนอ้ยทที่สร้างขึ้นมาใหม่

$O_{point}$ คือ จุดข้อมูลของคลาสส่วนน้อยที่นำไปใช้ในการหา ระยะห่างเทียบกับจุดใกล้เคียง

$Random[0,1]$ คือ การสุ่มค่าระหว่าง 0 ถึง 1

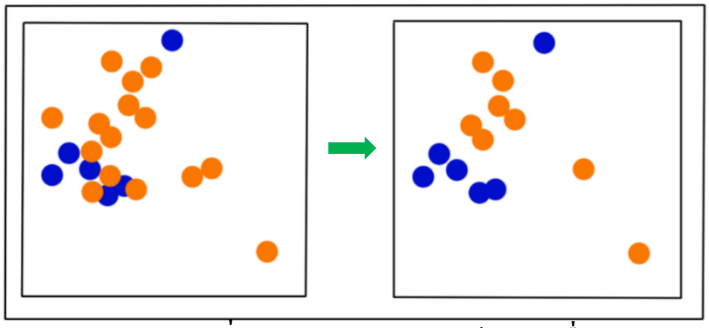
$distance(x, y, \dots, z)$ คือ ระยะห่างระหว่างจุดตั้งต้นกับจุดใกล้เคียงในแอตทริบิวต์ x, y ถึง z



**ภาพประกอบที่ 2.9** การใช้ SMOTE โดย k=5. Sจะทำการเลือกโดย K-NN (ซ้าย) และสร้างข้อมูลแบบสุ่มค่าที่อยู่ระหว่างจุดที่ถูกเลือกอย่างสุ่มที่ภายหลังจากการถูกเลือกจากจาก K-NN (ขวา)

**2.12 Edited Nearest Neighbor Rule (ENN)**

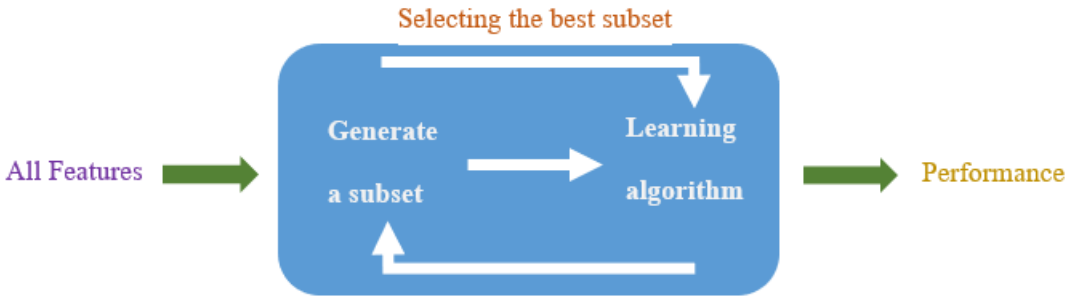
เป็นหนึ่งในวิธีการสุ่มลดเพื่อแก้ปัญหาข้อมูลที่ไม่สมดุลย์โดยจะทำการลบข้อมูลที่อยู่ในคลาสส่วนมากโดยจะจะทำ K-NN กับจุดข้อมูลที่มีคลาสส่วนมาก ถ้าจุดนั้นถูกทำนายว่าเป็นคลาสส่วนน้อยจะถูกลบออก ซึ่งจะทำให้ข้อมูลบริเวณเส้นแบ่งมีความนุ่มนวลมากขึ้น



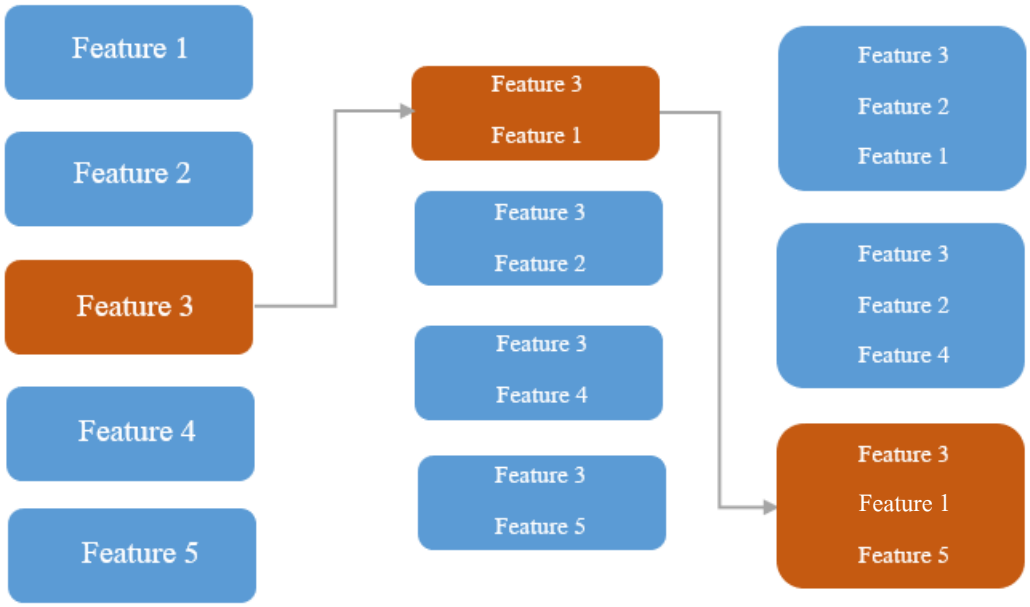
**ภาพประกอบที่ 2.10** ก่อนและหลังใช้ ENN เมื่อ k=5

**2.13 Forward Selection**

เป็นหนึ่งในวิธีการแบบห่อหุ้ม (Wrapper Methods)ซึ่งจะใช้กลุ่มคุณลักษณะย่อย (Feature Subsets) เพื่อตรวจสอบประสิทธิภาพของโมเดลขณะที่ใช้คุณลักษณะต่างๆจากนั้นจะเลือกกลุ่มย่อยที่ทำให้โมเดลมีประสิทธิภาพที่สุดที่สุด (ภาพประกอบที่ 2.11) Forward Selection จะหาค่าในแต่ละคุณลักษณะและเลือกคุณลักษณะที่ทำให้ประสิทธิภาพของโมเดล (AUC, Prediction Accuracy และอื่นๆ) ดีที่สุด จากนั้นจะทำการรวมกับคุณลักษณะอื่นทีละคุณลักษณะแล้วนำไปใช้กับโมเดลเพื่อหาชุดที่ดีที่สุดและทำไปเช่นนี้ไปเรื่อยๆจนกว่าจะครบทุกรูปแบบ (ภาพประกอบที่ 2.12)



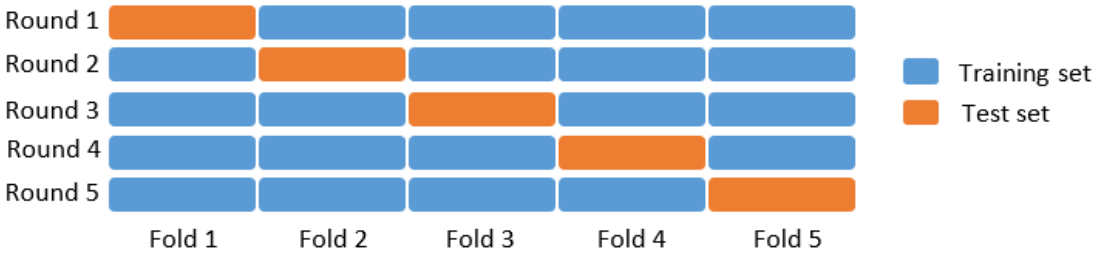

**ภาพประกอบที่ 2.11** วิธีการแบบห่อหุ้ม




**ภาพประกอบที่ 2.12** วิธีการคัดเลือกคุณลักษณะของ Forward Selection

**2.14 K- Fold cross validation**

เป็นวิธีการทางสถิติในการประเมินประสิทธิภาพ ข้อมูลจะถูกแบ่งย่อยๆและได้รับการฝึกฝนในหลายๆโมเดลโดยที่ k คือตัวเลขที่ผู้ใช้ระบุโดยปกติคือ 5 เมื่อทำการตรวจสอบข้อมูลจะถูกแบ่งออกเป็น 5 ส่วนแบบสุ่ม จากนั้นแบบจำลองแรกได้รับการฝึกโดยใช้ชุดข้อมูลย่อยแรกเป็น Test set และใช้ชุดข้อมูลย่อยที่เหลือ (2-5) เป็น Training set แบบจำลองสร้างขึ้นโดยใช้ข้อมูลในชุดข้อมูลย่อยที่ 2-5 จากนั้นจะประเมินความแม่นยำในชุดข้อมูลย่อยที่ 1 จากนั้นจึงสร้างโมเดลอื่นขึ้นมาคราวนี้ใช้ชุดข้อมูลย่อยที่ 2 เป็น Test set และข้อมูลในชุดข้อมูลย่อยที่ 1, 3, 4 และ 5 เป็น Training set กระบวนการนี้จะทำซ้ำโดยใช้ชุดข้อมูลย่อยที่ 3, 4 และ 5 เป็นชุด Test set โดยในแต่ละรอบจะบันทึกค่า Validation Error ไว้ด้วยเพื่อนำไปสรุปผลหลังจบกระบวนการ Cross Validation ทั้งหมด




**ภาพประกอบที่ 2.13** การแยกข้อมูลและใช้ข้อมูลของ Five-fold cross validation

### 2.15 Standardization (z-Score Normalization)

เป็นการปรับสเกลของข้อมูลเพื่อให้ง่ายในการเรียนรู้ของเครื่องโดยจะปรับให้ ค่าเฉลี่ย (Mean) = 0 และ ส่วนเบี่ยงเบนมาตราฐาน (Standard Deviation) = 1 (Unit Variance) โดยจะหาค่าที่ถูก Standardization ได้จากสมการ(2.9)

$$X' = \frac{x-\bar{x}}{\sigma} \tag{2.9}$$

### 2.16 งานวิจัยที่เกี่ยวข้อง

#### 2.16.1 งานวิจัยในประเทศ

นาย ธนวัฒน์ ซื่อสัตย์ และคณะ, (2558), งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีความสัมพันธ์กับผลการรักษาผู้ป่วยกล้ามเนื้อหัวใจขาดเลือดเฉียบพลันผลพบว่าปัจจัยที่มีความสัมพันธ์กับผลการรักษาได้แก่ ตัวแปรเพศ อายุ ระดับความรู้สึกตัว ประวัติการสูบบุหรี่ ประวัติการดื่มเครื่องดื่มแอลกอฮอล์ สัญญาณชีพจร อัตราการหายใจ การเป็นโรคเบาหวาน สิทธิการรักษาพยาบาล และเมื่อพิจารณาปัจจัยที่ส่งผลต่อโอกาสที่จะมีผลการรักษาที่ดีขึ้น พบว่าผู้ป่วยที่อายุน้อย อาการแรกรับมาแบบรู้สึกตัว มีอัตราการหายใจที่ปกติ ไม่มีโรคเบาหวาน และเป็นโรคหัวใจขาดเลือดชนิด Unstable angina จะส่งผลทางบวกต่อโอกาสที่จะมีผลการรักษาที่ดีขึ้น

**2.16.2 งานวิจัยนอกประเทศ**

Helen B. Hubert และคณะ, (2526), งานวิจัยนี้คือการติดตามผลในช่วง 26 ปีของ Framingham Heart Study พบว่าโรคอ้วนเป็นตัวแปรอิสระอย่างมีนัยยะสำคัญโดยเฉพาะเพศหญิงในการบ่งชี้ CVD

P. W. Wilson, W. P. Castelli, W. B. Kannel, (2530), ใช้ปัจจัยเสี่ยง อายุ, เพศ, ระดับคอเลสเตอรอลรวม, HDL, ความดันช่วงบน, การสูบบุหรี่, การแพ้กลูโคสและการขยายตัวของหัวใจ (การเจริญเติบโตมากเกินไปของหัวใจห้องล่างซ้ายในคลื่นไฟฟ้าหัวใจหรือหัวใจที่โตขึ้นเมื่อเอ็กซเรย์ทรวงอก) ในการทำนายความเสี่ยงในการเกิดโรคหลอดเลือดหัวใจโดยมีความเสี่ยงตั้งแต่น้อยกว่า 1 % ถึงมากกว่า 80% ปรากฏขึ้นจากการทำนาย

Richard O. Cannon III, Stephen E. Epstein, (2531), ได้ทำการทดลองเกี่ยวกับจะมีอาการเจ็บหน้าอกเมื่อเป็นโรคหลอดเลือดหัวใจหรือไม่พบว่ามีผู้ป่วยที่เป็นโรคหลอดเลือดหัวใจ แต่มีอาการเจ็บหน้าอกเหมือนโรคหลอดเลือดหัวใจตีบ

Stanley S. Franklin และคณะ, (2544), ใช้ความดันช่วงล่าง, ความดันช่วงบน และความดันชีพจรในการทำนายความเสี่ยงของโรคหลอดเลือดหัวใจ พบว่าเมื่ออายุมากขึ้นมีการค่อยๆ เปลี่ยนตัวทำนายความเสี่ยงในการเกิดโรคหลอดเลือดหัวใจ จากความดันช่วงล่าง เป็นเป็นความดันช่วงบน จากนั้นเป็นความดันชีพจร

Christopher J O'Donnell, Roberto Elosua, (2551), ได้ทำการสรุปการค้นพบปัจจัยเสี่ยงในการเกิดโรคหัวใจและหลอดเลือดจาก Framingham Heart Study ว่ามีดังนี้ ไขมัน (HDL, LDL), ภาวะความดันสูง, การสูบบุหรี่, โรคเบาหวาน, ความอ้วน และการไม่เคลื่อนไหวร่างกาย

Gary F. Mitchell และคณะ, (2553), ใช้แบบจำลองหลายตัวแปรที่ปรับตามอายุ, เพศ, ความดันช่วงบน, การใช้ยาลดความดันโลหิต, ความเข้มข้นของHDL, คลอเลสเตอรอลรวม, การสูบบุหรี่และการมีโรคเบาหวานเป็นพื้นฐาน และถ้ามีการเพิ่ม pulse wave velocity เข้าไปจะทำให้เพิ่มความเสี่ยงความเสี่ยงขึ้น 0.7%

Maruf Ahmed Tamal และคณะ, (2562), ใช้คุณลักษณะจำนวนมากเช่น อาการเจ็บแน่นหน้าอก, อายุ, ปวดแน่นบริเวณแขน, มึนหัว, มีความดันสูง, เจ็บร้อนที่หัวใจ, มีการออกกำลังกาย, ความเครียด, เพศ, มีการหายใจสั้นและถี่, ผู้ปกครองมีการหัวใจวายก่อนอายุ60ปี, สูบบุหรี่, เป็นเบาหวาน และการใช้ยาเสพติด ในการหารูปแบบของโมเดลที่ดีที่สุดในการทำนายการเกิด

โรคหัวใจ พบว่า อาการเจ็บแน่นหน้าอกเป็นคุณลักษณะที่สำคัญที่สุดและSVMเป็นรูปแบบที่แม่นยำที่สุดในชุดข้อมูลนี้

R P Fleet และคณะ, (2537), ได้ตรวจสอบกลุ่มตัวอย่างที่มีอาการเจ็บหน้าอกพบว่ามีผู้ป่วย 30% ที่มีอาการแต่ไม่ได้เป็นโรคหลอดเลือดหัวใจ

Cheryl N. Carmi และคณะ, (2551), ได้ทำการตรวจกลุ่มตัวอย่างที่มีอาการเจ็บหน้าอกแต่ไม่ได้เป็นโรคหัวใจและหลอดเลือด ได้มีการกล่าวถึงผู้ป่วยที่มีอาการเจ็บหน้าอกแต่ไม่ได้เป็นโรคหลอดเลือดหัวใจแต่เป็นโรคเครียด

R. Alejo และคณะ, (2553), ได้ศึกษา Edited Nearest Neighbor Rule ในการเพิ่มประสิทธิภาพของโครงข่ายประสาทเทียมพบว่าการลดลงของชุดฝึกทำให้ภาระการคำนวณต่ำลงแต่ยังไม่สูญเสียประสิทธิภาพและมีแนวโน้มที่จะปรับปรุงความแม่นยำในการจำแนก

Marcelo Beckmann, Nelson F. F. Ebecken, Beatriz S. L. Pires de Lima, (2558), ได้ทำการเสนอแนวทางในการใช้ K-NN ในการสุ่มตัวอย่างเพื่อปรับสมดุลย์ข้อมูลพบว่าวิธีการสุ่มตัวอย่างของ K-NN มีประสิทธิภาพดีกว่าวิธีการสุ่มตัวอย่างอื่น ๆ

Ajinkya More, (2559), ได้ทบทวนเทคนิคการสุ่มตัวอย่างพื่อจัดการกับชุดข้อมูลที่ไม่สมดุลและศึกษาผลกระทบต่อประสิทธิภาพการจำแนกพบว่าวิธีการ SMOTE + ENN ร่วมกับ Logistic Regression และ Balance Cascade ให้ประสิทธิภาพที่ดีที่สุด

# บทที่ 3
# วิธีดำเนินการวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงการทดลอง (Experiment Research) เพื่อศึกษาปัญหาและรวบรวมข้อมูลสำหรับการพัฒนาแบบจำลองการพยากรณ์การเกิดโรคหลอดเลือดหัวใจด้วยปัจจัยที่วัดได้จากภายนอกโดยใช้โครงข่ายประสาทเทียม ผู้วิจัยได้ดำเนินงานวิจัยโดยมีขั้นตอน ดังนี้

3.1 ประชากรและกลุ่มตัวอย่าง

3.2 ขั้นตอนการดำเนินการการวิจัย

3.3 เครื่องมือที่ใช้ในการวิจัย

3.4 วิธีวิเคราะห์ข้อมูล

## 3.1 ประชากรและกลุ่มตัวอย่าง

3.1.1 ประชากร

ประชากรที่ใช้ในการวิจัยครั้งนี้เป็นประชากรชาย-หญิงชาวยุโรป-อเมริกันที่ไม่แสดงอาการของโรคหัวใจ มีอายุระหว่าง 28-74 ที่อาศัยอยู่ในเมืองเฟลมมิงแฮม รัฐแมสซาชูเซตส์ ประเทศสหรัฐอเมริกา ในปี พ.ศ. 2491

3.1.2 กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลของ The Original Cohort จาก Framingham Heart Study โดยเป็นประชากรชาย-หญิงชาวยุโรป-อเมริกันที่ไม่แสดงอาการของโรคหัวใจ มีอายุระหว่าง 28-74 ปี ที่อาศัยอยู่ในเมืองเฟลมมิงแฮม รัฐแมสซาชูเซตส์ ประเทศสหรัฐอเมริกา ในปี พ.ศ. 2491 โดยได้ส่งจดหมายเชิญแบบสุ่มจากสองในทุกๆสามครอบครัวที่มีสมาชิกอายุ 30–59 ปีอาศัยอยู่ในเมืองเฟลมมิงแฮม รัฐแมสซาชูเซตส์ จากผู้ติดต่อ 6507 รายชายและหญิง 4494 (69%) ตกลงที่จะเข้าร่วมและกลุ่มอาสาสมัครเพิ่มเติม ( $n$ = 715) เข้าร่วมด้วย รวมมีกลุ่มตัวอย่าง 5209 คน แบ่งเป็น เพศชาย 2336 คนและเพศหญิง 2873 คน

**3.2 ขั้นตอนการดำเนินการการวิจัย**

ผู้วิจัยได้ใช้ชุดข้อมูลของ The Original Cohort จาก Framingham Heart Study ซึ่งเป็นการวิจัยระยะยาว โดยได้ดำเนินการวิจัยตามขั้นตอนหลักดังนี้

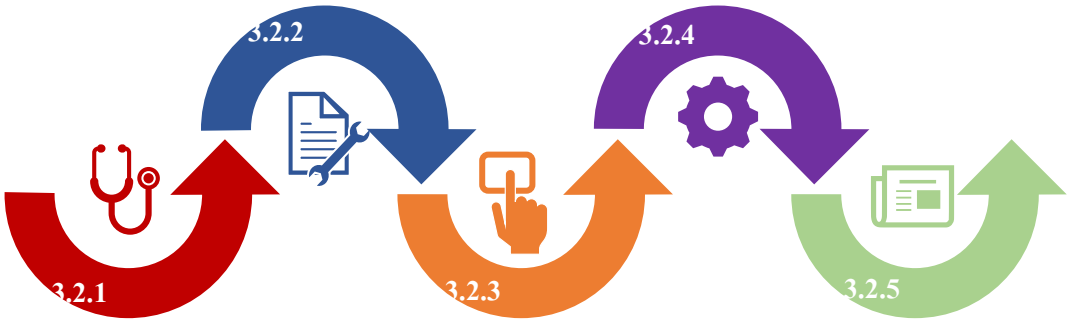
3.2.1 นำเข้าชุดข้อมูล

3.2.2 เตรียมข้อมูลให้เหมาะสมกับการฝึกโครงข่ายประสาทเทียม

3.2.3ใช้เทคนิค Forward Selection ในการคัดเลือกคุณลักษณะด้วยโมเดล 1-20 เซลล์ประสาทใน Hidden Layer

3.2.4 พัฒนาแบบจำลองและทดสอบด้วยเทคนิค K-Fold Cross Validation

3.2.5 สรุปผล




**ภาพประกอบที่ 3.1**. แสดงขั้นตอนงานวิจัยหลัก5ขั้นตอน

3.2.1 นำเข้าชุดข้อมูล

ดึงชุดข้อมูล The Original Cohort จาก Framingham Heart Study จาก วิชาเรียน Bio statistic 503 ของมหาวิทยาลัยวอชิงตัน โดยมีรายละเอียดแสดงดังตารางที่ 3.1 โดยมีคำอธิบาย

**ตารางที่ 3.1** แสดงรายละเอียดข้อมูลที่ใช้ในการวิจัย

ชื่อดังนี้

| Name | Type | Missing | Average | Min | Max |
|---|---|---|---|---|---|
| lexam | Integer | 0 | 14.03532 | 2 | 16 |
| surv | Integer | 0 | 0.382223 | 0 | 1 |
| cause | Integer | 0 | 1.563064 | 0 | 9 |
| cexam | Integer | 0 | 2.561912 | 0 | 16 |
| chd | Integer | 0 | 0.116145 | 0 | 1 |
| cva | Integer | 0 | 0.072567 | 0 | 1 |
| ca | Integer | 0 | 0.103475 | 0 | 1 |
| oth | Integer | 0 | 0.266078 | 0 | 1 |
| sex | Integer | 0 | 1.551545 | 1 | 2 |
| age | Integer | 0 | 44.06873 | 28 | 62 |
| ht | Real | 6 | 64.81318 | 51.5 | 76.5 |
| wt | Integer | 6 | 153.0867 | 67 | 300 |
| scl1 | Polynominal | 2037 | - | 96 | 200 |
| scl2 | Integer | 626 | 228.1778 | 115 | 568 |
| Dbp | Integer | 0 | 85.35861 | 50 | 160 |
| Sbp | Integer | 0 | 136.9096 | 82 | 300 |
| mrw | Integer | 6 | 119.9575 | 67 | 268 |
| smok | Integer | 36 | 9.366518 | 0 | 60 |

1. lexam คือ ครั้งสุดท้ายที่ทดสอบในแต่ละตัวอย่าง

2. surv คือ ตัวอย่างยังมีชีวิตอยู่ในขณะทดสอบครั้งที่ 16 หรือไม่

3. cause คือ สาเหตุการเสียชีวิต แบ่งได้ดังนี้

- 0 = ยังมีชีวิตอยู่

- 1 = เสียชีวิตจากโรคหลอดเลือดหัวใจแบบฉับพลัน
- 2 = เสียชีวิตจากสาเหตุโรคหลอดเลือดหัวใจ อื่น ๆ
- 3 = เสียชีวิตจากโรคหลอดเลือดสมองแบบฉับพลัน (CVA)
- 4 = เสียชีวิตจากโรคหลอดเลือดสมองอื่น ๆ
- 5 = เสียชีวิตจากมะเร็ง
- 6 = เสียชีวิตจากสาเหตุอื่น ๆ
- 9 = ไม่สามารถหาสาเหตุการเสียชีวิตได้

4. cexam คือ ครั้งที่ตรวจพบโรคหลอดเลือดหัวใจ,
5. chd คือ เสียชีวิตจากสาเหตุที่ 1-2
6. cva คือ เสียชีวิตจากสาเหตุที่ 3-4
7. ca คือ เสียชีวิตจากสาเหตุที่ 5
8. oth คือ เสียชีวิตจากสาเหตุที่ 3-9
9. sex คือ เพศ แบ่งได้ดังนี้
    - 1 = Male
    - 2 = Female
10. age คือ อายุ
11. ht คือ ความสูง (นิ้ว)
12. wt คือ น้ำหนัก (ปอนด์)
13. scl1 คือ คอเรสเตอรอลในการเก็บตัวอย่างครั้งที่ 1 (มิลลิกรัม/100 มิลลิลิตร)
14. scl2 คือ คอเรสเตอรอลในการเก็บตัวอย่างครั้งที่ 2 (มิลลิกรัม/100 มิลลิลิตร)
15. Dbp คือ ความดันช่วงล่าง
16. Sbp คือ ความดันช่วงบน
17. mrw คือ น้ำหนักที่ควรเป็นเมื่อวัดจากความสูง
18. smok คือ จำนวนบุหรี่ที่สูบต่อ 1 วัน

3.2.2 เตรียมข้อมูลให้เหมาะสมกับการฝึกโครงข่ายประสาทเทียม

3.2.2.1 Data Transformation

ทำการแปลงข้อมูลความสูงและน้ำหนักในมาตราอิมพีเรียลเป็นมาตราเมตริกและใส่ไว้ในคอลัมส์ Ht(m) และ Wt(kg) ตามลำดับ จากนั้นแปลงค่าของ smoke ให้เป็น Binominal ในคอลัมส์ Smoking และใช้ข้อมูลทั้งสองทำการคำนวนเพื่อสร้างคอลัมส์ BMI จากนั้นทำการเพิ่มคอลัมส์ผลลัพธ์ที่จะถูกทำเครื่องหมายโดยใช้ชื่อว่า 10-years โดยใช้ตารางการตรวจโรคของ Framingham Heart Study (ภาพประกอบที่ 3.2) เทียบกับคอลัมส์ cexam

| Exam Dates and Age Ranges as of 2019 - rev 9.11.2019 | | | | |
|---|---|---|---|---|
| Original Cohort (idtype = 0) | | | | |
| **Exam** | **Exam Date Range** | **Age Range** | **Mean Age** | **Attendees** |
| Exam 1 | 1948 - 1953 | 28 - 74 | 44 | 5209 |
| Exam 2 | 1950 - 1955 | 31 - 65 | 46 | 4792 |
| Exam 3 | 1952 - 1956 | 32 - 67 | 48 | 4416 |
| Exam 4 | 1954 - 1958 | 34 - 69 | 50 | 4541 |

**ภาพประกอบที่ 3.2** ตารางการตรวจโรคของFramingham Heart Study

3.2.2.2 Data Cleaning

จากตารางที่ 1 จะพบว่าคอลัมส์ scl1 และ scl2 มีข้อมูลหายจำนวนมากและเป็นข้อมูลที่ไม่ได้ใช้จึงทำการลบทั้ง 2 คอลัมส์ออกก่อน จากนั้นจึงลบแถวที่มีข้อมูลไม่ครบออกทั้งหมด และเนื่องจากงานวิจัยฉบับนี้สนใจเพียงปัจจัยที่วัดได้จากภายนอกเท่านั้นจึงจะทำการลบข้อมูลอื่นเพื่อลดภาระในการคำนวน

3.2.2.3 Balancing

ใช้เทคนิค SMOTE+ENN ด้วยค่า k = 5 ในการแก้ไขข้อมูลไม่สมดุลโดยแสดงข้อมูลต้นฉบับและข้อมูลที่ผ่านกระบวนการแล้วได้ดังภาพประกอบที่ 3.3-3.5

**ภาพประกอบที่ 3.3** ข้อมูลต้นฉบับ

**ภาพประกอบที่ 3.4** ข้อมูลภายหลังจากการใช้ SMOTE + ENN

| | | | | | |
|---|---|---|---|---|---|
| 10-years | Binominal | 0 | Negative: false | Positive: true | Values: false (4893), true (270) |
| sex | Integer | 0 | Min: 1 | Max: 2 | Average: 1.552 |
| age | Integer | 0 | Min: 28 | Max: 62 | Average: 44.063 |
| dbp | Integer | 0 | Min: 50 | Max: 160 | Average: 85.364 |
| sbp | Integer | 0 | Min: 82 | Max: 300 | Average: 136.917 |
| Ht(m) | Real | 0 | Min: 1.308 | Max: 1.943 | Average: 1.646 |
| Wt(kg) | Real | 0 | Min: 30.391 | Max: 136.078 | Average: 69.439 |
| Smoking | Integer | 0 | Min: 0 | Max: 1 | Average: 0.516 |
| BMI | Real | 0 | Min: 14.124 | Max: 56.684 | Average: 25.585 |

**ภาพประกอบที่ 3.5** รายละเอียดข้อมูลภายหลังจากการใช้ SMOTE + ENN

3.2.3 ใช้เทคนิค Forward Selection ในการคัดเลือกคุณลักษณะด้วยโมเดล 1-30 เซลล์ประสาทใน Hidden Layer

ทำการแยกข้อมูลเป็นTraining set และ Test set ด้วยสัดส่วน 7:3 เราจะใช้ Training set ในการฝึกโครงข่ายประสาทเทียมและใช้ Test set ในการทดสอบความแม่นยำ เริ่มโดยทำ Standardization กับข้อมูล Training set โดยจะเก็บค่าเฉลี่ย และส่วนเบี่ยงเบนมาตราฐานไว้และใช้ทำ Standardization ข้อมูล Test set เช่นกัน จากนั้นจะทำการฝึกโครงข่ายประสาท

เทียม 1000 รอบ ใช้updater เป็น Adam, optimization method เป็น Stochastic Gradient Descent และใช้ activation function เป็น ReLU ในการวัดประสิทธิภาพนั้นจะทำการเปลี่ยนจำนวนเซลล์ประสาทใน Hidden Layer ตั้งแต่ 1-30 เพื่อหา จำนวนเซลล์ประสาทที่ให้ผลดีที่สุด โดยวัดจากค่าความถูกต้อง (Accuracy) แสดงได้ดังสมการที่ (3.1)

accuracy: 84.32%

| | true no |
|---|---|
| pred. no | 822 |
| pred. yes | 184 |

**ภาพประกอบที่ 3.6** ตัวอย่างความถูกต้อง

3.2.4 พัฒนาแบบจำลองและทดสอบด้วยเทคนิค K-Fold cross validation

เราจะสร้างโครงข่ายประสาทเทียมดังขั้นตอนที่ 3.2.3 ครั้งแต่ในครั้งนี้เราจะใช้เทคนิค K-Fold Cross Validation มาสร้างเพื่อกำจัดความลำเอียงที่อาจเกิดขึ้นจากการแบ่งข้อมูล โดยจะกำหนด k = 10 และทำการ Standardization ไปพรร้อมกันด้วย ในขั้นตอนนี้จะทำการการวัดค่า ความถูกต้อง, ค่า AUC ดังสมการที่ (3.2), Recall ดังสมการที่ (3.3) และ Specificity ดังสมการที่ (3.4)

3.2.5 สรุปผลการทดลอง

ทำการสรุปผลการทดลองจากการเก็บค่าคุณลักษณะที่ใช้ ความแม่นยำในการพยากรณ์ และจำนวนเซลล์ประสาทที่ใช้

## 3.3 เครื่องมือที่ใช้ในการวิจัย

3.3.1 The Original Cohort จาก Framingham Heart Study

3.3.2 ระบบพัฒนา

3.3.2.1 ฮาร์ดแวร์ที่ใช้ในการวิจัยประกอบด้วย

3.3.2.1.1 คอมพิวเตอร์ ซีพียู Intel® Core™ i7-7700HQ จำนวน 1 เครื่อง

3.3.2.1.2 ฮาร์ดดิสก์ความจุ 1 TB และ หน่วยความจำ 12 GB

3.3.2.2 ซอฟต์แวร์ที่ใช้ในการวิจัยประกอบด้วย

3.3.2.2.1 ระบบปฏิบัติการ Microsoft Windows 10

3.3.2.2.2 โปรแกรม RapidMiner Studio

3.3.2.2.2 โปรแกรม Microsoft Excel

**3.4 วิธีวิเคราะห์ข้อมูล**

ในการวิจัยในครั้งนี้ ผู้วิจัยได้ดำเนินการวิเคราะห์ข้อมูลต่าง ๆ ดังนี้

3.4.1 Accuracy เป็นการวัดความแม่นยำโดยรวม โดยมีสมการ

$$Accuracy = \frac{Correct\ Predictions}{Number\ of\ Examples} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.1)$$

โดยที่

TP = จำนวนของ True Positives FP = จำนวนของ False Positives

TN = จำนวนของ True Negatives FN = จำนวนของ False Negatives

3.4.2 AUC Score เกิดจากการหาพื้นที่ใต้กราฟที่มีการพลอตระหว่าง TPR (True Positive Rate) และ FPR (False Positive Rate) ค่ายิ่งเข้าใกล้ 1 มากยิ่งดี เพราะโมเดลจะมีประสิทธิภาพโดยมีสมการ

$$AUC = \int TPR\ dFPR \quad (3.2)$$

3.4.3 Recall จะใช้ในการป้องกันผลลบปลอมในการทดสอบนั้นยิ่งการทดสอบไวมากเท่าไรโอกาสการได้ผลลบจะน้อยลงเท่านั้นและดังนั้น ถ้าความไวอยู่ที่ 100% โอกาสได้ผลลบปลอมจะอยู่ที่ 0% โดยมีสมการ

$$Recall = \frac{TP}{TP+FN} \quad (3.3)$$

โดยที่

TP = จำนวนของ True Positives FN = จำนวนของ False Negatives

3.4.4 Specificity จะใช้ในการป้องกันผลบวกปลอมในการทดสอบนั้นยิ่งการทดสอบยิ่งจำเพาะมากเท่าไรโอกาสการได้ผลบวกก็น้อยลงเท่านั้น และดังนั้น ถ้าความจำเพาะอยู่ที่ 100% โอกาสได้ผลบวกปลอมจะอยู่ที่ 0% โดยมีสมการ

$$Specificity = \frac{TN}{TN+FP} \tag{3.4}$$

โดยที่

TN = จำนวนของ True Negatives FP = จำนวนของ False Positives

# บทที่ 4

## ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาคุณปัจจัยภายนอกที่มีความแน่นอนซึ่งสามารถใช้ในการพยากรณ์การเกิดโรคหลอดเลือดหัวใจได้และเพื่อสร้างแบบจำลองในการพยากรณ์การเกิดโรคหลอดเลือดหัวใจ ในงานวิจัยนี้เริ่มด้วยการทำความสะอาดข้อมูลจาก Framingham Heart Study เพื่อกำจัดข้อมูลที่ไม่สมบูรณ์จากนั้นได้ทำการแก้ไขข้อมูลไม่สมดุลย์ด้วยเทคนิค SMOTE+ENN แล้วจึงทำการการคัดเลือกคุณลักษณะที่เป็นปัจจัยภายนอกที่สามารถวัดได้ที่มีผลต่อการพยากรณ์ด้วยเทคนิค Forward Selection จากคุณลักษณะดังนี้ เพศ, อายุ, ส่วนสูง(เมตร), น้ำหนัก(กิโลกรัม), ความดันช่วงบน (Sbp), ความดันช่วงล่าง (Dbp), การสูบบุหรี่, ค่า BMI เพื่อนำมาใช้ในการพยากรณ์และทำการสร้างแบบจำลองและทดสอบแบบจำลองด้วยเทคนิค K-Fold cross validation โดยวัดค่าความแม่นยำ ด้วยค่า ความถูกต้อง (Accuracy), ค่า AUC, Recall และ Specificity โดยมีผลการคัดเลือกคุณลักษณะรวมถึงจำนวนเซลล์ประสาท และผลความแม่นยำในการทดสอบแบบจำลองด้วย K-Fold cross validation ปรากฏดังนี้

**4.1 ผลการคัดเลือกคุณลักษณะและจำนวนเซลล์ประสาทใน Hidden Layer ด้วยวิธี Forward Selection**

**ตารางที่ 4.1** ความถูกต้อง และคุณลักษณะเมื่อใช้จำนวนเซลล์ประสาทต่างกัน

| numbers of neuron | accuracy | feature |
|---|---|---|
| 1 | 82.28 | Sex, Age, Dbp, Smoking |
| 2 | 82.45 | Age, Dbp |
| 3 | 82.49 | Age, Hight, Dbp, BMI |
| 4 | 84.20 | Sex, Age, Sbp, Smoking, BMI |
| 5 | 83.30 | Sex, Age, Weight, Sbp, Dbp |
| 6 | 80.74 | Age, BMI |
| 7 | 87.32 | Age, Hight, Weight, Dbp, Smoking, BMI |
| 8 | 84.59 | Sex, Age, Weight, Sbp, Dbp |

**ตารางที่ 4.1** (ต่อ)

| | | |
|---|---|---|
| 9 | 84.93 | Age, Weight, Dbp, Smoking |
| 10 | 89.33 | Age, Hight, Weight, Dbp, BMI |
| 11 | 88.34 | Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI |
| **12** | **90.61** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| 13 | 86.34 | Sex, Age, Dbp, Smoking, BMI |
| 14 | 89.84 | Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI |
| **15** | **90.95** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| 16 | 86.59 | Age,Sbp, Dbp, Smoking, BMI |
| 17 | 87.11 | Sex, Age,Sbp,Dbp,BMI |
| **18** | **90.65** | **Age, Hight, Weight, Sbp, Dbp, BMI** |
| **19** | **90.95** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **20** | **90.18** | **Age, Hight, Weight, Sbp, BMI** |
| **21** | **90.78** | **Sex, Age, Hight, Weight, Sbp, Dbp, BMI** |
| **22** | **91.33** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **23** | **91.42** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **24** | **91.2** | **Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **25** | **91.46** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **26** | **91.37** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **27** | **92.95** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| **28** | **92.14** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |
| 29 | 88.98 | Age, Weight, Sbp, Dbp, Smoking |
| **30** | **92.14** | **Sex, Age, Hight, Weight, Sbp, Dbp, Smoking, BMI** |

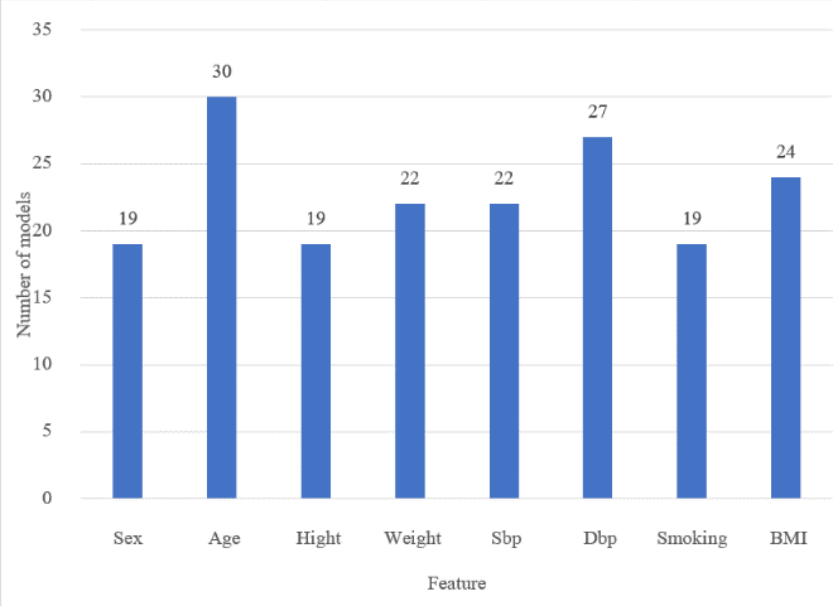
*Dbp = Diastolic blood pressure

*Sbp = Systolic blood pressure

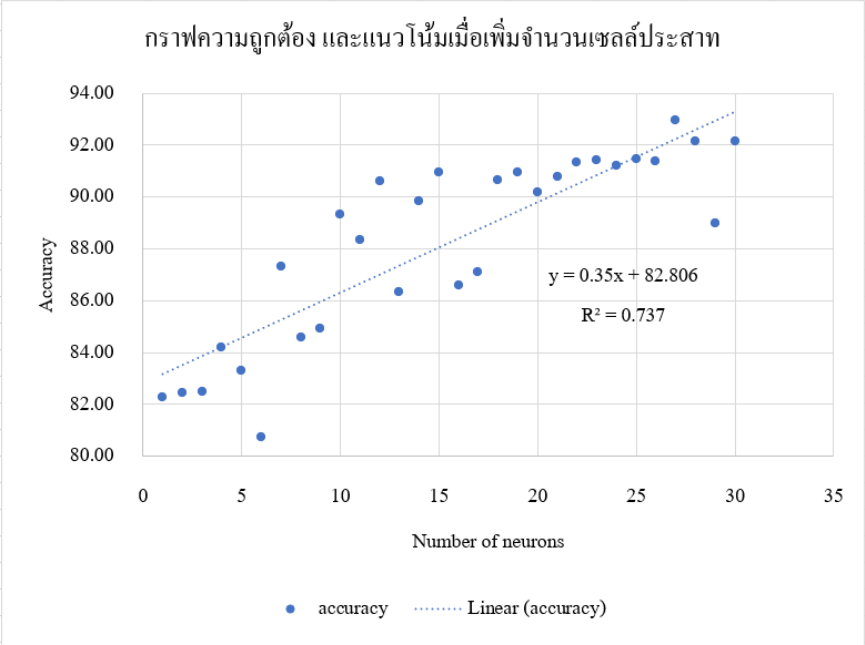
4.1.1 อภิปรายผลการคัดเลือกคุณลักษณะ

จากผลการคัดเลือกคุณลักษณะ (ตารางที่ 4.1) จากชุดข้อมูล The Original Cohort ที่ผ่านการแก้ไขข้อมูลไม่สมดุลย์แล้ว พบว่าจำนวนเซลล์ประสาทที่ทำให้มีค่าความถูกต้องสูงมากกว่า 90% จำนวน 14 แบบจำลอง คือ 12, 15, 18 – 28 และ 30 เซลล์ประสาท

จากการสังเกตจะพบว่าในทุกแบบจำลองจะใช้ อายุ เป็นปัจจัยสำคัญในการพยากรณ์ซึ่งเป็นไปตามแบบพยากรณ์ความเสี่ยงของ Framingham Heart Study (ภาพประกอบที่ 2.3) ที่ใช้อายุในการพยากรณ์ และมีปัจจัยที่ถูกใช้มากรองลงมาแต่ยังมากกว่าค่าเฉลี่ยคือความดันช่วงล่าง และค่า BMI ตามลำดับ โดยมีแบบจำลองใช้ความดันช่วงล่างในการพยากรณ์เป็นจำนวนมากถึง 27 จาก 30 แบบจำลอง และใช้ค่าBMI เป็นจำนวนมากถึง 24 จาก 30 (ภาพประกอบที่ 4.1) ในขณะที่มีค่าเฉลี่ย 22.75 และยังพบว่าเมื่อเพิ่มจำนวนของเซลล์ประสาทในแบบจำลองจะทำให้ความแม่นยำมีแนวโน้มเพิ่มขึ้น (ภาพประกอบที่ 4.2)




**ภาพประกอบที่ 4.1** จำนวนแบบจำลองต่อคุณลักษณะ




**ภาพประกอบที่ 4.2** กราฟความถูกต้อง และแนวโน้มเมื่อเพิ่มจำนวนเซลล์ประสาท

## 4.2 ผลลัพธ์ประสิทธิภาพเมื่อทำการลดความลำเอียงด้วยเทคนิค K- Fold Cross validation

**ตารางที่ 4.2** ความถูกต้อง, AUC, Recall และ Specificity หลังจากสร้างและทดสอบด้วย K-Fold cross validation

| number of neurons | Accuracy | AUC | Recall/Sensitivity | Specificity |
|---|---|---|---|---|
| 12 | 88.23 | 0.951 | 91.16 | 84.63 |
| 15 | 89.93 | 0.958 | 92.35 | 85.74 |
| 18 | 89.79 | 0.963 | 90.83 | 88.5 |
| 19 | 90.04 | 0.961 | 92.18 | 87.42 |
| 20 | 89.84 | 0.961 | 90.72 | 88.76 |
| 21 | 91.13 | 0.97 | 92.39 | 89.59 |
| 22 | 91.45 | 0.97 | 93.44 | 89.02 |
| **23** | **91.92** | **0.973** | **93.39** | **90.1** |
| 24 | 91.35 | 0.97 | 92.95 | 89.39 |
| 25 | 90.84 | 0.966 | 93 | 88.19 |

**ตารางที่ 4.2** (ต่อ)

| | | | | |
|---|---|---|---|---|
| 26 | 91.25 | 0.97 | 93.25 | 88.79 |
| 27 | 91.88 | 0.971 | 93.98 | 89.3 |
| **28** | **92.04** | **0.974** | **93.44** | **90.33** |
| 30 | 91.81 | 0.973 | 93.65 | 89.56 |

4.2.1 อภิปรายผลการสร้างและทดสอบแบบจำลองด้วย K-Fold cross validation

จากผลการสร้างและทดสอบแบบจำลอง (ตารางที่ 4.2) จากผลการคัดเลือกคุณลักษณะที่มีความถูกต้องมากกว่า 90% จำนวน 14 แบบจำลอง พบว่ามีแบบจำลอง 2 แบบจำลอง คือ 23 และ 28 เซลล์ประสาท ที่มีค่า AUC มากกว่า 0.9 ค่า Recall มากกว่า 90% และ ค่า Specificity มากกว่า 90% โดยที่ทั้ง 2 แบบจำลองมีค่าใกล้เคียงกัน โดยมีแบบจำลองจำนวน 28 เซลล์ประสาทเป็นแบบจำลองที่ดีที่สุด โดยมีค่าความถูกต้อง 92.04% ค่า AUC 0.974 ค่า Recall 93.44% และค่า Specificity 90.33% จะแสดง Confusion Metrix ทั้ง 14 แบบจำลองที่ถูกคัดเลือก ดังภาพด้านล่างนี้

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 2967 | 380 | 88.65% |
| pred. true | 539 | 3919 | 87.91% |
| class recall | 84.63% | 91.16% | |

**ภาพประกอบที่ 4.3** Confusion Matrix ของแบบจำลอง 12 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3006 | 329 | 90.13% |
| pred. true | 500 | 3970 | 88.81% |
| class recall | 85.74% | 92.35% | |

**ภาพประกอบที่ 4.4** Confusion Matrix ของแบบจำลอง 15 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3103 | 394 | 88.73% |
| pred. true | 403 | 3905 | 90.65% |
| class recall | 88.51% | 90.84% | |

**ภาพประกอบที่ 4.5** Confusion Matrix ของแบบจำลอง 18 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3065 | 336 | 90.12% |
| pred. true | 441 | 3963 | 89.99% |
| class recall | 87.42% | 92.18% | |

**ภาพประกอบที่ 4.6** Confusion Matrix ของแบบจำลอง 19 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3065 | 336 | 90.12% |
| pred. true | 441 | 3963 | 89.99% |
| class recall | 87.42% | 92.18% | |

**ภาพประกอบที่ 4.7** Confusion Matrix ของแบบจำลอง 20 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3141 | 327 | 90.57% |
| pred. true | 365 | 3972 | 91.58% |
| class recall | 89.59% | 92.39% | |

**ภาพประกอบที่ 4.8** Confusion Matrix ของแบบจำลอง 21 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3121 | 282 | 91.71% |
| pred. true | 385 | 4017 | 91.25% |
| class recall | 89.02% | 93.44% | |

**ภาพประกอบที่ 4.9** Confusion Matrix ของแบบจำลอง 22 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3159 | 284 | 91.75% |
| pred. true | 347 | 4015 | 92.04% |
| class recall | 90.10% | 93.39% | |

**ภาพประกอบที่ 4.10** Confusion Matrix ของแบบจำลอง 23 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3134 | 303 | 91.18% |
| pred. true | 372 | 3996 | 91.48% |
| class recall | 89.39% | 92.95% | |

**ภาพประกอบที่ 4.11** Confusion Matrix ของแบบจำลอง 24 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3092 | 301 | 91.13% |
| pred. true | 414 | 3998 | 90.62% |
| class recall | 88.19% | 93.00% | |

**ภาพประกอบที่ 4.12** Confusion Matrix ของแบบจำลอง 25 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3113 | 290 | 91.48% |
| pred. true | 393 | 4009 | 91.07% |
| class recall | 88.79% | 93.25% | |

**ภาพประกอบที่ 4.13** Confusion Matrix ของแบบจำลอง 26 เซลล์ประสาท

| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3131 | 259 | 92.36% |
| pred. true | 375 | 4040 | 91.51% |
| class recall | 89.30% | 93.98% | |

**ภาพประกอบที่ 4.14** Confusion Matrix ของแบบจำลอง 27 เซลล์ประสาท

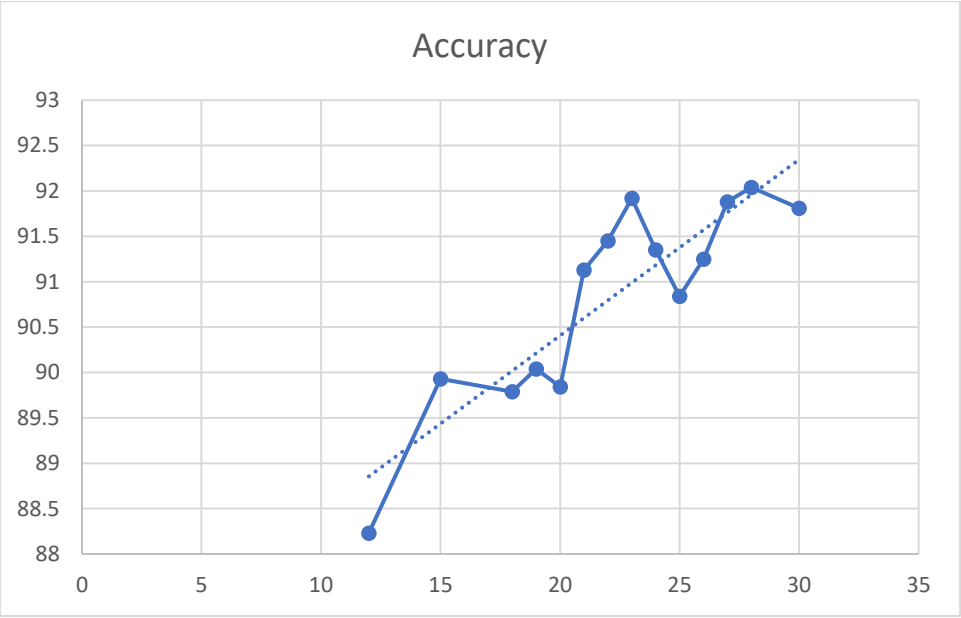
| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3167 | 282 | 91.82% |
| pred. true | 339 | 4017 | 92.22% |
| class recall | 90.33% | 93.44% | |

**ภาพประกอบที่ 4.15** Confusion Matrix ของแบบจำลอง 28 เซลล์ประสาท

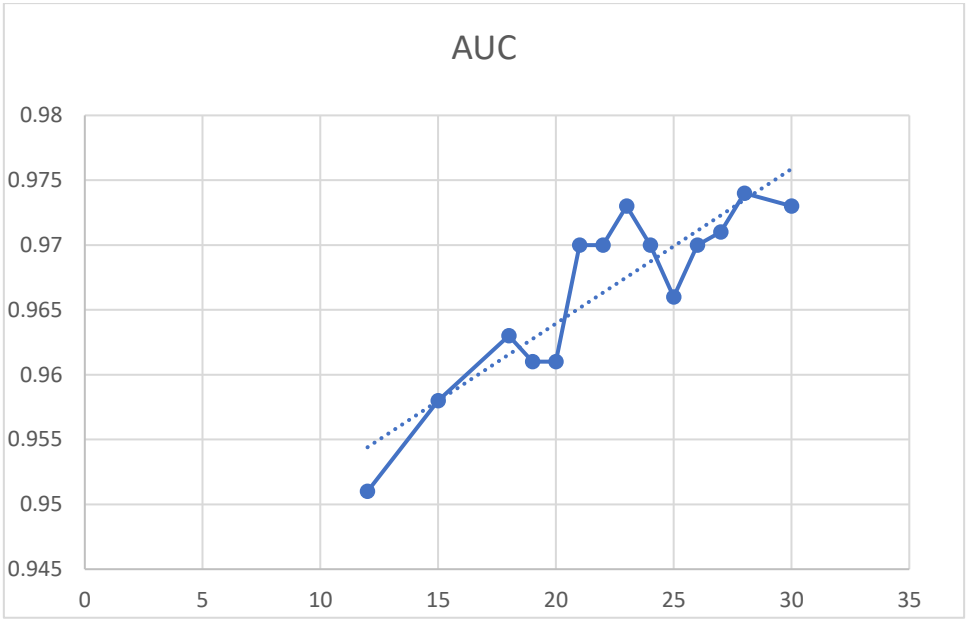
| | true false | true true | class precision |
|---|---|---|---|
| pred. false | 3140 | 273 | 92.00% |
| pred. true | 366 | 4026 | 91.67% |
| class recall | 89.56% | 93.65% | |

**ภาพประกอบที่ 4.16** Confusion Matrix ของแบบจำลอง 30 เซลล์ประสาท

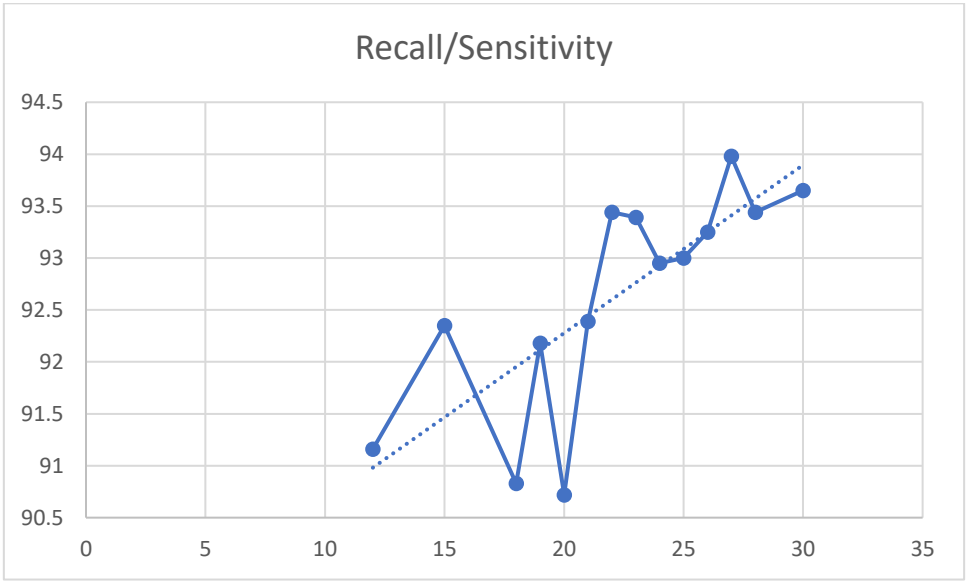
โดยจากการสังเกตค่าต่างๆจะพบว่ามีแนวโน้มเพิ่มขึ้นเมื่อมีจำนวนเซลล์ประสาทเพิ่มมากขึ้น แต่ค่าของ Recall นั้นถึงแม้จะมีแนวโน้มเพิ่มขึ้นแต่ยังมีจุดที่ค่าตกลงจนทำจุดต่ำสุดใหม่ที่ต่ำกว่าจุดต่ำสุดเดิม (ภาพประกอบที่ 4.19) ซึ่งค่าอื่นไม่ปรากฏเหตุการณ์นี้เกิดขึ้น



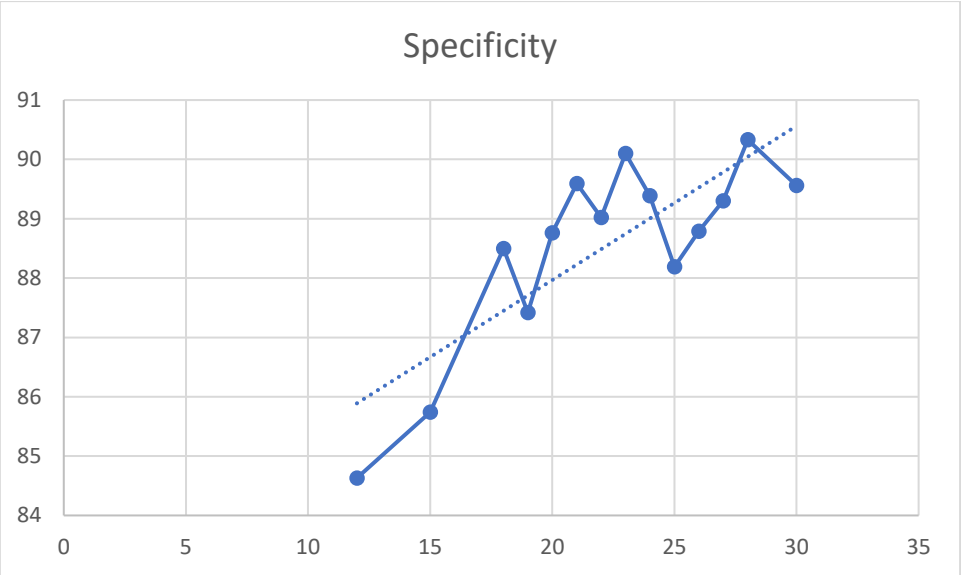

**ภาพประกอบที่ 4.17** แนวโน้มของความถูกต้อง




**ภาพประกอบที่ 4.18** แนวโน้มของค่า AUC




**ภาพประกอบที่ 4.19** แนวโน้มของค่า Recall




**ภาพประกอบที่ 4.20** แนวโน้มของค่า Specificity

งานวิจัยนี้เป็นการสร้างแบบจำลองสำหรับการพยากรณ์การเกิดโรคโดยใช้ปัจจัยที่สามารถวัดได้จากภายนอก โดยมีวัตถุประสงค์เพื่อ สร้างแบบจำลองที่สามารถใช้เพียงปัจจัยที่สามารถวัดได้จากภายนอกในการพยากรณ์เพื่อให้นำไปใช้ได้ในวงกว้างโดยลดความซับซ้อนและปัจจัยที่ต้องมีอุปกรณ์เฉพาะในการตรวจโดยมีประสิทธิภาพการพยากรณ์สูงสุดคิด เป็นร้อยละ 92.04% โดยมี AUC ที่ 0.974 ค่า Recall ที่ 93.44% และ ค่า Specificity ที่ 90.33% โดยใช้คุณลักษณะทั้ง 8 คุณลักษณะ คือ เพศ, อายุ, ส่วนสูง(เมตร), น้ำหนัก(กิโลกรัม), ความดันช่วงบน (Sbp), ความดันช่วงล่าง (Dbp), การสูบบุหรี่, ค่า BMI และมีการใช้เซลล์ประสาทจำนวน 28 แบบจำลองใน Hidden layer 1 ชั้นซึ่งสามารถนำผลที่ได้ไปประยุกต์ใช้ในการสร้างระบบสนับสนุนการตัดสินใจเพื่อเพิ่มการตระหนักรู้ในการเกิดโรคหลอดเลือดหัวใจ และการดูแลสุขภาพเพื่อป้องกันไม่ให้เกิดโรคต่อไป

# บทที่ 5

# สรุป อภิปรายผลและข้อเสนอแนะ

## 5.1 สรุป

ผลการทดลองแสดงให้เห็นว่าประเภทของปัจจัยและจำนวนของเซลล์ประสาทมีผลต่อประสิทธิภาพในการพยากรณ์ของแบบจำลอง (ตารางที่ 4.1) อย่างไรก็ตามด้วยความลำเอียงจากการแบ่งข้อมูลทำให้มีการเปลี่ยนแปลงความแม่นยำเล็กน้อยเมื่อทดสอบด้วยเทคนิค K-Fold Cross validation โดยในจำนวนแบบจำลองการพยากรณ์ความเสี่ยงการเกิด โรคหลอดเลือดหัวใจ ที่สร้างขึ้นมีแบบจำลองที่มีประสิทธิภาพในการพยากรณ์โรคในระดับที่มากกว่า 90% จำนวน 10 แบบจำลองแต่เมื่อคิดร่วมกับค่า AUC ที่มากกว่า 0.9 ค่า Recallและ Specificity มากกว่า 90% แล้วจะเหลือเพียง 2 แบบจำลองซึ่งแบบจำลองที่ดีที่สุดในการทดลองนี้คือแบบจำลองที่ใช้ 28 เซลล์ประสาท และยังคงมีแนวโน้มที่เพิ่มขึ้นเมื่อเพิ่มจำนวนเซลล์ประสาท

## 5.2 อภิปรายผล

แบบจำลองทุกแบบจำลองจะใช้อายุเป็นปัจจัยในการพยากรณ์ซึ่งสอดคล้องกับ ATP III Hard CHD Rick Score แต่กลับพบว่ามีการใช้ความดันช่วงล่างในการพยากรณ์เป็นจำนวนมากซึ่งขัดแย้งกับ ATP III Hard CHD Rick Score แต่กลับไปสอดคล้องกับผลลัพธ์ของการทดสอบของ Wilson Peter W. F. (Original FRS) ที่ถูกระบุว่ามีความสามารถในการจำแนกที่น้อยกว่าแทน (Damen, J. A.,2019) ซึ่งอาจเกิดจากการหายไปของปัจจัยหลักที่ใช้พยากรณ์ซึ่งคือคลอเรสเตอรอลรวมและ HDL อย่างไรก็ตามยังไม่สามารถอธิบายได้ว่าเหตุใดแบบจำลองจึงมีการเลือกใช้ปัจจัยที่ไม่ปรับปรุงประสิทธิภาพอย่างมีนัยสำคัญ (Goff David C.,2014) มากกว่าปัจจัยที่มีผล ข้อจำกัดสำคัญของการวิจัยนี้อยู่ที่ข้อมูลที่มีอยู่นั้นค่อนข้างมีความเก่าและมีความเจาะจงในเชื้อชาติซึ่งอาจทำให้ผลการพยากรณ์ไม่สอดคล้องกับปัจจุบัน

## 5.3 ข้อเสนอแนะ

การศึกษาในอนาคตขอเสนอแนะให้มีการเก็บข้อมูลที่ครอบคลุมในหลายเชื้อชาติ และต่างภูมิภาค เพื่อเปรียบเทียบประสิทธิภาพของการพยากรณ์ ศึกษาการใช้ปัจจัยอื่นๆที่มีผลทดแทนเมื่อไม่มีปัจจัยหลักในการพยากรณ์ ถึงกระนั้นการพยากรณ์ที่มีความแม่นยำสูงและใช้ง่ายนั้นทำได้ยาก การศึกษานี้ก็ได้แสดงการพยากรณ์ที่มีความแม่นยำระดับหนึ่งโดยใช้ปัจจัยที่วัดได้จากภายนอกแล้วจึงสามารถเป็นเหตุให้ยอมรับได้เพื่อนำไปพัฒนาการพยากรณ์เพื่อพยากรณ์ความเสี่ยงในการ

เป็น โรคหลอดเลือดหัวใจ ในอนาคตช่วง 10 ปีของผู้ใช้เพื่อช่วยในการเพิ่มความตระหนักรู้ความเสี่ยงในการเกิดโรคแก่ผู้ใช้และสามารถใช้ได้ง่ายเนื่องจากใช้เพียงตัวแปรจากปัจจัยที่วัดได้จากภายนอกจึงไม่มีความกำกวม

มีประชากรที่เป็นโรคหลอดเลือดหัวใจเพิ่มขึ้นในทุก ๆ ปีแต่ยังมีประชากรจำนวนมากที่ไม่สามkรถเข้าถึงการตรวจโรค เพื่อเพิ่มการตระหนักรู้การพยากรณ์ที่แม่นยำและใช้งานได้ง่ายจะมีส่วนช่วยได้อย่างมาก งานวิจัยนี้มีสามารถใช้เพียงข้อมูลปัจจัยที่สามารถวัดได้จากภายนอกในการสร้างแบบจำลองการเป็นโรคที่มีประสิทธิภาพในการพยากรณ์ โรคหลอดเลือดหัวใจ ในอนาคตหวังว่างานวิจัยนี้จะนำไปสู่การช่วยพยากรณ์ที่แม่นยำและใช้งานได้ง่ายมากยิ่งขึ้น หากไม่สามารถสร้างความตระหนักรู้ในการเป็นโรคหลอดเลือดหัวใจแล้วการเพิ่มขึ้นของผู้ป่วยโรคนี้จะนำไปสู่การเป็นโรคอื่น ๆ ที่รุนแรงกว่าได้จึงหวังว่าในอนาคตการวิจัยการพยากรณ์โดยใช้ปัจจัยอย่างง่ายนี้จะเป็นส่วนช่วยในการลดปริมาณการเกิดโรคหลอดเลือดหัวใจ

# บรรณานุกรม


กองระบาดวิทยา กรมควบคุมโลก กระทรวงสาธารณะสุข. (2562, Dec 27). สถานการณ์โรคหลอดเลือดหัวใจ Coronary Artery Disease (CAD) ปี พ.ศ. 2561. [Online]. Retrieved from: https://ddc.moph.go.th/uploads/files/1081120191227091554.pdf.

ซื่อสัตย์, ธ., น้อมกุศล, จ., อินทร์พิมพ์, พ., พลเยี่ยม, พ., & สุระภี, ส. (2020). ปัจจัยที่มีความสัมพันธ์ต่อผลการรักษาผู้ป่วยกล้ามเนื้อหัวใจขาดเลือดเฉียบพลัน. วารสารวิจัยสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น | KKU Journal for Public Health Research, 102–112.

โรงพยาบาลศิริราช ปิยมหาราชการุณย์. (2563, Dec 8). หลอดเลือดหัวใจตีบ ภัยใกล้ตัว (Coronary Heart Disease). [Online]. Retrieved from: https://www.siphhospital.com/th/news/article/share/850/Coronaryheartdisease.

Ahmed, M., Saiful, M., Ahmmed, M., Aziz, M., Miah, P., & Rezaul, K. (2019). Heart Disease Prediction based on External Factors: A Machine Learning Approach. International Journal of Advanced Computer Science and Applications, 10.

Alejo, R., Sotoca, J. M., Valdovinos, R. M., & Toribio, P. (2010). Edited Nearest Neighbor Rule for Improving Neural Networks Classifications. In L. Zhang, B.-L. Lu, & J. Kwok (Eds.), Advances in Neural Networks—ISNN 2010 (pp. 303–310). Springer.

Bach, M., Werner, A., & Palt, M. (2019). The Proposal of Undersampling Method for Learning from Imbalanced Datasets. Procedia Computer Science, 159, 125–134.

Beckmann, M., Ebecken, N. F. F., & Lima, B. S. L. P. de. (2015). A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 7(4), 104–116.

CDC. (2019, December 9). Coronary Artery Disease | cdc.gov. Centers for Disease Control and Prevention. [Online]. Retrieved from: https://www.cdc.gov/heartdisease/coronary_ad.htm

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chia, Y. C., Gray, S. Y. W., Ching, S. M., Lim, H. M., & Chinna, K. (2015). Validation of the Framingham general cardiovascular risk score in a multiethnic Asian population: A retrospective cohort study. BMJ Open, 5(5), e007324.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P., & CHD Risk Prediction Group. (2001). Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. JAMA, 286(2), 180–187.

D'Agostino, Ralph B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation, 117(6), 743–753.

Damen, J. A., Pajouheshnia, R., Heus, P., Moons, K. G. M., Reitsma, J. B., Scholten, R. J. P. M., Hooft, L., & Debray, T. P. A. (2019). Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: A systematic review and meta-analysis. BMC Medicine, 17(1), 109.

Dua, S. & Chowriappa, P. (2012). Data mining for bioinformatics. Boca Raton: CRC Press.

Fleet, R. P., Dupuis, G., Marchand, A., Burelle, D., & Beitman, B. D. (1994). Panic disorder, chest pain and coronary artery disease: Literature review. The Canadian Journal of Cardiology, 10(8), 827–834.

Franklin, S. S., Larson, M. G., Khan, S. A., Wong, N. D., Leip, E. P., Kannel, W. B., & Levy, D. (2001). Does the relation of blood pressure to coronary heart disease risk change with aging? The Framingham Heart Study. Circulation, 103(9), 1245–1249.

Goff David C., Lloyd-Jones Donald M., Bennett Glen, Coady Sean, D'Agostino Ralph B., Gibbons Raymond, Greenland Philip, Lackland Daniel T., Levy Daniel, O'Donnell Christopher J., Robinson Jennifer G., Schwartz J. Sanford, Shero Susan T., Smith Sidney C., Sorlie Paul, Stone Neil J., & Wilson Peter W. F. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. Circulation, 129(25_suppl_2), S49–S73.

Grundy, S. M. (2002). Third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. Circulation, 106(25), 3143-3421.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3(null), 1157–1182.

Haasenritter, J., Stanze, D., Widera, G., Wilimzig, C., Abu Hani, M., Sönnichsen, A. C., Bösner, S., Rochon, J., & Donner-Banzhoff, N. (2012). Does the patient with chest pain have a coronary heart disease? Diagnostic value of single symptoms and signs – a meta-analysis. Croatian Medical Journal, 53(5), 432–441.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction 2nd ed. New York: Springer.

Hubert, H. B., Feinleib, M., McNamara, P. M., & Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: A 26-year follow-up of participants in the Framingham Heart Study. Circulation, 67(5), 968–977.

Khemphila, A., & Boonjing, V. (2011). Heart Disease Classification Using Neural Network and Feature Selection. 2011 21st International Conference on Systems Engineering, 406–409.

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [Cs].

Lakrat, R., Sookaneknun, P., & Sutapak, U. (2019). Mortality Rate of Coronary Heart Disease at Detudom Royal Crown Prince Hospital, Ubonratchathani in 2013 and 2016. Isan Journal of Pharmaceutical Sciences, IJPS (Isan J Pharm Sci), 15(1), 93–104.

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. Lancet, 383(9921), 999–1008.

Mitchell, G. F., Hwang, S.-J., Vasan, R. S., Larson, M. G., Pencina, M. J., Hamburg, N. M., Vita, J. A., Levy, D., & Benjamin, E. J. (2010). Arterial stiffness and cardiovascular events: The Framingham Heart Study. Circulation, 121(4), 505–511.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. ArXiv:1608.06048 [Cs, Stat].

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on International Conference on Machine Learning, 807–814.

National Health Service. (2020, Mar 10). Coronary heart disease causes. [Online]. Retrieved from: https://www.nhs.uk/conditions/coronary-heart-disease/causes/.

National Health Service. (2020, Mar 10). Coronary heart disease overview. [Online]. Retrieved from: https://www.nhs.uk/conditions/coronary-heart-disease/.

National Health Service. (2020, Mar 10). Coronary heart disease symptoms. [Online]. Retrieved from: https://www.nhs.uk/conditions/coronary-heart-disease/symptoms/.

National Heart, Lung, and Blood Institute. (2019, Aug 26). Coronary Heart Disease. [Online]. Retrieved from: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease.

Nielsen, M. A. (2015). Neural networks and deep learning. San Francisco, CA: Determination press.

O'Donnell, C. J., & Elosua, R. (2008). [Cardiovascular risk factors. Insights from Framingham Heart Study]. Revista Espanola De Cardiologia, 61(3), 299–310.

Tsao, C. W., & Vasan, R. S. (2015). Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. International Journal of Epidemiology, 44(6), 1800–1813.

Wilson Peter, W. F., D'Agostino, Ralph B., Levy, D., Belanger, Albert M., Silbershatz, H., & Kannel, William B. (1998a). Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation, 97(18), 1837–1847.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3), 408–421.

# ประวัติผู้วิจัย

**ชื่อ-สกุล** กนกภัณฑ์ โชติชะวารานนท์

**วัน เดือน ปีเกิด** 8 ธันวาคม 2538

**สถานที่เกิด** กรุงเทพมหานคร

**ประวัติการศึกษา** ปี พ.ศ. 2557

- มัธยมศึกษา (ปลาย) จากโรงเรียนมัธยมสาธิตวัดพระศรีมหาธาตุ มหาวิทยาลัยราชภัฏพระนคร

ปี พ.ศ. 2561

- ปริญญาตรี วิทยาศาสตรบัณฑิต สาขา วิทยาการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย

**ประสบการณ์ทำงาน** ปี 2564 – ปัจจุบัน

- นักวิชาการคอมพิวเตอร์ปฏิบัติการ โรงพยาบาลเวชการุณรัศมิ์ สำนักการแพทย์ กรุงเทพมหานคร