การเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล

Efficiency comparison of cardiovascular risk prediction models using data mining algorithms

นงเยาว์ ในอรุณ[1*]
Nongyao Nai-arun[1*]

Received: 20 January 2021 ; Revised: 15 February 2021 ; Accepted: 1 March 2021

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ (1) สร้างแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล ได้แก่ โครงข่ายประสาทเทียม ฟอเรสต์แบบสุ่ม เค-เนียร์เรสเนเบอร์ นาอีฟเบย์ และต้นไม้ตัดสินใจ (2) ใช้อัลกอริทึมทั้ง 5 วิธี พร้อมการเลือกคุณสมบัติของแอตทริบิวต์ และ (3) เปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation โดยเครื่องมือที่ใช้ในการวิจัยคือโปรแกรม MySQL และ RapidMiner Studio และชุดข้อมูลเป็นคนที่ผ่านการคัดกรองผู้ป่วยโรคหัวใจและหลอดเลือดที่รวบรวมข้อมูลมาจากสำนักงานสาธารณสุขจังหวัดสระบุรี ระหว่างปี พ.ศ. 2561-2562 จากโรงพยาบาลในจังหวัดสระบุรี 12 แห่ง และโรงพยาบาลส่งเสริมสุขภาพตำบล 126 แห่ง จำนวน 31,929 คน ผลการวิจัย พบว่า แบบจำลองที่มีประสิทธิภาพการทำนายดีที่สุดคือ แบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ มีค่าความถูกต้อง 99.29% และต่ำสุดคือ แบบจำลองต้นไม้ตัดสินใจ มีค่าความถูกต้อง 70.39% งานวิจัยนี้ สรุปผลว่า แบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติของแอตทริบิวต์เป็นแบบจำลองที่เหมาะสมที่สุด ที่ควรนำไปพัฒนาเว็บแอปพลิเคชันเพื่อการวินิจฉัยโรคเบื้องต้นเกี่ยวกับความเสี่ยงโรคหัวใจและหลอดเลือด

คำสำคัญ: โรคหัวใจและหลอดเลือด เหมืองข้อมูล โครงข่ายประสาทเทียม

Abstract

This research aimed (1) to create cardiovascular risk diagnosis prediction models using algorithms including Neural Network, Random Forest, Naïve Bayes, K-Nearest Neighbors and Decision Tree (2) five algorithms were used with attribute selection and (3) comparing the model performance using 10f Fold cross validation method. Tools usee were MySQL and RapidMiner Studio programs. The data see comprised people who had been screened as patients with cardiovascular disease that were collected from the Saraburi Provincial Public Health Office during 2018-2019 from 12 Saraburi hospitals and 126 health promoting hospitalsreIt was found that the model with the best prediction performance was the neural network model with attribute selection having 99.29% accuracy, and the lowest was the decision tree model with 70.39% accuracy. This research concluded that the neural network model with attribute selection of the best qualificationt should be further developed for early diagnosis of cardiovascular risk web applications.

Keywords: Cardiovascular Disease, Data Mining, Neural Network

[1] ผู้ช่วยศาสตราจารย์, สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา, จังหวัดพระนครศรีอยุธยา 13000

[1] Assist. Prof., Department of Information Technology, Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Phranakhon Si Ayutthaya 13000

* Corresponding author; Department of Information Technology, Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University. nongyaonaiarun@gmail.com

บทนำ

องค์การอนามัยโลก (World Health Organization, 2019) เปิดเผยว่าในปี พ.ศ. 2559 มีผู้เสียชีวิตปีละ 57 ล้านคน ซึ่งกลุ่มโรคหัวใจและหลอดเลือดเป็นสาเหตุของการตายของคนทั่วโลกเป็นอันดับหนึ่ง มีผู้เสียชีวิตจำนวน 9.2 ล้านคนโดยประมาณ ในประเทศไทยจากการรายงานของกรมควบคุมโรค กระทรวงสาธารณสุข ปี พ.ศ. 2560 มีผู้เสียชีวิตจากโรคหัวใจและหลอดเลือดประมาณร้อยละ 12 ของสาเหตุของการเสียชีวิตทั้งหมด และยังพบว่าผู้เสียชีวิตจากโรคหัวใจและหลอดเลือดประมาณ 20,746 คน คิดเป็นร้อยละ 21.8 ต่อประชากร 10,000 คน และในปี พ.ศ. 2561 พบอัตราความชุกของผู้ป่วยจากโรคหัวใจและหลอดเลือดที่สูงกว่าในปี พ.ศ. 2557-2560 มีอัตราความชุกประมาณ 1,396,400 ต่อประชากร 10,000 คน จากข้อมูลทั้งการตายและป่วยด้วยโรคหัวใจและหลอดเลือด แสดงให้เห็นว่าโรคหัวใจและหลอดเลือดยังคงมีความรุนแรงเพิ่มขึ้น เพราะมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่องข้อมูล จากสถานการณ์ปัจจุบันและรูปแบบการบริการด้านโรคไม่ติดต่อเรื้อรัง (NCDs) ของกรมการแพทย์ในปี พ.ศ. 2557 พบประเทศไทยมีค่าใช้จ่ายในการรักษาพยาบาลเฉลี่ยของผู้ป่วยโรคหัวใจและหลอดถึงประมาณ 6,906 ล้านบาทต่อปี และยังเป็นสาเหตุของการสูญเสียปีสุขภาวะในอันดับต้น ของประชากรไทยวัยทำงาน ส่งผลกระทบต่อคุณภาพชีวิตของประชากร ทำให้เกิดความสูญเสียทางเศรษฐกิจจากการเสียชีวิตก่อนวัยอันควร ทำให้ส่งผลกระทบในระดับต่างๆ ได้แก่ ส่วนบุคคล ครอบครัว สังคม และประเทศชาติ (กรมควบคุมโรค, 2562)

ปัจจุบันมีการใช้เทคโนโลยีสารสนเทศในกระทรวงสาธารณสุขมีการเก็บรวบรวมข้อมูลสุขภาพ ทุกหน่วยบริการจะมีการบันทึกข้อมูลการให้บริการประจำวันในโปรแกรมระบบสารสนเทศของหน่วยบริการ (Hospital Information System-HIS) ซึ่งมีหลากหลายโปรแกรม และใช้ประโยชน์จากข้อมูลที่บันทึกไว้จัดทำเป็นสารสนเทศเพื่อใช้พัฒนาการบริการของหน่วยบริการ จากนั้นให้มีการใช้ประโยชน์ข้อมูลให้มากขึ้น จึงมีการรวบรวมข้อมูลที่หน่วยบริการบันทึกไว้ มารวบรวมไว้ที่ระดับที่สูงขึ้น เช่น อำเภอ จังหวัด กระทรวงฯ เพื่อจัดทำเป็นสารสนเทศในการปฏิบัติตามภารกิจของแต่ละระดับที่เกี่ยวข้องเป็นศูนย์รวมข้อมูลในแต่ละระดับ ในเอกสารฉบับนี้ใช้คำว่า “คลังข้อมูลสุขภาพ (Health Data Center-HDC)” แทนการรวบรวมข้อมูลจากหน่วยบริการเป็นศูนย์รวมข้อมูลสุขภาพ ตามที่กระทรวงสาธารณสุขได้กำหนดแนวทางการพัฒนาระบบข้อมูลข่าวสารสุขภาพให้จัดเก็บข้อมูลเป็นฐานข้อมูลรายบุคคลในระดับต่างๆ ประกอบด้วยฐานข้อมูลระดับสถานีอนามัย ศูนย์สุขภาพชุมชน รวมทั้งฐานข้อมูลระดับโรงพยาบาล และฐานข้อมูลนั้นได้รับการออกแบบให้มีโครงสร้างได้ระหว่างหน่วยงาน โดยให้สถานบริการส่งออกข้อมูลด้านการแพทย์และสุขภาพ ตามมาตฐานโครงสร้าง 43 แฟ้ม ซึ่งประกอบด้วยข้อมูลผู้ป่วยนอก ข้อมูลผู้ป่วยใน และข้อมูลด้านการป้องกัน ส่งเสริม และฟื้นฟู เพื่อให้สอดคล้องกับการนำไปใช้ประโยชน์ร่วมกันทั้งระดับสถานบริการ ระดับจังหวัด และส่วนกลาง สามารถเชื่อมโยงข้อมูล นำไปใช้ประโยชน์ได้อย่างมีประสิทธิภาพ โดยโปรแกรมที่ใช้ในการรวบรวมข้อมูล คือ HOSxP เป็นโปรแกรมสำหรับสถานพยาบาล โรงพยาบาลส่งเสริมสุขภาพตำบล และโรงพยาบาลซึ่งมีเป้าหมายที่จะพัฒนาระบบสารสนเทศที่มีประสิทธิภาพมาก สามารถนำไปใช้งานได้จริงในระดับส่งเสริมสุขภาพตำบลไปจนถึงโรงพยาบาลศูนย์ ข้อมูลจะมาจากโครงสร้างมาตรฐาน 43 แฟ้มที่สถานบริการทุกที่ในจังหวัดส่งข้อมูลทุกเดือนแล้ว Upload เข้าไปรวบรวมที่คลังข้อมูลระดับจังหวัด (Health Data Center : HDC) (กระทรวงสาธารณสุข, 2561)

มาตรการเชิงรุกในการป้องกันและควบคุมอุบัติการณ์โรคหัวใจและหลอดเลือดของกรมควบคุมโรค กระทรวงสาธารณสุข ได้มอบหมายให้โรงพยาบาลและโรงพยาบาลส่งเสริมสุขภาพตำบลทำการรวบรวมข้อมูล ด้วยการใช้แบบประเมินความเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือด โดยใช้อิทธิพลของปัจจัยเสี่ยงต่างๆ มีตัวแปร ได้แก่ เพศ ประวัติการสูบบุหรี่ ประวัติการเป็นโรคเบาหวาน อายุ ส่วนสูง เส้นรอบเอว ค่าความดันโลหิต และค่าโคเรสเตอรอลในเลือด เพื่อค้นหากลุ่มเสี่ยงต่อการเกิดโรคดังกล่าว

อัลกอริทึมการทำเหมืองข้อมูล (Data Mining) เป็นอัลกอริทึมในการวิเคราะห์ข้อมูลขนาดใหญ่ เพื่อค้นหารูปแบบ (Pattern) หรือกฎ (Rule) ที่มีในฐานข้อมูลขนาดใหญ่ และเป็นกระบวนการดึงข่าวสาร ค้นหาความรู้ที่น่าสนใจและเป็นประโยชน์จากฐานข้อมูลขนาดใหญ่ (Han & Kamber, 2006) การทำเหมืองข้อมูลเกี่ยวข้องกับทฤษฎีและหลักการจากสาขาวิชาต่างๆ ได้แก่ ระบบฐานข้อมูล การจดจำรูปแบบ เทคโนโลยีคลังข้อมูล การวิเคราะห์ทางสถิติ เครือข่ายประสาทเทียม การเรียนรู้ของเครื่องจักร การค้นคืนข้อมูล การประมวลผลภาพ และการวิเคราะห์ข้อมูลเชิงพื้นที่ของเหตุการณ์ (Witten & Frank, 2011) วิธีการทำเหมืองข้อมูลอาศัยเทคนิคการวิเคราะห์ที่ซับซ้อนกว่าการวิเคราะห์ทางสถิติและการสืบค้นแบบสอบถามเชิงโครงสร้างทั่วไป เช่น ภาษา SQL ฯลฯ การทำเหมืองข้อมูลที่นิยม ได้แก่ การหาความสัมพันธ์ (Association) การจัดกลุ่ม (Clustering) การจำแนกประเภทข้อมูล (Classification) การวิเคราะห์ข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data Analytics) เป็นต้น (เอกสิทธิ์ พัชรวงศ์ศักดา, 2563)

จากเหตุผลดังกล่าว ผู้วิจัยจึงมีแนวคิดในการนำข้อมูลจากคลังข้อมูลระดับจังหวัด (HDC) ในส่วนของการ

คัดกรองความเสี่ยงโรคหัวใจและหลอดเลือด กรณีศึกษาสำนักงานสาธารณสุขจังหวัดสระบุรี มาใช้ในการสร้างแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล ได้แก่ Neural Network, Random Forest, Naïve Bayes, K-Nearest Neighbors and Decision Tree แล้วทำการเปรียบเทียบประสิทธิภาพของแบบจำลองเพื่อให้ได้องค์ความรู้ใหม่และแบบจำลองที่เหมาะสมที่สุดที่นำไปพัฒนาต่อสำหรับการวินิจฉัยโรคเบื้องต้นเกี่ยวกับความเสี่ยงโรคหัวใจและหลอดเลือดต่อไป

การทบทวนวรรณกรรม

ผู้วิจัยได้ทำการศึกษาและค้นคว้าเอกสาร แนวคิด ทฤษฎี เกี่ยวกับโรคหัวใจและหลอดเลือก อัลกอริทึมเหมืองข้อมูลและงานวิจัยที่เกี่ยวข้อง ดังนี้

1. โรคหัวใจและหลอดเลือด

โรคหัวใจและหลอดเลือด(Cardiovascular Disease) เกิดจากมีเนื้อเยื่ออยู่ในผนังของหลอดเลือด ทำให้มีไขมันสะสม เกิดการพอกตัวที่หนาขึ้น ทำให้หลอดเลือดตีบและเกิดการแข็งตัว จนการไหลเวียนของเลือดลดลง ตีบตัน เลือดที่จะไปเลี้ยงกล้ามเนื้อหัวใจมีจำนวนน้อยลง มีผลทำให้เกิดภาวะกล้ามเนื้อหัวใจขาดเลือด อาการของโรคหัวใจและหลอดเลือด ได้แก่ การเจ็บแน่นหน้าอก เหมือนมีอะไรมากดทับหน้าอก ระยะเวลา 15 นาที ถึง 30 วินาที มีอาการหายใจเหนื่อย หอบ หายใจไม่ออก นอนราบไม่ได้ เวียนศรีษะ หน้ามืดจะเป็นลม รวมทั้งมีอาการหมดสติเพราะว่าเลือดไปเลี้ยงสมองไม่เพียงพอ อาการเตือนของโรคหัวใจและหลอดเลือด คือ เจ็บกลางหน้าอกที่บริเวณเหนือลิ้นปี่ขึ้นมาเล็กน้อย เจ็บแบบจุกแน่นๆ เหมือนมีอะไรมากดทับหรือบีบไว้ อาการเจ็บจะร้าวไปที่คอหรือขากรรไกร ขณะออกกำลังกายมักจะเจ็บมากที่ไหล่ซ้าย หรือบางคนมีอาการจุกแน่นลิ้นปี่เหมือนอาหารไม่ย่อย และอาจทำให้เสียชีวิต

หน่วยงานในกระทรวงสาธารณสุขได้รับมอบหมายให้ทำงานร่วมกับภาครัฐบาลและเอกชน เพื่อทำงานด้านการป้องกันโรค การส่งเสริมสุขภาพ และการปรับพฤติกรรมทางสุขภาพ ที่มีผลต่อการเกิดโรคเรื้อรัง (NCDs) ได้แก่ โรคความดันโลหิตสูง โรคเบาหวาน โรคหัวใจและหลอดเลือด ซึ่งเป็นโรคที่เกิดจากพฤติกรรมการใช้ชีวิต เพื่อทำให้คนไทยไม่เจ็บป่วย มีสุขภาพดี และช่วยลดค่าใช้จ่ายทางการแพทย์ได้ จากการศึกษาปัจจัยเสี่ยงที่ทำให้เกิดโรคหัวใจและหลอดเลือด ประกอบด้วย ภาวะจากความดันโลหิตสูง ภาวะโรคเบาหวาน ภาวะไขมันในเลือดสูง การสูบบุหรี่ การดื่มเครื่องดื่มแอลกอฮอล์ การเป็นโรคอ้วนและลงพุง ไม่ออกกำลังกาย ความเครียด พักผ่อนไม่เพียงพอ และการรับประทานรสเค็ม มัน หวานมากเกินไป ประชาชนจึงควรตระหนักหรือประเมินความเสี่ยงของตัวเอง รวมทั้งควรตรวจสุขภาพอย่างสม่ำเสมอ (กรมควบคุมโรค, 2562)

แนวทางการประเมินความเสี่ยงต่อโรคหัวใจและหลอดเลือด ตามแนวทางของกรมควบคุมโรค กระทรวงสาธารณสุข เพื่อให้บุคลากรทางการแพทย์ใช้ประเมินความเสี่ยงต่อโรค ในบุคคลที่ไม่เคยเป็นโรคนี้มาก่อน (Primary Prevention) ด้วยแนวทาง 2 ส่วน คือ (1) การประเมินโอกาสเสี่ยงต่อกลุ่มเป็นโรคเบาหวานและโรคความดันโลหิตสูงสามารถใช้ได้ในประชาชนทั่วไป และกลุ่มเสี่ยงสูงต่อการเกิดโรคเบาหวานและโรคความดันโลหิตสูงและผู้มีภาวะอ้วนที่มีอายุ 35 ปีขึ้นไป โดยการคัดกรองด้วยวาจา (Verbal Screening) และ (2) การบริการหลังการประเมินความเสี่ยง โดยวิธีการประเมินความเสี่ยง มีขั้นตอน คือ 1) โรงพยาบาลและโรงพยาบาลส่งเสริมสุขภาพตำบล หรือสถานบริการทำการตรวจหาค่าโคเรสเตอรอลในเลือด 2) เลือกตารางว่าเป็นโรคเบาหวานหรือไม่ 3) เลือกเพศชายหรือหญิง 4) เลือกการสูบบุหรี่ว่าสูบหรือไม่สูบ 5) เลือกอายุ 6) เลือกค่าความดันโลหิต 7) เลือกส่วนสูง และ 8) เลือกเส้นรอบเอว

2. อัลกอริทึมเหมืองข้อมูล

อัลกอริทึมเหมืองข้อมูล (Data Mining) เป็นการวิเคราะห์หรือสืบค้นความรู้ที่เป็นประโยชน์และเป็นที่น่าสนใจที่อยู่ในฐานข้อมูลขนาดใหญ่ หรือเป็นวิธีการที่ใช้จัดการกับข้อมูลจำนวนมาก โดยจะนำข้อมูลที่มีอยู่มาทำการวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการพยากรณ์หรือทำนายสิ่งที่เกิดขึ้นใหม่ ซึ่งวิธีการสืบค้นหาความรู้หรือความจริงที่แฝงอยู่ในฐานข้อมูล เป็นกระบวนการขุดค้นสิ่งที่ยังไม่ทราบมาก่อนที่มีอยู่ในฐานข้อมูลนั้น ซึ่งการทำเหมืองข้อมูลเป็นกระบวนการทำงานที่สกัดหาข้อมูล (Extract Data) จากฐานข้อมูลขนาดใหญ่เพื่อให้ได้สารสนเทศที่มีประโยชน์ที่ยังไม่ทราบมาก่อน (สายชล สินสมบูรณ์ทอง, 2560) อัลกอริทึมการทำเหมืองข้อมูลแบ่งเป็นหลายประเภทตามลักษณะของการทำงาน ในการวิจัยนี้ใช้อัลกอริทึมการจำแนกประเภทข้อมูล (Classification) ซึ่งเป็นเทคนิคในการวิเคราะห์จากการจดจำหรือเรียนรู้จากรูปแบบของข้อมูลในอดีต และมาสร้างเป็นแบบจำลองเพื่อใช้คาดการณ์หรือทำนาย (Predict) ค่าให้กับข้อมูลใหม่ (Chakrabarti, *et al.*, 2009) ในงานวิจัยนี้ใช้ 5 อัลกอริทึม ได้แก่

2.2.1 โครงข่ายประสาทเทียม (Neural Network)

แบบจำลองที่สร้างขึ้นจะจำลองการทำงานของสมองมนุษย์ที่มีการทำงานหลายหลายรูปแบบ ประกอบด้วยเซลล์ประสาท ต่างๆ ที่เชื่อมโยงกันและการส่งกระแสไฟฟ้าเพื่อให้เซลล์ประสาทถัดไปทำงาน แนวคิดของอัลกอริทึมโครงข่ายประสาทเทียม ใช้การคำนวณที่เรียกว่า ฟังก์ชันการถ่ายโอน

(Transfer Function) ค่าถ่วงน้ำหนัก (Weight) และค่าไบแอส (Bias) เป็นส่วนประกอบในการจำลองคุณสมบัติของเซลล์ประสาท และเซลล์ประสาทหลายๆ ตัวถูกเชื่อมต่อกันทำให้เกิดลักษณะของโครงข่ายเป็นชั้นๆ (Layer) แบ่งออกเป็น 3 ชั้น คือ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นนำออก (Output Layer) โดยมีลักษณะโครงสร้างดัง Figure 1

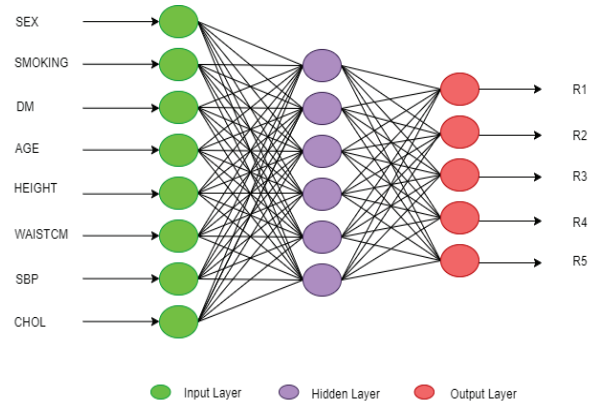


Figure 1 Neural network algorithm

2.2.2 ฟอเรสต์แบบสุ่ม (Random Forest)

อัลกอริทึมฟอเรสต์แบบสุ่มพัฒนาขึ้นมาโดย (Breiman, 2001) เป็นวิธีการจำแนกประเภทข้อมูลซ้ำหลายๆ ครั้ง โดยจะเพิ่มขั้นตอนการทำการสุ่มเลือกตัวแปรจากตัวแปรทั้งหมดและใช้เฉพาะตัวแปรที่สุ่มได้ในการสร้างโมเดลหลายๆ ครั้ง เช่น สมมุติว่าจะสร้าง K โมเดล โดยจะทำการสุ่มชุดข้อมูลเริ่มต้นแบบใส่กลับ จนได้ข้อมูล K ชุด (D1, D2, …, Dk) แต่ละชุดข้อมูล ทำการสุ่มตัวแปรมาบางส่วนได้ข้อมูล K ชุด (S1, S2, …, Sk) แล้วนำมาสร้างโมเดลการจำแนกประเภท K โมเดล (M1, M2, …, Mk) ใช้แต่ละโมเดลในการทำนายผลที่ได้จากการทำนายของทั้ง K โมเดล (Ali, *et al*., 2012) และนับโหวตมากที่สุด วิธีการทำงานดัง Figure 2

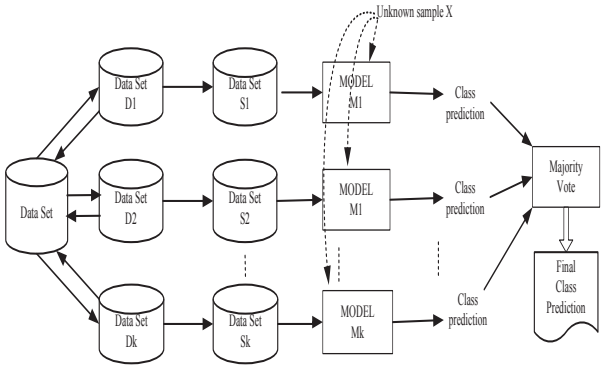


Figure 2 Random forest algorithm

2.2.3 เค-เนียร์เรสเนเบอร์ (K-Nearest Neighbors)

เป็นอัลกอริทึมที่มีหลักการทำงานและวิธีการจำแนกข้อมูลที่ไม่ซับซ้อน โดยพิจารณาจากชุดข้อมูลใกล้เคียงกับข้อมูลที่กำลังสนใจ เรียกชุดข้อมูลเหล่านี้ว่า เพื่อนบ้านใกล้ที่สุด (Nearest Neighbor) มีวิธีการเลือกชุดข้อมูลเพื่อนบ้านใกล้ที่สุดกับค่าของชุดข้อมูลที่กำลังพิจารณามาจำนวน K ตัว การคำนวณค่าความคล้ายคลึงด้วยการใช้ค่าระยะทางน้อยที่สุด โดยส่วนมากมักจะใช้วิธีการวัดระยะทาง (Distance) ด้วยยูคลิเดียน (Euclidean) ดังสมการที่ (1) (Han, *et al*., 2011)

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (1)$$

โดยที่ n คือ จำนวนตัวแปรทั้งหมด, $dist(X_1, X_2)$ คือ ค่าระยะทางระหว่างสองตัวแปร, X_1 คือ ค่าตัวแปรที่ และ X_2 คือ ค่าตัวแปรที่ 2

2.2.4 นาอีฟเบย์ (Naïve Bayes)

เป็นอัลกอริทึมในการจำแนกประเภทข้อมูลโดยใช้หลักสถิติในการทำนายความน่าจะเป็น (Probability) ของข้อมูล ด้วยหลักการทฤษฎีของเบย์ (Bayesian Theorem) สามารถทำนายค่าคลาสเป้าหมายของตัวอย่างด้วยการพิจารณาค่าความน่าจะเป็นมากที่สุดระหว่างทุกค่าคลาสที่เป็นไปได้ การคำนวณค่าความน่าจะเป็นโดยรวม จะต้องคำนวณค่าความน่าจะเป็นของแต่ละคุณสมบัติของแต่ละคลาส ดังสมการที่ (2)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad (2)$$

โดยที่ $P(A \mid B)$ คือ ความน่าจะเป็นของ A เมื่อกำหนด B, $P(B \mid A)$ คือ ความน่าจะเป็นของ B เมื่อกำหนด A, $P(A)$ คือ ความน่าจะเป็นการเกิดเหตุการณ์ A และ $P(B)$ คือ ความน่าจะเป็นการเกิดเหตุการณ์ B

2.2.5 ต้นไม้ตัดสินใจ (Decision Tree)

เป็นอัลกอริทึมที่ใช้วิธีการแตกแขนงจากโหนดราก (Root Node) เป็นโหนดภายใน (Branch Node) แตกออกไปตามเงื่อนไขหรือข้อมูล จนไปสู่โหนดใบ (Leaf Node) เป็นแบบจำลองที่มีการเชื่อมโยงระหว่างสิ่งที่สนใจกับผลสรุปที่อาจเกิดขึ้นจากค่าของเหตุการณ์ (Jones, 2008) โหนดภายในของต้นไม้ตัดสินใจจะประกอบเป็นคุณลักษณะของข้อมูล ซึ่งเมื่อสอดคล้องกับข้อมูลใดก็จะใช้คุณลักษณะนั้นเป็นตัวตัดสินใจว่าข้อมูลจะไปทิศทางใด โหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายใน

และสุดท้ายคือ โหนดใบ เป็นกลุ่มผลลัพธ์ในการจำแนกประเภทข้อมูล ผลลัพธ์ที่ได้สามารถแปลงเป็นกฎ (Rule) ได้ การสร้างจะเริ่มพิจารณาที่โหนดรากเป็นอันดับแรกก่อนจะดำเนินการพิจารณา โหนดใบและกิ่งก้านที่แตกแขนงต่อไป โดยต้องคำนวณหาข้อมูลที่เหมาะสมที่จะเป็น โหนดราก ซึ่งพิจารณาจากค่า Information Gain ที่มากที่สุด ที่ได้จากการคำนวณค่า Entropy เพื่อให้การจำแนกและแยกแยะข้อมูลให้อยู่ในกลุ่มเดียวกันมากที่สุด หลังจากที่ได้โหนดรากแล้วก็จะสร้าง Decision Tree ในลำดับต่อไป จนกระทั่งถึงโหนดใบ และได้ Decision Tree ที่สมบูรณ์ ดังสมการที่ 3 และ 4

$$E(S) = \sum_{i=1}^{n} - P(V_i) \log_2 P(V_i) \quad (3)$$

โดยที่ $E(S)$ คือ ค่า Entropy ของเซต (S), S คือ ข้อมูลทั้งหมด, $P(V_i)$ คือ ค่าความน่าจะเป็นของข้อมูลที่สนใจ และ V_i คือ คุณลักษณะข้อมูลที่สนใจ

$$Gain(S, A) = E(S) - \sum_{v \in values(A)} \frac{|S_v|}{S} E(S_v) \quad (4)$$

โดยที่ $Gain(S, A)$ คือ ค่า Gain ของเหตุการณ์ที่สนใจ, $E(S)$ คือ ค่า Entropy ของเซต (S) ก่อนการแบ่งกลุ่มย่อย, $E(S_v)$ คือ ค่า Entropy ของเซตข้อมูลกลุ่มย่อย v, S_v คือ จำนวนข้อมูลในเซตข้อมูลกลุ่มย่อย v, S คือ จำนวนข้อมูลในเซต (S) และ A คือ ตัวแปรที่สนใจ

3. งานวิจัยที่เกี่ยวข้อง

บุณยานุช ใหมเง่า และคณะ (2560) พัฒนาโปรแกรมสำหรับช่วยวิเคราะห์อัตราเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือด ด้วย โครงข่ายประสาทเทียม (Neural Network) และต้นไม้ตัดสินใจ (Decision Tree) และเพื่อพัฒนาโปรแกรมวิเคราะห์อัตราเสี่ยงการเกิดโรคหัวใจและหลอดเลือด ซึ่งในการทดลองสร้างโมเดลของทั้งสองอัลกอริทึมใช้ โปรแกรม WEKA ผลการทดลองสร้างโมเดลพบว่า อัลกอริทึมที่สามารถคาดการณ์จำแนกข้อมูลอัตราเสี่ยงของการเป็น โรคหัวใจและหลอดเลือดได้ดีที่สุดคือ อัลกอริทึมโครงข่ายประสาทเทียม ซึ่งมีค่าที่โปรแกรมจำแนกได้ ถูกต้องสูงสุดคิดเป็นร้อยละ 97.297

สราวุธ มีศรี และคณะ (2560) ทำการศึกษาการวินิจฉัยโรคหัวใจโดยใช้ตัวจำแนกผสม ชุดข้อมูลโรคหัวใจจำนวน 303 เรคคอร์ด 14 แอตทริบิวต์ ในขั้นตอนแรกถูกจำแนกด้วยอัลกอริทึมวิธีเบย์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้ที่สุดเคตัว และได้นำมารวมกันโดยใช้โครงข่ายประสาทเทียมด้วยการเรียนรู้แบบแพร่กระจายย้อนกลับ จากนั้นในขั้นตอนที่สองผลลัพธ์จากตัวจำแนกเดี่ยวจะเป็นข้อมูลนำเข้าสำหรับการจำแนกด้วยโครงข่ายประสาทเทียม จากผลการทดลองตัวจำแนกผสมให้ความถูกต้องมากกว่า ซึ่งค่าความถูกต้อง 86.16% และอัตราบวกเท็จดีกว่าตัวจำแนกอื่นๆ

Assari *et al*. (2017) นำเสนอการวินิจฉัยโรคหัวใจโดยใช้เทคนิคการทำเหมืองข้อมูล โดยได้อธิบายว่าที่ผ่านมาโรคหัวใจได้รับการระบุว่าเป็นสาเหตุสำคัญของการเสียชีวิตทั่วโลก อย่างไรก็ตามโรคนี้ถือได้ว่าเป็นโรคที่สามารถป้องกันได้และควบคุมได้มากที่สุดในเวลาเดียวกัน การทดลองใช้เทคนิคการทำเหมืองข้อมูลเดียวกันนำไปสู่ผลลัพธ์ที่ต่างกันในชุดข้อมูลที่แตกต่างกัน การศึกษาครั้งนี้ได้รับความช่วยเหลือผู้เชี่ยวชาญด้านการดูแลสุขภาพในการวินิจฉัยโรคหัวใจและประเมินปัจจัยเสี่ยง เทคนิคการทำเหมืองข้อมูลได้ถูกนำไปใช้กับชุดข้อมูลที่สัมพันธ์กัน ได้มีการพัฒนาแบบจำลองขึ้นโดยใช้กฎที่แยกออกมาด้วยโปรแกรม Visual Studio ใช้เทคนิคเหมืองข้อมูล 4 เทคนิค ได้แก่ Decision Tree, Bayesian Network, K-nearest Neighbors, Support Vector Machines ผลการวิจัยสรุปว่าเทคนิค Support Vector Machines มีความแม่นยำสูงสุด (84.33%)

Suksawatchon *et al*. (2018) พัฒนาระบบผู้เชี่ยวชาญด้านการวิเคราะห์ความเสี่ยงด้านสุขภาพสำหรับผู้ดูแลครอบครัวคนพิการโดยใช้เทคนิคเหมืองข้อมูล บทความนี้มีการนำเสนอระบบวิเคราะห์ความเสี่ยงด้านสุขภาพหรือ HRASซึ่งเป็นระบบผู้เชี่ยวชาญใหม่ในการระบุระดับความเสี่ยงต่อสุขภาพใน 3 ด้าน ได้แก่ ด้านสุขภาพจิต ร่างกาย และสังคม ระบบ HRAS รวบรวมข้อมูลสุขภาพผ่านแบบสอบถามออนไลน์และแสดงผลการวิเคราะห์ด้วยวิธี RAC ใช้อัลกอริทึมการจำแนกข้อมูลด้วย Rule-base และสร้าง RAC ประเมินด้วยวิธี K-Fold Cross Validation และผู้เชี่ยวชาญ ผลการประเมินพบว่า Neural Network มีประสิทธิภาพดีที่สุดโดยรวมซึ่งมีความแม่นยำสูงกว่า 90% ในชุดข้อมูลสุขภาพทั้งหมด ดังนั้น Neural Network จึงเป็นลักษณนามที่เหมาะสมที่สุดสำหรับงานนี้

Nai-arun & Moungmai (2020) นำเสนอบทความวิจัยการแบบจำลองการทำนายการวินิจฉัยความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้เทคนิคเหมืองข้อมูล ได้แก่ Decision Tree, Logistic Regression, Back-propagation Neural Network, K-nearest Neighbors, Random Forest และ Naïve Bayes ด้วยโปรแกรม WEKA ผลสรุปได้ว่าเทคนิค Logistic Regression (99.940%) ให้ค่าความถูกต้องการทำนายดีที่สุด รองลงมาคือ Back-propagation Neural Network (98.105%) และ Radon Forest (95.627%) ตามลำดับ

วิธีดำเนินการวิจัย

การทำวิจัยครั้งนี้เป็นการวิจัยแบบประยุกต์ (Applied Research) โดยขั้นตอนแรกผู้วิจัยได้ดำเนินการขอจริยธรรมในมนุษย์ ณ มหาวิทยาลัยนเรศวร และได้รับรองโครงการวิจัยตามแนวทางหลักจริยธรรมการวิจัยในคนที่เป็นมาตรฐานสากล หมายเลขโครงการ 095/62 ซึ่งได้ผ่านการรับรองจริยธรรมในมนุษย์ เมื่อวันที่ 6 มีนาคม 2563 เรียบร้อยแล้ว หลังจากนั้นได้ดำเนินการตามขั้นตอนการวิจัยและกรอบแนวคิด (Conceptual Framework) ดัง Figure 3

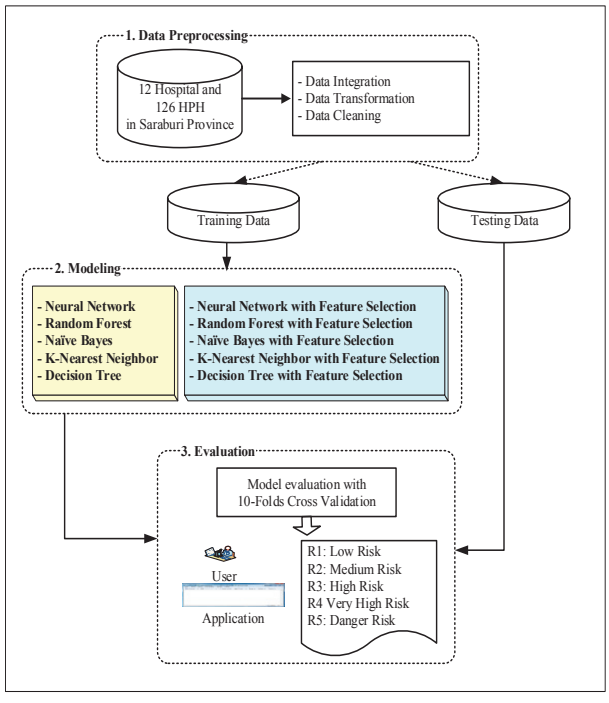


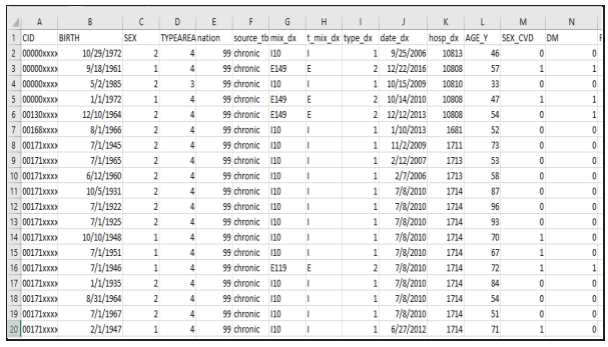
Figure 3 Conceptual framework

จาก Figure 3 แสดงกรอบแนวคิดการวิจัยและขั้นตอนการดำเนินการวิจัย แบ่งออกเป็น 3 ขั้นตอน ดังนี้

1. การจัดเตรียมข้อมูล (Data Preprocessing)

การรวมข้อมูล (Data Integration) ผู้วิจัยได้ทำการรวบรวมข้อมูลจากสำนักงานสาธารณสุขจังหวัดสระบุรี ที่อยู่ระหว่างปี พ.ศ.2561-2562 เป็นข้อมูลการคัดกรองความเสี่ยงโรคหัวใจและหลอดเลือด มาจากโรงพยาบาล 12 แห่ง และโรงพยาบาลส่งเสริมสุขภาพตำบล 126 แห่ง ในจังหวัดสระบุรี ใช้ตามแนวทางการประเมินความเสี่ยงต่อการเกิดโรคหัวใจและขาดเลือด ของกรมควบคุมโรค กระทรวงสาธารณสุข ที่กำหนดให้ โรงพยาบาลส่งเสริมสุขภาพตำบล (รพ.สต.) และโรงพยาบาลต่างๆ ใช้ทำการคัดกรองผู้ป่วยโรคหัวใจและหลอดเลือด

การแปลงข้อมูล (Data Transformation) มีขั้นตอน คือ (1) ใช้โปรแกรม MySQL ในการจัดการข้อมูลทั้งหมดที่ได้มาจากสำนักงานสาธารณสุขจังหวัดสระบุรี ทำการรวมและคิวรีข้อมูลจากไฟล์ข้อมูลที่ได้มา ดัง Figure 4



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CID	BIRTH	SEX	TYPEAREA	nation	source_tb	mix_dx	t_mix_dx	type_dx	date_dx	hosp_dx	AGE_Y	SEX_CVD	DM
2	00000xxxx	10/29/1972	2	4	99	chronic	I10	I	1	9/25/2006	10813	46	0	0
3	00000xxxx	9/18/1961	1	4	99	chronic	E149	E	2	12/22/2016	10808	57	1	1
4	00000xxxx	5/2/1985	2	3	99	chronic	I10	I	1	10/15/2009	10810	33	0	0
5	00000xxxx	1/1/1972	1	4	99	chronic	E149	E	2	10/14/2010	10808	47	1	1
6	00130xxxx	12/10/1964	2	4	99	chronic	E149	E	2	12/12/2013	10808	54	0	1
7	00168xxxx	8/1/1966	2	4	99	chronic	I10	I	1	1/10/2013	1681	52	0	0
8	00171xxxx	7/1/1945	2	4	99	chronic	I10	I	1	11/2/2009	1711	73	0	0
9	00171xxxx	7/1/1965	2	4	99	chronic	I10	I	1	2/12/2007	1713	53	0	0
10	00171xxxx	6/12/1960	2	4	99	chronic	I10	I	1	2/7/2006	1713	58	0	0
11	00171xxxx	10/5/1931	2	4	99	chronic	I10	I	1	7/8/2010	1714	87	0	0
12	00171xxxx	7/1/1922	2	4	99	chronic	I10	I	1	7/8/2010	1714	96	0	0
13	00171xxxx	7/1/1925	2	4	99	chronic	I10	I	1	7/8/2010	1714	93	0	0
14	00171xxxx	10/10/1948	1	4	99	chronic	I10	I	1	7/8/2010	1714	70	1	0
15	00171xxxx	7/1/1951	1	4	99	chronic	I10	I	1	7/8/2010	1714	67	1	0
16	00171xxxx	7/1/1946	1	4	99	chronic	E119	E	2	7/8/2010	1714	72	1	1
17	00171xxxx	1/1/1935	2	4	99	chronic	I10	I	1	7/8/2010	1714	84	0	0
18	00171xxxx	8/31/1964	2	4	99	chronic	I10	I	1	7/8/2010	1714	54	0	0
19	00171xxxx	7/1/1967	2	4	99	chronic	I10	I	1	7/8/2010	1714	51	0	0
20	00171xxxx	2/1/1947	1	4	99	chronic	I10	I	1	6/27/2012	1714	71	1	0

Figure 4 Information from the Saraburi Provincial Public Health Office

(2) ทำการแปลงค่าความเสี่ยงโรคหัวใจและหลอดเลือด (RISK_SCORE) ข้อมูลค่าระดับความเสี่ยงที่ได้มาเป็นตัวเลข ด้วยสูตรของสำนักงานสาธารณสุข (Thai CV Risk Score) ดังนี้

```
cvd_func(AGE_Y,SEX_CVD,L_SBP,DM,L_SMOKING,L_CHOL,L_
WAIST_CM,L_HEIGHT)
DECLARE FullScore DECIMAL(20,9) DEFAULT 0 ;
DECLARE cvd_score DECIMAL(5,2) DEFAULT 0 ;
IF chol > 0 THEN
 SET FullScore = (0.08183*age) + (0.39499*sex) + (0.02084*sbp) +
0.69974*dm) + (0.00212*chol) + (0.41916*smoking) ;
 SET cvd_score = (1-POWER(0.978296,EXP(FullSco
re-7.04423)))*100 ;
END IF ;
IF chol = 0 THEN
 SET FullScore = (0.079*age) + (0.128*sex) + (0.019350987*sbp) +
(0.58454*dm) + (3.512566 * (waist_cm/height)) + (0.459*smoking) ;
 SET cvd_score =(1-POWER(0.978296,EXP(FullSco
re-7.720484)))*100 ;
END IF ;
IF cvd_score > 0 THEN
 RETURN cvd_score ;
 ELSE
 RETURN 0 ;
END IF ;
```

หลังจากนั้นทำการแปลงค่าให้อยู่ในรูปแบบคลาส 5 คลาส ตามหลักการของระบบการเฝ้าระวัง ควบคุม ป้องกันโรคหัวใจและหลอดเลือด (กระทรวงสาธารณสุข, 2561) ได้แก่

R1 ความเสี่ยงต่ำ มีค่า RISK_SCORE <=10

R2 ความเสี่ยงปานกลาง มีค่า RISK_SCORE = 11-20

R3 ความเสี่ยงสูง มีค่า RISK_SCORE = 21-30

R4 ความเสี่ยงสูงมาก มีค่า RISK_SCORE = 31-40

R5 ความเสี่ยงอันตราย มีค่า RISK_SCORE >= 41

(3) ทำการปรับเปลี่ยนข้อมูลทั้งหมดด้วยโปรแกรม MySQL ได้ชุดข้อมูลทั้งหมด 44,674 คน ดัง Figure 5

	SEX	AGE	DM	SMOKING	SBP	CHOL	WAISTCM	HEIGHT	CLASS
2	F	66	Y	N	147	122	86	156	R3
3	F	62	N	N	102	232	96	165	R1
4	F	45	Y	N	160	153	92	150	R1
5	F	73	N	N	134	179	82	148	R2
6	F	42	N	N	140	202	70	155	R1
7	F	47	Y	N	118	206	78	160	R1
8	F	57	N	N	122	194	90	154	R1
9	F	43	Y	N	120	230	95	150	R1
10	F	53	N	N	139	177	93	150	R1
11	F	55	N	N	140	223	86	152	R1
12	F	38	Y	N	147	191	102	152	R1
13	F	52	N	N	125	223	78	156	R1
14	M	50	N	Y	138	182	86	163	R1
15	F	42	Y	N	146	129	104	157	R1
16	F	48	N	N	128	180	102	165	R1
17	F	52	N	N	121	174	95	160	R1
18	F	46	N	N	128	186	79	152	R1
19	F	52	Y	N	123	209	89	160	R1
20	F	52	Y	N	133	198	82	158	R1
21	F	48	N	N	137	168	82	165	R1
22	F	50	N	N	110	180	86	160	R1
23	F	45	Y	N	115	189	93	160	R1

Figure 5 Data Transformation

การทำความสะอาดข้อมูล (Data Cleaning) นำข้อมูลทั้งหมด 44,674 คน นำมาทำความสะอาด (Data Cleaning) ด้วยการขจัดข้อมูลที่ผิดพลาด (Missing Value) ข้อมูลที่ไม่สมบูรณ์ และข้อมูลที่มีค่าของคุณลักษณะบางอย่างขาดหาย ตัดทิ้งไปจำนวน 12,745 คน เหลือข้อมูลที่มีความสมบูรณ์จำนวน 31,929 คน ที่ประกอบด้วยทั้งหมด 9 แอตทริบิวต์ ได้แก่ แอตทริบิวต์นำเข้า (Input) จำนวน 8 แอตทริบิวต์ และแอตทริบิวต์ผลลัพธ์ (Class) จำนวน 1 แอตทริบิวต์ ดัง Table 1

Table 1 Attribute details

No.	Attribute	Detail	Value of attribute
1	SEX	เพศ	Nominal scale M: ชาย F: หญิง
2	SMOKING	ประวัติการสูบบุหรี่	Nominal scale Y: สูบบุหรี่ N: ไม่สูบบุหรี่
3	DM	ประวัติการเป็นโรคเบาหวาน	Nominal scale Y: เคยเป็นโรคเบาหวาน N: ไม่เคยเป็นโรคเบาหวาน
4	AGE	อายุ (ปี)	Numerical scale
5	HEIGHT	ส่วนสูง (ซม.)	Numerical scale
6	WAISTCM	เส้นรอบเอว (ซม.)	Numerical scale
7	SBP	ค่าความดันโลหิต (mmHg หรือมิลลิเมตรปรอท)	Numerical scale
8	CHOL	ค่าโคเรสเตอรอลรวม (mg/dL หรือ มิลลิกรัม/เดซิลิตร)	Numerical scale
9	CLASS	ค่าระดับความเสี่ยงโรคหัวใจและหลอดเลือด	Ordinal scale R1: ความเสี่ยงต่ำ R2: ความเสี่ยงปานกลาง R3: ความเสี่ยงสูง R4: ความเสี่ยงสูงมาก R5: ความเสี่ยงอันตราย

2. การสร้างแบบจำลอง (Modeling)

การสร้างแบบจำลองแบ่งออกเป็น 2 ขั้นตอน ได้แก่ (1) สร้างแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือด โดยใช้อัลกอริทึมเหมืองข้อมูล ได้แก่ โครงข่ายประสาทเทียม ฟอเรสต์แบบสุ่ม เค-เนียร์เรสเนเบอร์ นาอีฟเบย์ และต้นไม้ตัดสินใจ ด้วยโปรแกรม RapidMiner Studio และได้ทดลองกำหนดค่าพารามิเตอร์ต่างๆ ในโปรแกรม ดัง Figure 6 และ Table 2 จนได้ผลลัพธ์ที่เหมาะสมและให้ค่าประสิทธิภาพดีที่สุด

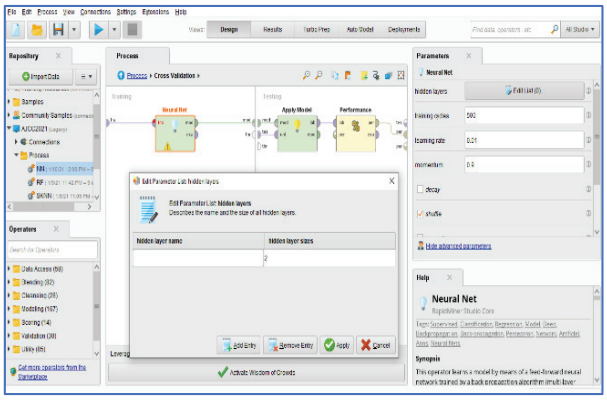


Figure 6 Parameters in the RapidMiner Studio program

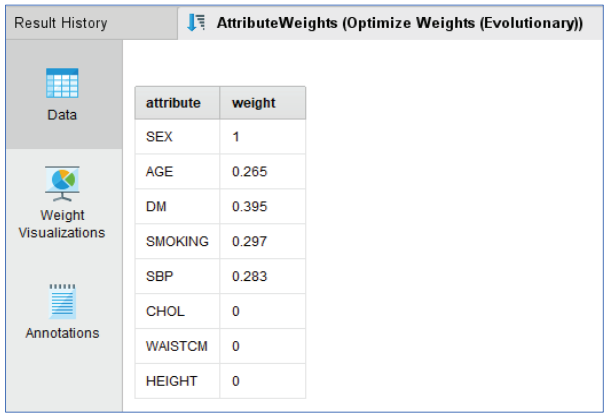


Figure 7 Optimize weights (Evolutionary)

Table 2 Parameter setting details

Algorithm	Parameter setting
Neural Network	Hidden layer sizes = 2 Training cycles = 500 Learning rate =0.01 Momentum = 0.9
Random Forest	Number of trees = 100 Criterion = gain_ratio Maximal depth = 10 Voting Strategy = Confidence vote
K-Nearest Neighbors	k = 4 Measure type = MixedMeasures Mixed measure = MixedEuclidean Distance
Naïve Bayes	Laplace correction
Decision Tree	Number of trees = 100 Criterion = gain_ratio Maximal depth = 10 Confidence = 0.1 Minimal gain = 0.01 Minimal leaf size = 2 Minimal size for split = 4 Number of prepruning alternatives = 3

(2) ใช้อัลกอริทึมทั้ง 5 วิธีพร้อมการเลือกคุณสมบัติของแอตทริบิวต์ (Attribute Selection) ด้วยโปรแกรม RapidMiner Studio โดยเทคนิคในการเลือกแอตทริบิวต์ ใช้โอเปอเรเตอร์ Optimize Weights (Evolutionary) เป็นการใช้สำหรับคำนวณค่าน้ำหนักของแอตทริบิวต์ต่างๆ ของแต่ละแอตทริบิวต์และเลือกแอตทริบิวต์ที่เหมาะสม โดยมีค่าน้ำหนักอยู่ระหว่าง 0-1 ซึ่งค่าน้ำหนัก 0 หมายถึง ไม่มีความสำคัญ และค่าน้ำหนัก 1 หมายถึงสำคัญที่สุด ตัวอย่างดัง Figure 7

3. ประเมินผลโมเดล (Evaluation)

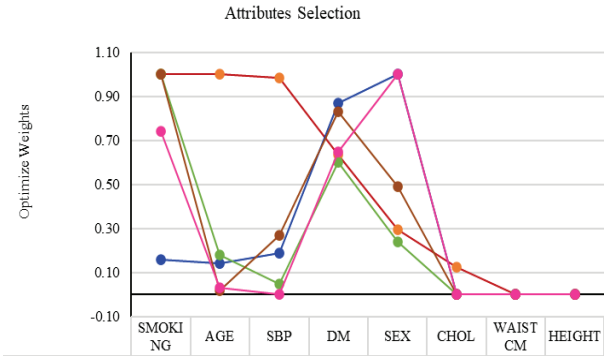
การประเมินผลเพื่อทำการเปรียบเทียบหาประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล ด้วยวิธี 10-Fold Cross Validation โดยทำการแบ่งข้อมูลออกเป็น 10 ชุดเท่าๆ กัน และใช้ข้อมูลสำหรับสร้างแบบจำลองเป็นชุดการเรียนรู้ (Training Data) จำนวน 9 ชุด ข้อมูลสำหรับทดสอบแบบจำลองเป็นชุดทดสอบ (Testing Data) จำนวน 1 ชุด หลังจากนั้นทำการวนรอบจำนวน 10 รอบ หลังจากนั้นทำการเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy)

ผลการวิจัย

การวิจัยนี้ได้ผลการวิจัย ประกอบด้วย ผลการเลือกคุณสมบัติแอตทริบิวต์ ผลการเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy และ Confusion Matrix ดังนี้

1. ผลการเลือกคุณสมบัติของแอตทริบิวต์

ผลการสร้างแบบจำลองด้วยการใช้อัลกอริทึม 5 อัลกอริทึม ได้แก่ โครงข่ายประสาทเทียม ฟอเรสต์แบบสุ่ม เค-เนียร์เรสเนเบอร์ นาอีฟเบย์ และต้นไม้ตัดสินใจ และอัลกอริทึม 5 วิธีพร้อมการเลือกคุณสมบัติของแอตทริบิวต์ (Attribute Selection) ดัง Figure 8



	SMOKING	AGE	SBP	DM	SEX	CHOL	WAIST CM	HEIGHT
NN with AS	1	1	0.98	0.63	0.29	0.13	0	0
RF with AS	0.16	0.14	0.19	0.87	1	0	0	0
KNN with AS	1	0.18	0.05	0.60	0.24	0	0	0
NB with AS	1	0.02	0.27	0.83	0.49	0	0	0
DT with AS	0.74	0.03	0.00	0.65	1	0	0	0

Figure 8 Efficiency attributes selection results

จาก Figure 8 แสดงผลการเลือกแอตทริบิวต์ที่สำคัญในการสร้างแบบจำลองร่วมกับ 5 อัลกอริทึม ได้แก่ Neural Network with Attribute Selection (NN with AS), Random Forest with Attribute Selection (RF with AS), K-Nearest Neighbors with Attribute Selection (KNN with AS), Naïve Bayes with Attribute Selection (NB with AS) และ Decision Tree with Attribute Selection (DT with AS) เพื่อคำนวณหาค่าน้ำหนัก (Weight) ของแอตทริบิวต์ต่าง ๆ และแต่ละอัลกอริทึมเลือกแอตทริบิวต์ที่สำคัญในการสร้างแบบจำลองแตกต่างกัน ผลการวิจัย พบว่า แอตทริบิวต์ที่มีความสำคัญทำการเรียงลำดับได้ดังนี้ SMOKING, DM, SEX, AGE, SBP, CHOL, WAISTCM และ HEIGHT ตามลำดับ

2. ผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง

เปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation หลังจากนั้นทำการหาค่าความถูกต้อง (Accuracy) เพื่อทำการเปรียบเทียบประสิทธิภาพ ผลแสดงดัง Figure 9

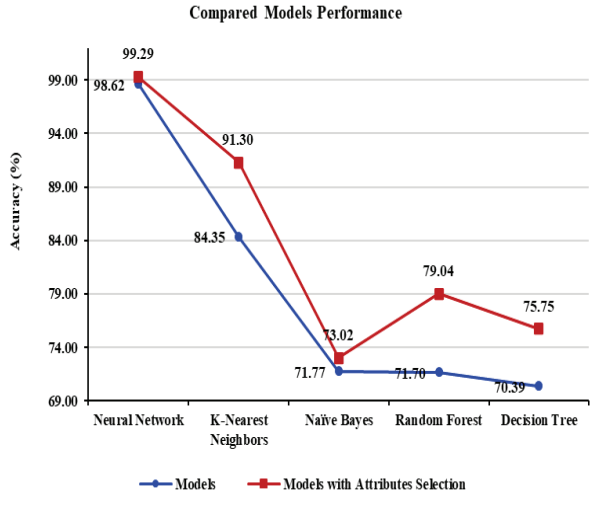


Figure 9 Results of the model performance efficiency

จาก Figure 9 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือด โดยใช้ 5 อัลกอริทึมเหมืองข้อมูล และใช้อัลกอริทึมทั้ง 5 วิธี พร้อมการเลือกคุณสมบัติของแอตทริบิวต์ รวมเป็นทั้งหมด 10 แบบจำลอง ผลการวิจัย พบว่า แบบจำลองที่มีประสิทธิภาพการทำนายสูงที่สุดคือ แบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ มีค่าความถูกต้อง 99.29% และต่ำสุดคือ แบบจำลองต้นไม้ตัดสินใจ มีค่าความถูกต้อง 70.39%

3. ผลการเปรียบเทียบ Confusion Matrix

ผลการวัดประสิทธิภาพของแบบจำลอง ด้วยการใช้ค่าในตาราง Confusion Matrix เพื่อประเมินผลลัพธ์การทำนายและการเปรียบเทียบกับค่าจริง ของแบบจำลองที่ให้ค่าความถูกต้องสูงสุด คือ แบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ และแบบจำลองโครงข่ายประสาทเทียม ดัง Table 3 และ 4

Table 3 The confusion matrix of a neural network model with selection of propertics

R1		Predicted value				
		R2	R3	R4	R5	
Actual value	R1	14983	57	0	0	0
	R2	29	8215	16	0	0
	R3	0	41	3986	29	0
	R4	0	0	22	2027	18
	R5	0	0	0	14	2492

Table 4 The confusion matrix of a neural network model

R1		Predicted value				
		R2	R3	R4	R5	
Actual value value	R1	14978	62	0	0	0
	R2	81	8112	67	0	0
	R3	0	68	3950	38	0
	R4	0	0	46	1967	54
	R5	0	0	0	25	2481

สรุปผลและอภิปรายผลการวิจัย

จากผลการวิจัยพบว่า แบบจำลองที่สร้างอัลกอริทึมโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ (Neural Network with Attribute Selection)ได้ค่าประสิทธิภาพการ

ทำนายความเสี่ยงโรคหัวใจและหลอดเลือดดีที่สุด ซึ่งได้ผลคล้ายกับงานวิจัยของบุณยานุช ใหมเง่าและคณะ และ Suksawatchon และคณะ ที่ได้ผลการทดลองว่าอัลกอริทึมโครงข่ายประสาทเทียมได้ค่าการจำแนกและมีประสิทธิภาพการวิเคราะห์ในงานนั้นๆ ได้สูงสุด ด้วยเหตุผลที่ว่าวิธีการของโครงข่ายประสาทเทียมเป็นการเรียนรู้แบบต้องมีผู้สอน โดยจะหาค่าผลลัพธ์ของโครงข่ายแล้วนำมาเปรียบเทียบกับค่าจริงของข้อมูล หลังจากนั้นทำการคำนวณค่าความผิดพลาด (Error) แล้วใช้ค่าความผิดพลาดนี้เป็นการเรียนรู้ของโครงข่ายแบบย้อนกลับ เพื่อทำการปรับค่าน้ำหนัก (Weight) เส้นเชื่อมต่อ แล้วทำการเรียนรู้ซ้ำๆ จนกระทั่งค่าความผิดพลาดน้อยลงมากที่สุด จะทำให้ได้ค่าน้ำหนักที่เหมาะสมที่ในสุด พร้อมทั้งได้ทำการคัดเลือกแอตทริบิวต์ที่เหมาะสมและสำคัญมาใช้ในการสร้างแบบจำลอง จึงทำให้ได้แบบจำลองที่ดีที่สุดในการนำไปใช้พัฒนาระบบแอปพลิเคชันต่างๆ ในการคัดกรองผู้ป่วยใหม่ที่มีความเสี่ยงโรคหัวใจและหลอดเลือด เพื่อให้บุคลากรทางการแพทย์ใช้ประเมินโอกาสเสี่ยงต่อโรคหัวใจและหลอดเลือดในบุคคลที่ไม่เคยเป็นโรคนี้มาก่อน และเมื่อพบผู้ป่วยที่มีความเสี่ยงก็จะนำไปสู่การตรวจและรักษาโรคได้อย่างทันท่วงที และทำให้ลดอัตราการเสียชีวิตลงได้

จากผลการวิจัยที่ได้ในการคัดเลือกแอตทริบิวต์ที่สำคัญในการนำมาสร้างแบบจำลองร่วมกับ 5 อัลกอริทึม เพื่อคำนวณหาค่าน้ำหนักของแอตทริบิวต์ต่างๆ พบว่าแอตทริบิวต์ที่มีความสำคัญโดยเรียงลำดับได้ดังนี้ ประวัติการสูบบุหรี่ ประวัติการเป็นโรคเบาหวาน เพศ อายุ ค่าความดันโลหิต ค่าโคเรสเตอรอลรวม เส้นรอบเอว และส่วนสูง แต่อย่างไรก็ตามแนวทางการประเมินความเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือดของกระทรวงสาธารณสุขก็ให้ใช้ปัจจัยเสี่ยงทั้งหมดดังกล่าวในการคัดกรองผู้ป่วยเป็นปัจจัยสำคัญ รวมทั้งการปรับเปลี่ยนพฤติกรรมด้านสุขภาพที่ไม่ถูกต้องเหมาะสม เช่น การรับประทานอาหารที่ไม่สมดุลหรือมากเกินพอดี ไม่ออกกำลังกาย ความเครียด และการพักผ่อนที่ไม่เพียงพอ การสูบบุหรี่ การดื่มเครื่องดื่มที่มีแอลกอฮอล์ ส่งผลให้อ้วน น้ำหนักเกิน มีความดันโลหิตสูง ไขมันในเลือดสูง โรคเบาหวาน และเป็นผลที่นำไปสู่ความเสี่ยงการเกิดโรคหัวใจและหลอดเลือด

ข้อเสนอแนะการทำวิจัยครั้งต่อไป ควรนำอัลกอริทึมเหมืองข้อมูลอื่นๆ มาสร้างโมเดล เช่น เทคนิคการเรียนรู้ร่วมกัน (Ensemble Learning) ได้แก่ อัลกอริทึม Vote, Bagging, Boosting เพื่อหาประสิทธิภาพการทำนายที่เหมาะสมกว่า เพื่อก่อให้เกิดประโยชน์และสามารถนำไปใช้สำหรับการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดได้ดี

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณสำนักงานสาธารณสุขจังหวัดสระบุรี ที่ให้ความอนุเคราะห์ข้อมูลของการคัดกรองผู้ป่วยโรคหัวใจและหลอดเลือด โดยผู้วิจัยได้หนังสือขอความอนุเคราะห์และผ่านการอนุญาตให้นำข้อมูลมาใช้ในการศึกษาและการวิจัยครั้งนี้แล้ว และขอขอบพระคุณคณะกรรมการจริยธรรมในมนุษย์เครือข่ายภูมิภาค มหาวิทยาลัยนเรศวร ที่ได้รับรองโครงการวิจัยตามแนวทางหลักจริยธรรมการวิจัยในคนที่เป็นมาตรฐานสากล ของหมายเลขโครงการ 095/62 ซึ่งได้ผ่านการรับรองจริยธรรมในมนุษย์ เมื่อวันที่ 6 มีนาคม 2563

เอกสารอ้างอิง

กรมควบคุมโรค กระทรวงสาธารณสุข. (2562). *แนวทางการเฝ้าระวังโรค ของกระทรวงสาธารณสุข*. องค์การรับส่งสินค้าและวัสดุภัณฑ์.

กระทรวงสาธารณสุข. (2561). *ระบบการเฝ้าระวัง ควบคุม ป้องกันโรคหัวใจและหลอดเลือด* โรงพิมพ์สำนักพระพุทธศาสนาแห่งชาติ.

บุณยานุช ใหมเง่า และคณะ. (2560). โปรแกรมช่วยวิเคราะห์อัตราเสี่ยงต่อการเป็นโรคหัวใจและหลอดเลือด. *วารสารวิชาการเทคโนโลยีอุตสาหกรรม. มหาวิทยาลัยราชภัฏสวนสุนันทา. 5*(1), 55-65.

สายชล สินสมบูรณ์ทอง. (2560). *การทำเหมืองข้อมูล เล่ม 1 : การค้นหาความรู้จากข้อมู*. (พิมพ์ครั้งที่ 2). จามจุรีโปรดักส์.

สราวุธ มีศรี, ศุภกานต์ พิมลธเรศ และ อัจฉรา มหาวีรวัฒน์. (2560). *การวินิจฉัยโรคหัวใจโดยใช้ตัวจำแนกผสม*. [ปริญญานิพนธ์วิทยาศาสตร์มหาบัณฑิต ไม่ได้ตีพิมพ์]. จุฬาลงกรณ์มหาวิทยาลั..

เอกสิทธิ์ พัชรวงศ์ศักดา. (2563). *Big Data and Machine Learning*: ไอดีซี พรีเมียร์.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision tree. *International Journal of Computer Science, 9*(5), 272-278.

Assari, R., Azimi, P., & Taghva, M. R. (2017). Heart Disease Diagnosis Using Data Mining Techiques. *International Journal of Economics & Management Science, 6*(3), 72-79.

Breiman, L. (2001). Random forests. *Journal of Machine Learning, 45*, 5–32.

Chakrabarti, S., Cox, E., Frank, E., Guting, R. H., Han, J., Jiang, X., Kamber, M. ₿ Lightstone, S. S. (2009). *Data mining: Know it all*. Morgan Kaufmann.

Han, J. & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.)*. Morgan Kaufmann.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining concepts and techniques*. (3rd ed.). Morgan Kaufmann.

Jones, M. T. (2008). *Artificial intelligence*. Infinity Science Press.

Nai-arun N., & Moungmai M. (2020). Diagnostic Prediction Models for Cardiovascular Disease Risk using Data Mining Techniques. *Journal of ECTI TRANSACTIONS ON COMPUTER AND INFORMATION TECHNOLOGY, 14*(2), 113-121.

Suksawatchon, U., Suksawatchon J., & Lawang W. (2018). Health Risk Analysis Expert System for Family Caregiver of Person with Disabilities using Data Mining Techniques. *Journal of ECTI TRANSACTIONS ON COMPUTER AND INFORMATION TECHNOLOGY. 12*(1), 62-72.

Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniqus.* (2nd ed.). USA: Morgan Kaufmann.

World Health Organization. (2019, November 15th). *Cardiovascular Diseasrom* https://www.who.in/Cardiovascuar_Disease/.