การหาประสิทธิภาพแบบจำลองการทำนายระดับความอ้วน จากพฤติกรรมการกิน และสภาพร่างกาย

[ธัญชนก กิ่งปรุ]1 [สุชารัตน์ กองฉลาด]2 [ศรัณยพร ฉิมกูล]3 [น้ำทิพย์ บวรอารักษ์สกุล]4 [ณัชพล ชูผล]5 และ[พิชญสินี กิจวัฒนาถาวร]6

[Thanchanok Kingpru]1, [Sucharat Kongchalart]2, [Saranyaporn Chimkun]3 [Namthip Bovornaraksakun]4 [Natchapol Choopol]5 and [Pitchayasini Kitwatthanathawon]6

สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี

[B6501297@g.sut.ac.th]1, SucharatKongchalart@gmail.com]2 [saranyaporn4616@gmail.com]3 [B6530990@g.sut.ac.th]4 [sahapol1608@gmail.com]5 and [pichak@sut.ac.th]6

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการทำนายทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย โดยใช้อัลกอริทึมเหมืองข้อมูล ได้แก่ โครงข่ายประสาทเทียม, ป่าสุ่ม, นาอีฟเบย์, และต้นไม้ตัดสินใจ และเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation โดยเครื่องมือที่ใช้ในการวิจัยคือโปรแกรม WEKA และชุดข้อมูลเป็นข้อมูลสำหรับการประมาณระดับโรคอ้วนในบุคคลจากประเทศเม็กซิโก เปรู และโคลัมเบีย โดยพิจารณาจากพฤติกรรมการกิน และสภาพร่างกายของพวกเขา ข้อมูลประกอบด้วยจำนวน 2111 ชุดข้อมูล 17 คุณลักษณะ ผลการวิจัยพบว่า เทคนิคป่าสุ่ม (Random Forest) ให้ค่าความถูกต้อง (Accuracy) มากที่สุด ที่ 95.64% ค่าความแม่นยำ (Precision) ที่ 95.90% ค่าความระลึก (Recall) ที่ 95.60% ค่าความถ่วงดุล (F–measure) ที่ 95.70% และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error : RMSE) 11.48 % จึงสรุปได้ว่าวิธีเทคนิคป่าสุ่ม (Random Forest) มีประสิทธิภาพในการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกายดีที่สุด

***คำสำคัญ**: [แบบจำลอง] [เหมืองข้อมูล] [การทำนาย] [ระดับความอ้วน] [พฤติกรรมการกิน] [สภาพร่างกาย]*

บทนำ

ประชากรที่ป่วยเป็นโรคอ้วนเพิ่มสูงขึ้นอย่างรวดเร็วทั่วโลก ซึ่งโรคอ้วน คือ ภาวะที่ร่างกายมีการสะสมไขมันมากเกินกว่าปกติหรือมากเกินกว่าที่ร่างกายจะเผาผลาญ จึงสะสมพลังงานที่เหลือเอาไว้ในรูปของไขมันตามอวัยวะต่าง ๆ อาจมีความเสี่ยงต่อการเกิดปัญหาสุขภาพ และเป็นสาเหตุของการเกิดโรคเรื้อรังต่าง ๆ ตามมา สาเหตุที่ทำให้เกิดโรคอ้วน แบ่งออกเป็น ปัจจัยภายใน และปัจจัยภายนอก ซึ่งส่วนใหญ่แล้วผู้ที่เป็นโรคอ้วน มักมีสาเหตุจากปัจจัยภายนอก เพราะมีพฤติกรรมการรับประทานที่ตามใจตนเอง จนทำให้รับประทานเกินความต้องการของร่างกาย จากผลการศึกษาของสหพันธ์โรคอ้วน World Obesity Federation ซึ่งเป็นองค์กรในสังกัดองค์การอนามัยโลกหรือ WHO ระบุว่า ปัจจุบันประชากรโลกที่มีภาวะน้ำหนักเกินหรือเป็นโรคอ้วนมีจำนวนประมาณ 2,600 ล้านคนหรือ 38% ของจำนวนประชากรโลกทั้งหมด 8,000 ล้านคน องค์การอนามัยโลก หรือ WHO พบว่า 1 ใน 3 ของประชากรที่เป็นโรคอ้วน หรือประมาณกว่า 600 ล้านคน มีอาการป่วยจากสาเหตุของโรคอ้วน เช่น เบาหวาน ความดันโลหิตสูง หลอดเลือดหัวใจ หลอดเลือดสมอง และรวมถึงมะเร็งบางชนิด ทางสหพันธ์โรคอ้วน ได้คาดว่า ในปี 2035 ตัวเลขจะเพิ่มขึ้นเป็นมากกว่า 4,000 ล้านคน หรือ 51% ของจำนวนประชากรโลกทั้งหมด และสัดส่วนประชากรที่มีภาวะอ้วนรุนแรงจะเพิ่มจากจาก 1 คนต่อประชากร 7 คนในปัจจุบัน เป็น 1 คนต่อประชากร 4 คนในปี 2035 ประชากรโลกเกินครึ่งจะมีภาวะน้ำหนักเกินมาตรฐาน และโรคอ้วน หากรัฐบาลของประเทศต่าง ๆ ไม่แก้ไขปัญหานี้อย่างเร่งด่วน ผศ.พญ.ศานิต วิชานศวกุล อาจารย์ประจำหน่วยโภชนศาสตร์ ภาควิชาอายุรศาสตร์โรงพยาบาลธรรมศาสตร์เฉลิมพระเกียรติ กล่าวว่า อ้วนเป็นโรคที่ป้องกันได้ และควรได้รับการรักษาก่อนจะมีภาวะแทรกซ้อนจากโรคต่าง ๆ ไม่เฉพาะคนไข้ ครอบครัว ระบบสาธารณสุขต้องเข้ามาช่วยเหลือรักษาตั้งแต่เนิ่น จากปัญหาข้างต้น ผู้วิจัยจึงให้ความสำคัญในการนำทฤษฎีการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นกระบวนการวิเคราะห์ข้อมูลเพื่อค้นหารูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ๆ และในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น การพยากรณ์ผู้ป่วยเพื่อพยากรณ์การณ์การอุบัติของโรคต่าง ๆ มาวิเคราะห์เพื่อพยากรณ์หรือคาดการณ์โอกาสการเกิดโรคอ้วนโดยสร้างแบบจำลองสำหรับการคาดการณ์ระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย และเปรียบเทียบประสิทธิภาพของเทคนิคการการทำเหมืองข้อมูล งานวิจัยนี้ได้ใช้เทคนิคการทำเหมืองข้อมูลที่หลากหลายเพื่อพยากรณ์ และเปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล ซึ่งเทคนิคที่จะนำมาใช้ 5 เทคนิค ได้แก่ เทคนิคการเรียนรู้เชิงลึก (Deep Learning) คือวิธีการเรียนรู้แบบอัตโนมัติด้วยการ เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Neurons) โดยนำระบบโครงข่ายประสาท (Neural Network) มาซ้อนกัน หลายชั้น (Layer) และทำการเรียนรู้ข้อมูลตัวอย่าง ซึ่งข้อมูล ดังกล่าวจะถูกนำไปใช้ในการตรวจจับรูปแบบ (Pattern) หรือจัด หมวดหมู่ข้อมูล (Classify the Data), เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลสำหรับการจำแนกข้อมูล (Classification Rules) โดยเป็นการนำข้อมูลมาสร้างแบบจำลองพยากรณ์เพื่อการทำนายหรือการจำแนกออกเป็นประเภทต่าง ๆ โดยมีโครงสร้างในลักษณะที่เป็นต้นไม้ (Sinsomboonthong, 2015) โครงสร้าง

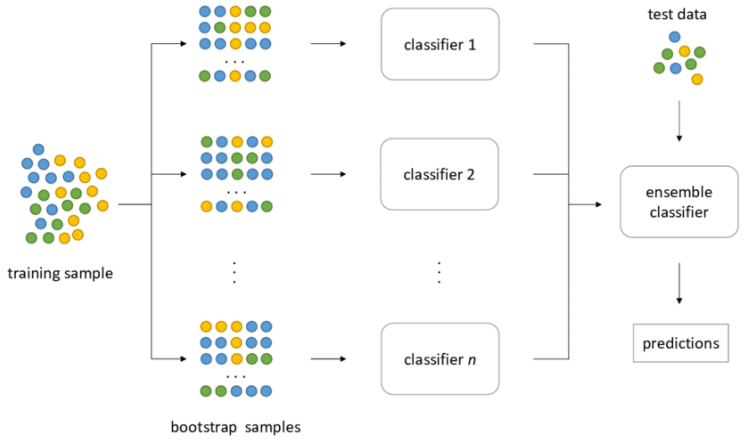
ของต้นไม้ตัดสินใจประกอบด้วยโหนดราก (Root Node) กิ่ง (Branch) และโหนดใบ (Leaf Node), เทคนิคป่าสุ่ม (Random Forest) มีหลักการคือการเทรนแบบจำลองที่เหมือนกันหลาย ๆ ครั้ง (หลาย Instance) บนข้อมูลชุดเดียวกัน โดยแต่ละครั้งของการเทรนจะเลือกส่วนของข้อมูลที่เทรนไม่เหมือนกัน แล้วเอาการตัดสินใจของแบบจำลองเหล่านั้นมาโหวตกันว่า Class ไหนถูกเลือกมากที่สุด, เทคนิคนาอีฟเบย์ (Naïve Bayes) หรือการเรียนรู้แบบเบยส์เป็นวิธีการเรียนรู้ที่อาศัยความน่าจะเป็นตามทฤษฎีของเบย์ส์ (Bayes' Theorem) เป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ (Suwanco et al., 2017) เหมาะกับกรณีของเซ็ตตัวอย่างที่มีจำนวนมาก และลักษณะของตัวอย่างไม่ขึ้นต่อกัน, การ (Linear Regression) คือการนำเอาข้อมูลหรือตัวแปรมาหาความสัมพันธ์กันโดยความสัมพันธ์ของข้อมูลจะออกมาในรูปแบบของการเรียงกันเป็นเส้นตรงหรือใกล้เคียง ผู้วิจัยได้นำเทคนิคที่กล่าวไปข้างต้นนี้ในการสร้างแบบจำลองเพื่อคาดการณ์ระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย

วัตถุประสงค์การวิจัย

1. สร้างแบบจำลองเพื่อคาดการณ์ระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย

ทฤษฎีที่เกี่ยวข้อง

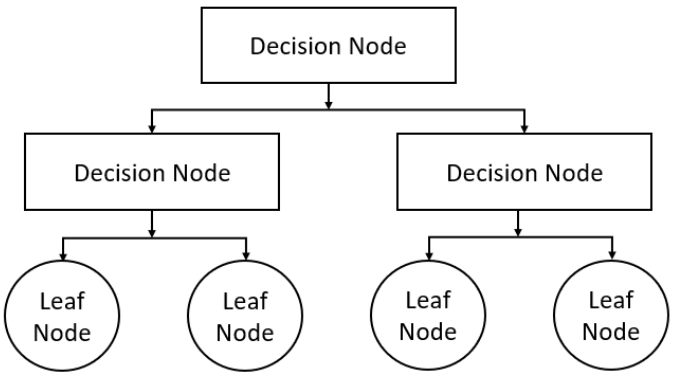
เทคนิคป่าสุ่ม (Random Forest) เป็นอัลกอริทึมประเภทหนึ่งของอัลกอริทึมต้นไม้ตัดสินใจที่มีลักษณะแบบ Unpruned หรือ Regression Trees ซึ่งถูกสร้างขึ้นโดยการสุ่มเลือกตัวอย่างข้อมูล. หลักการของ Random Forest คือการสร้าง model จาก Decision Tree หลาย ๆ model (ตั้งแต่ 10 model ถึงมากกว่า 1000 model) โดยแต่ละ model จะได้รับ data set ที่ไม่เหมือนกันซึ่งเป็น subset ของ data set ทั้งหมด. เมื่อทำการ prediction, แต่ละ Decision Tree ทำการ prediction แต่ละตัว และคำนวณผล prediction ด้วยการ vote output ที่ถูกเลือกโดย Decision Tree มากที่สุด หรือหาค่าเฉลี่ยจาก output ของแต่ละ Decision Tree (ในกรณี regression) ดังตัวอย่างในภาพที่ 1



ภาพที่ 1 : เทคนิคป่าสุ่ม (Random Forest)

หมายเหตุ.จาก https://medium.com/analytics-vidhya/lets-talk-about-random-forests-524ae1138d8b

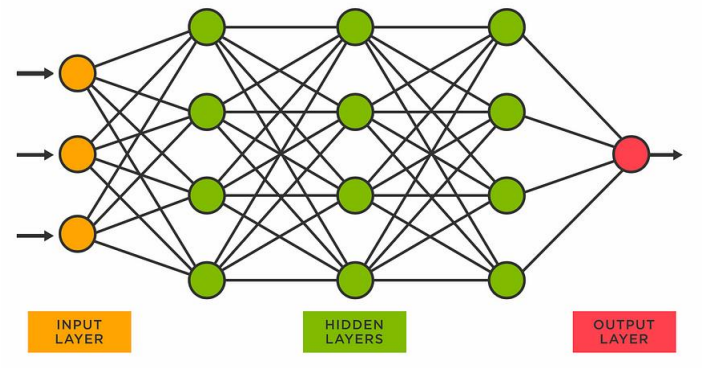
เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นอัลกอริทึมที่ใช้วิธีการแตกแขนงจากโหนดราก (Root Node) เป็นโหนดภายใน (Branch Node) แตกออกไปตามเงื่อนไขหรือข้อมูล จนไปสู่โหนดใบ (Leaf Node) เป็นแบบจำลองที่มีการเชื่อมโยงระหว่างสิ่งที่สนใจกับผลสรุปที่อาจเกิดขึ้นจากค่าของเหตุการณ์ (Jones, 2008) โหนดภายในของต้นไม้ตัดสินใจจะประกอบเป็นคุณลักษณะของข้อมูล ซึ่งเมื่อสอดคล้องกับข้อมูลใด ๆ ก็จะใช้คุณลักษณะนั้นเป็นตัวตัดสินใจว่าข้อมูลจะไปทิศทางใด โหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายใน และสุดท้ายคือ โหนดใบ เป็นกลุ่มผลลัพธ์ในการจำแนกประเภทข้อมูล ผลลัพธ์ที่ได้สามารถแปลงเป็นกฎ (Rule) ได้ การสร้างจะเริ่มต้นที่โหนดรากเป็นอันดับแรกก่อนจะดำเนินการพิจารณา โหนดใบ และกิ่งก้านที่แตกแขนงต่อไป โดยต้องคำนวณหาข้อมูลที่เหมาะสมที่จะเป็น โหนดราก ซึ่งพิจารณาจากค่า Information Gain ที่มากที่สุด ที่ได้จากการคำนวณค่า Entropy เพื่อให้การจำแนก และแยกแยะข้อมูลให้อยู่ในกลุ่มเดียวกันมากที่สุด หลังจากที่ได้โหนดรากแล้วก็จะสร้าง Decision Tree ในลำดับต่อไป จนกระทั่งได้ Decision Tree ที่สมบูรณ์ ดังตัวอย่างในภาพที่ 2



ภาพที่ 2 : เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มีโครงสร้างขนาดใหญ่ประกอบด้วยนิวรอน และชั้นซ่อนจำนวนมาก เป็นอัลกอริทึมที่ถูกสร้างขึ้นเพื่อการเรียนรู้ของเครื่องจักร แต่ระดับ Hidden Layer ของการเรียนรู้เชิงลึกมีมากกว่า ANN ซึ่งแต่ละเลเยอร์จะเปรียบเสมือนประกอบด้วยเซลล์ประสาท (Neural) จำนวนมากที่มีหน้าที่ในการประมวลผล โดยเลเยอร์แรกสุดท้ายจะทำหน้าที่ในการรับข้อมูล (Input Layer) และส่งข้อมูลที่ประมวลผลเสร็จแล้วไปยังเลเยอร์สุดท้าย (Output Layer) การส่งข้อมูลแบบนี้มีข้อดีคือแต่ละเลเยอร์อาจทำให้มีค่าถ่วงน้ำหนัก (Weight) ค่าความเอนเอียงของข้อมูล (Bias) และวิธีการประมวลผลทางคณิตศาสตร์ (Activation Function) เป็นอิสระต่อกันถ้าป้อนข้อมูล

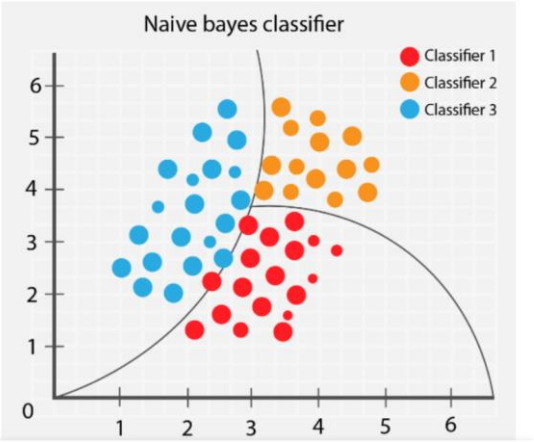
เข้าไปให้กับแบบจำลองมากเท่าไหร่ แต่ละเลเยอร์ก็จะสามารถสกัดคุณลักษณะที่มีความซับซ้อนมากขึ้นทำให้ระบบสามารถตัดสินใจได้ใกล้เคียงกับมนุษย์มากยิ่งขึ้น (สมศักดิ์ ศรีสวการย์ และ สมัย ศรีสวย, 2563) ดังตัวอย่างในภาพที่ 3



ภาพที่ 3 : เทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

หมายเหตุ. จาก https://medium.com/mit-6-s089-intro-to-quantum-computing/quantum-neural-networks-7b5bc469d984

เทคนิคนาอีฟเบย์ (Naïve Bayes) เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์ (Bayes' Theorem) (รุ่งโรจน์ บุญมา และ นิเวศ จิระวิชิตชัย, 2562) เพื่อหาสมมติฐานใดน่าจะถูกต้องที่สุด โดยใช้ความรู้ก่อนหน้า (Prior Knowledge) ได้แก่ ความน่าจะเป็นก่อนหน้าสำหรับสมมติฐานหนึ่ง ๆ ร่วมกับข้อมูล เช่น ความน่าจะเป็นที่สังเกตได้สำหรับสมมติฐานหนึ่ง ๆ เพื่อหาสมมติฐานที่ดีที่สุด การเรียนรู้แบบเบย์อาศัยหลักการของการคำนวณความน่าจะเป็นของแต่ละสมมติฐาน ในที่นี่คือคลาสเป้าหมายหรือผลลัพธ์การทำนายโดยการเรียนรู้แบบเบย์เป็นการเรียนรู้เพิ่มได้ เนื่องจากตัวอย่างใหม่ที่ได้มาถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นทำให้มีการเรียนรู้ที่เปลี่ยนไป วิธีการนี้แบบจำลองจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มี ซึ่งการทำนายค่าคลาสเป้าหมายของตัวอย่างใช้ความน่าจะเป็นมากที่สุดของทุกสมมติฐานจากทฤษฎีของเบย์ เราสามารถคำนวณความน่าจะเป็นของสมมติฐานต่าง ๆ ดังตัวอย่างในภาพที่ 4



ภาพที่ 4 : เทคนิคนาอีฟเบย์ (Naïve Bayes)

หมายเหตุ. จาก https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/

งานวิจัยที่เกี่ยวข้อง

สุพาพร บรรดาศักดิ์, เบญญาภา ศรีสว่าง, และสุภาวดี ทองคำ (สุพาพร, เบญญาภา และสุภาวดี, 2559) ศึกษาถึงปัจจัยต่าง ๆ ที่ส่งผลต่อการเป็นโรคข้อเข่าเสื่อม โดยจากการศึกษาพบว่าอัลกอริทึม Naive byes มีประสิทธิภาพค่าความถูกต้อง เท่ากับ 92.1466 %, 99.7382 % อัลกอริทึม Sequential Minimal Optimization (SMO)มีประสิทธิภาพค่าความถูกต้องเท่ากับ 87,4346 %, 99.7382 % อัลกอริทึม Decision Tree (J48) มีประสิทธิภาพค่าความ ถูกต้องเท่ากับ 74.3455 %, 99.4764 % อัลกอริทึม Neural network มีประสิทธิภาพค่าความถูกต้องเท่ากับ 89,0052 %, 98,9529 % จะเห็นได้ ว่าอัลกอริทึม Naive byes มีประสิทธิภาพค่าความถูกต้องสูงที่สุด และมีการพยากรณ์ที่เที่ยงตรงสามารถนำไปใช้วิเคราะห์ความเสียงคัดกรองผู้ที่เสี่ยงต่อการเป็นโรคข้อเข่าเสื่อมได้เป็นอย่างดี

นพรัตน์ นนท์ศิริ, ราตรี มนัสศิลา, และกริช สมกันธา (นพรัตน์, ราตรี และกริช, 2564) ได้สร้างแบบจำลองการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานจาก ผลการเปรียบเทียบพบว่า วิธีต้นไม้ตัดสินใจให้ค่าประสิทธิภาพสูงสุด โดยมีค่าความถูกต้อง 93.73% วิธีนาอีฟเบย์ค่าความถูกต้อง 88.92% วิธีความใกล้เคียงกันที่สุด และวิธีซัพพอร์ตเวกเตอร์แมชชีนค่าความถูกต้อง 86.97% และ 86.13% ตามลำดับ จะพบว่า วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการสร้าง แบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน เนื่องจากเป็นวิธีที่ไม่มีการแจกแจงหรือไม่ใช้พารามิเตอร์ซึ่งไม่ได้ ขึ้นอยู่กับสมมุติฐานการแจกแจงความน่าจะเป็น อีกทั้งสามารถจัดการกับข้อมูลที่มีมิติสูงได้อย่างแม่นยำ เหมาะสมที่จะนำแบบ จำลองไปพัฒนาระบบจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน ทางการแพทย์ในการวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานต่อไป เพื่อเป็นแนวทางในการสนับสนุนการตัดสินใจ

นงเยาว์ ในอรุณ (นงเยาว์, 2564) ได้สร้างแบบจำลองการทำนายความเสี่ยงโรคหัวใจ และหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล พบว่าแบบจำลองที่มีประสิทธิภาพการทำนายดีที่สุดคือ แบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ มีค่าความถูกต้อง 99.29% และต่ำสุดคือ แบบจำลองต้นไม้ตัดสินใจ มีค่าความถูกต้อง 70.39%

สุรวัชร ศรีเปารยะ, สายชล สินสมบูรณ์ทอง (สุรวัชร, สายชล, 2560) ได้เปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุด เพื่อวัดประสิทธิภาพการจำแนกกลุ่มโดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย จากการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรัง โดยเปรียบเทียบจากค่าความถูกต้อง และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย วิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความ ถูกต้อง คือ 100 % และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

เบญจภัค จงหมื่นไวย์ (เบญจภัค, 2558) ได้เปรียบเทียบปัจจัยข้อมูลผู้สูงอายุ (เพศ อายุ และโรคประจำตัว) ซึ่งโรคประจำตัว ผู้สูงอายุสามารถจัดกลุ่ม พบว่า แบบจำลอง Decision Tree ร้อยละค่าความถูกต้องของการทานายเท่ากับ 8.56% และ 91.43% เป็นค่าจริง แบบจำลอง Naïve Bayes ซึ่งพบว่า ร้อยละค่าความถูกต้องของการทานายเท่ากับ 10.27 % และ 89.72 % เป็นค่าจริง ดังนั้น การวัดประสิทธิภาพการทางานค่าทานายของผู้สูงอายุพบผู้สูงอายุที่ไม่มีโรคประจาตัวมากกว่าผู้สูงอายุที่มีโรคประจาตัวถึงร้อยละ 70.20 % และแบบจำลองที่เหมาะสมสาหรับผู้สูงอายุในกลุ่มนี้คือ Decision Tree

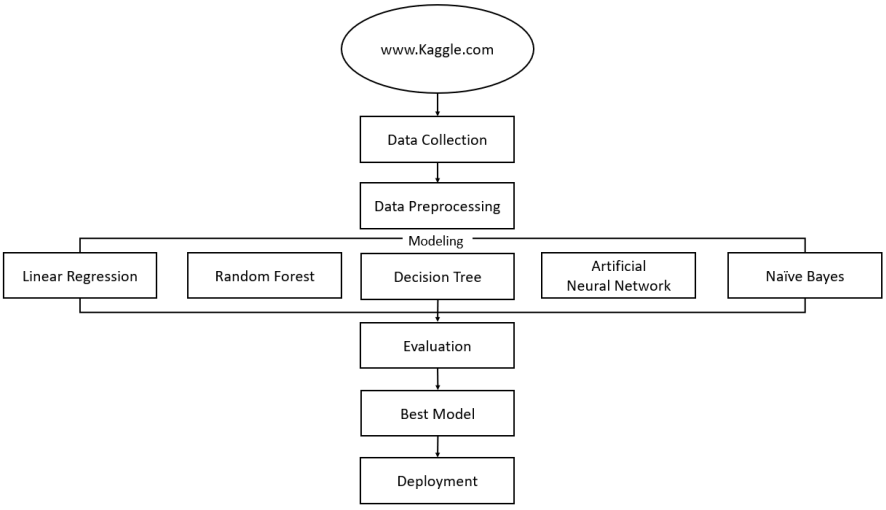
กฤตกนก ศรีพิมพ์สอ และกิตติพล วิแสง (กฤตกนก และกิตติพล, 2566) พัฒนาแบบจำลองสำหรับการพยากรณ์ผู้ป่วยโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล ผลการวิจัยพบว่าเทคนิคแรนดอมฟอเรส (Random Forest) ให้ค่าความถูกต้องในการทำนายผลเป็นโรค เบาหวานมากที่สุดที่อยู่ที่ 99.75% มีค่าความแม่นยำ (Precision) ที่ 98.50% ค่าความครบถ้วน (Recall) ได้ 98.50% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) ได้ 98.50% และค่าเส้นกราฟ (ROC) ได้ 95.20% สามารถนำผลที่ได้จากงานวิจัยนี้ไปประยุกต์ใช้ในการประกอบการรักษาผู้ป่วยโรคเบาหวานต่อไปในอนาคต

ณัฐพล แสนคำ, ทิพวัลย์ แสนคำ, และธนากร ปุรารัมย์ (ณัฐพล, ทิพวัลย์ และธนากร, 2560) ได้พัฒนาระบบสนับสนุนทางการแพทย์สำหรับคัดกรองผู้ป่วยโรคโตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลเปรียบเทียบประสิทธิภาพของอัลกอริทึมสำหรับทำนายโรคไตเรื้อรัง สรุปได้ว่า เทคนิค Random Forest ที่ใช้ชุดข้อมูลที่มีการเพิ่ม (Oversampling Data) มีประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงที่สุดจากทุก แบบจำลองที่ค่า 97.29% ค่าความแม่นตรง (Precision) ที่ 95,76% และค่าวัดประสิทธิภาพโดยรวม (F-Measure) เท่ากับ 97,44% และนำเทคนิคเหมืองข้อมูลนี้มาพัฒนาเป็นแบบจำลองในการทำนายโรคไต ผลการประเมินประสิทธิภาพในการทำนายโรคไตของระบบพบว่าสามารถทำนายโรคไตของข้อมูลใหม่ ได้ถูกต้อง 95.71% ทั้งนี้ เทคนิคต่าง ๆ และแบบจำลองที่ได้พัฒนาขึ้นจะสามารถนำไปต่อยอด เพื่อพัฒนาระบบสนับสนุนทางการแพทย์ที่มีประสิทธิภาพในอนาคต

	เทคนิคการทำเหมืองข้อมูลที่ใช้ในวิจัย										ข้อมูลที่ใช้ในงานวิจัย			
	เทคนิคป่าสุ่ม	เทคนิคนาอีฟเบย์	เทคนิคโลจิสติก	เทคนิคต้นไม้ตัดสินใจ	เทคนิคโครงข่ายประสาทเทียม	เทคนิคซัพพอรตเวกเตอร์แมชชีน	เทคนิคเพอร์เซ็ปตรอนแบบหลายชั้น	Sequential Minimal Optimization (SMO)	เค-เนียร์เรสเนเบอร์	วิธีฐานกฎ	ข้อมูลส่วนตัว	ข้อมูลเกี่ยวข้องกับโรค	พฤติกรรมการใช้ชีวิต	พฤติกรรมการกิน
สุพาพร บรรดาศักดิ์ และคณะ (2559)		X		X	X			X			X		X	
นพรัตน์ นนท์ศิริ และคณะ (2564)		X		X		X					X	X	X	
นงเยาว์ ในอรุณ (2564)	X	X		X	X				X					
เบญจภัค จงหมื่นไวย์ (2558)		X		X							X	X		
กฤตกนก ศรีพิมพ์สอ และกิตติพล วิแสง (2566)	X	X	X	X			X							
ณัฐพล แสนคำ และคณะ (2560)	X			X										
สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560)		X	X	X	X	X				X	X	X		
งานวิจัยนี้	X	X		X	X						X	X	X	X

วิธีดำเนินงาน

งานวิจัยนี้มีกรอบแนวคิดสำคัญเพื่อสร้างแบบจำลอง และเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลที่ใช้สำหรับการสร้างการหาประสิทธิภาพแบบจำลองการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย เพื่อช่วยคัดกรองผู้ป่วยที่มีโอกาสเกิดโรคอ้วน โดยมีขั้นตอนการดำเนินงาน ดังตัวอย่างในภาพที่ 5 ได้แก่



ภาพที่ 5 : ขั้นตอนการดำเนินงาน

1. ประชากรที่ใช้ในการวิจัยครั้งนี้ คือ บุคคลจากประเทศเม็กซิโก เปรู และโคลัมเบีย โดยพิจารณาจากพฤติกรรมการกิน และสภาพร่างกาย จำนวน 2,111 คน

2. การวิจัยครั้งนี้เป็นการนำข้อมูลที่คัดเลือกมาวิเคราะห์ตามกระบวนการเหมืองข้อมูลของ CRISP-DM (Cross - Industry Standard Process for Data Mining) ประกอบด้วย 6 ขั้นตอน ได้แก่ ขั้นทำความเข้าใจกับปัญหา (Business Understanding) ขั้นทำความเข้าใจ และรวบรวมข้อมูล ที่เกี่ยวข้อง (Data Understanding) ขั้นเตรียมข้อมูล (Data Preparation) ขั้นสร้างแบบจำลอง (Modeling) ขั้นประเมินวัดประสิทธิภาพของโมเดล (Evaluation) และขั้นนำผลลัพธ์ หรือองค์ความรู้ ที่ได้มาไปประยุกต์ใช้ (Deployment) (IBM, 2016)

3. ตัวแปรที่ศึกษา

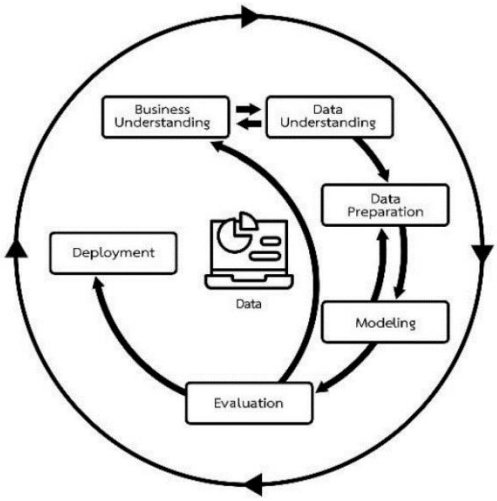
ตัวแปรอิสระ คือ ข้อมูลพื้นฐานของผู้ป่วย ได้แก่ เพศ อายุ ความสูง น้ำหนัก ประวัติการมีสมาชิกครอบครัวน้ำหนักเกิน การกินอาหารแคลอรี่สูง ความถี่ในการกินผัก จำนวนมื้ออาหาร การกินอาหารระหว่างมื้ออาหาร การสูบบุหรี่ การดื่มน้ำ การติดตามแคลอรี่ทุกวัน การออกกำลังกาย การใช้อุปกรณ์เทคโนโลยี การดื่มแอลกอฮอล์ บริการขนส่ง

ตัวแปรตาม คือ ผู้วิจัยได้กำหนดหน้าที่ให้กับคุณลักษณะที่ 17 ระดับโรคอ้วน (NObeyesdad) กำหนดหน้าที่เป็น “Label” หรือตัวแปรตาม (Dependent Variable) เพื่อระบุผลลัพธ์ของการพยากรณ์ระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย

4. เครื่องมือที่ใช้ในการวิจัย

การวิจัยครั้งนี้ เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่ โปรแกรมประยุกต์ Microsoft Excel ใช้ในการคัดเลือกข้อมูล และแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถวิเคราะห์ได้ตามอัลกอริทึม ของการทำเหมืองข้อมูลที่เลือกใช้ และโปรแกรม Weka , และโปรแกรม Rapid Miner Studio 8.2 ใช้ในการทดลองสร้างแบบจำลอง

การวิจัยนี้ผู้วิจัยได้นำข้อมูลพื้นฐานของนักศึกษาจากสำนักส่งเสริมวิชาการ และงานทะเบียนมหาวิทยาลัยราชภัฏเชียงราย โดยนำข้อมูลที่คัดเลือกมาวิเคราะห์ตามกระบวนการทำเหมืองข้อมูล (CRISP-DM) ซึ่งประกอบด้วย 6 ขั้นตอนหลัก ดังนี้ :



ภาพที่ 6 : กระบวนการมาตรฐานในการทำเหมืองข้อมูล

หมายเหตุ. จาก https://www.researchgate.net/figure/CRoss-Industry-Standard-Process-for-Data-Mining-source-wwwcrisp-dmorg_fig1_268274403

1. การทำความเข้าใจปัญหา (Business Understanding)

ผู้วิจัยทำการศึกษาข้อมูลที่เกี่ยวข้องกับสาเหตุที่ทำให้เกิดโรคอ้วน เกิดจากปัจจัยภายนอก เพราะมีพฤติกรรมการรับประทานที่ตามใจตนเอง จนทำให้รับประทานเกินความต้องการของร่างกาย ซึ่งเป็นปัญหาต่อสุขภาพร่างกาย จากผลการศึกษาของ องค์การอนามัยโลก หรือ WHO พบว่า 1 ใน 3 ของประชากรที่เป็นโรคอ้วน หรือประมาณกว่า 600 ล้านคน มีอาการป่วยจากสาเหตุของโรคอ้วน เช่น เบาหวาน ความดันโลหิตสูง หลอดเลือดหัวใจ หลอดเลือดสมอง และรวมถึงมะเร็งบางชนิด ทางสหพันธ์โรคอ้วน งานวิจัยนี้จึงต้องการศึกษา แบบจำลองที่

สามารถนำมาใช้สำหรับการพยากรณ์โอกาสการเกิดโรคอ้วนของผู้ป่วยด้วยเทคนิคเหมืองข้อมูล ซึ่งผลการศึกษาครั้งนี้สามารถนำไปสร้างเป็นระบบสารสนเทศเพื่อคัดกรองผู้ป่วยโรคอ้วน

2. การทำความเข้าใจเกี่ยวกับข้อมูล (Data Understanding)

งานวิจัยนี้นำชุดข้อมูลการทำนายโรคระดับความอ้วนจากพฤติกรรมการกิน จำนวนข้อมูลทั้งหมด 2111 ชุดข้อมูล และ 17 คุณลักษณะ โดยอยู่ในรูปแบบไฟล์ CSV เพื่อนำข้อมูลไปวิเคราะห์ ซึ่งมีรายละเอียดดังตารางที่ 1

ตารางที่1 : ข้อมูลที่ใช้ในงานวิจัย

No	Attribute	Description	Values	Type
1	Gender	เพศ (M = 50.59%, F = 49.41%)	M = ชาย, F = หญิง	Categorical
2	Age	อายุ (ค่าจริง)	ค่าจริง	Continuous
3	Height	ความสูง (ค่าจริง)	ค่าจริง	Continuous
4	Weight	น้ำหนัก (ค่าจริง)	ค่าจริง	Continuous
5	family_history_with_overweight	ประวัติการมีสมาชิกครอบครัวน้ำหนักเกิน (Yes = 81.76%, No = 18.24%)	Yes = ใช่, No =ไม่ใช่	Binary
6	FAVC	การกินอาหารแคลอรี่สูง (Yes = 88.39%, No = 11.61%)	Yes = ใช่, No =ไม่ใช่	Binary
7	FCVC	ความถี่ในการกินผัก (ค่าจริง)	ค่าจริง	Integer
8	NCP	จำนวนมื้ออาหาร (ค่าจริง)	ค่าจริง	Continuous
9	CAEC	การกินอาหารระหว่างมื้ออาหาร (Always = 2.51%, Frequently = 11.46%, Sometimes = 83.61%, No = 2.42%)	Always = เป็นประจำ Frequently = บ่อยครั้ง Sometimes = บางครั้ง No = ไม่เลย	Categorical
10	SMOKE	การสูบบุหรี่ (Yes = 2.08%, No = 97.92%)	Yes = ใช่, No =ไม่ใช่	Binary
11	CH2O	การดื่มน้ำ (ค่าจริง)	ค่าจริง	Continuous
12	SCC	การติดตามแคลอรี่ทุกวัน (Yes = 4.55%, No = 95.45%)	Yes = ใช่, No =ไม่ใช่	Binary
13	FAF	การออกกำลังกาย (ค่าจริง)	ค่าจริง	Continuous
14	TUE	การใช้อุปกรณ์เทคโนโลยี (ค่าจริง)	ค่าจริง	Integer

15	CALC	การดื่มแอลกอฮอล์ (Always = 0.05%, Frequently = 3.32%, Sometimes = 66.37%, No = 30.27%)	Always = เป็นประจำ Frequently = บ่อยครั้ง Sometimes = บางครั้ง No = ไม่เลย	Categorical
16	MTRANS	บริการขนส่ง (Automobile = 21.65%, Bike = 0.33%, Motorbike = 0.52%, Public_Transportation = 74.85%, Walking = 2.65%)	Automobile = รถยนต์ Bike = รถจักรยาน Motorbike = รถมอเตอร์ไซค์ Public_Transportation = ขนส่งสาธารณะ Walking = การเดิน	Categorical
17	NobeYesdad	ระดับโรคอ้วน (Insufficient_Weight = 12.88%, Normal_Weight = 13.60%, Obesity_Type_I = 16.63%, Obesity_Type_II = 14.07%, Obesity_Type_III= 15.35%, Overweight_Level_I = 13.74%, Overweight_Level_II = 13.74%)	Insufficient_Weight = น้ำหนักไม่เพียงพอ Normal_Weight = น้ำหนักปกติ Obesity_Type_I = โรคอ้วนประเภทที่ 1 Obesity_Type_II = โรคอ้วนประเภทที่ 2 Obesity_Type_III= โรคอ้วนประเภทที่ 3 Overweight_Level_I = น้ำหนักเกินระดับที่ 1 Overweight_Level_II = น้ำหนักเกินระดับที่ 2	Categorical

3. การเตรียมข้อมูล (Data Preparation)

ขั้นตอนเตรียมข้อมูลเป็นขั้นตอนที่ทำให้เกิดความเชื่อมั่นในคุณภาพข้อมูลที่นำมาใช้ แสดงถึงความเชื่อมั่นของข้อมูลก่อนจะนำไปสร้างแบบจำลองการพยากรณ์ในครั้งนี้ โดยผู้วิจัยได้ทำการเตรียมข้อมูลกับชุดข้อมูลทั้งหมด 2 ขั้นตอนดังนี้

3.1 การคัดเลือกข้อมูล (Data Selection)

ผู้วิจัยได้ศึกษาปัจจัยที่ส่งผลทำให้เกิดการเป็นโรคอ้วน ผู้วิจัยจึงได้ทำการคัดเลือกตัวแปรจากชุดข้อมูล ดังแสดงในตารางที่ 1 จำนวน 17 คุณลักษณะ ประกอบด้วย 1. เพศ 2. อายุ 3. ความสูง 4. น้ำหนัก 5. ประวัติการมีสมาชิกครอบครัวน้ำหนักเกิน 6. การกินอาหารแคลอรี่สูง 7. ความถี่ในการกินผัก 8. จำนวนมื้ออาหาร 9. การกินอาหารระหว่างมื้ออาหาร 10. การสูบบุหรี่ 11. การดื่มน้ำ 12. การติดตามแคลอรี่ทุกวัน 13. การออกกำลังกาย

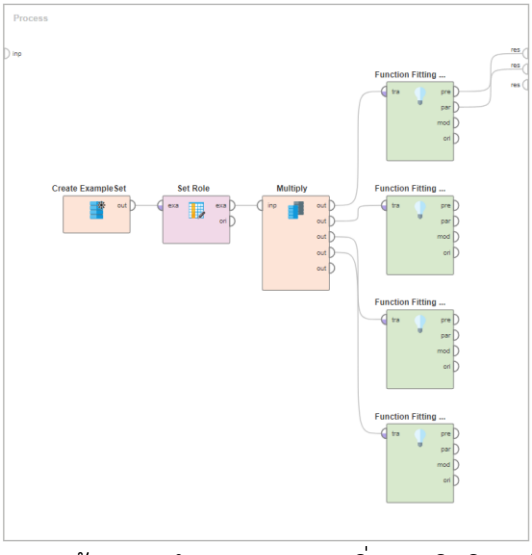
14. การใช้อุปกรณ์เทคโนโลยี 15. การดื่มแอลกอฮอล์ 16. บริการขนส่ง ซึ่งทั้ง 16 คุณลักษณะนี้จะทำหน้าที่เป็นตัวแปรอิสระ (Independent Variable)

3.2 กำหนดหน้าที่ให้กับตัวแปร

ผู้วิจัยได้กำหนดหน้าที่ให้กับคุณลักษณะที่ 17 ระดับโรคอ้วน (NObeyesdad) กำหนดหน้าที่เป็น “ป้าย” (Label) หรือตัวแปรตาม (Dependent Variable) เพื่อระบุผลลัพธ์ของการพยากรณ์ระดับความอ้วน จากพฤติกรรมการกิน และสภาพร่างกาย

4. การสร้างแบบจำลอง (Modeling)

ขั้นตอนนี้เป็นการสร้างแบบจำลอง ผู้วิจัยได้ใช้ WEKA และชุดข้อมูลเป็นข้อมูลสำหรับการประมาณระดับโรคอ้วนในบุคคลจากประเทศเม็กซิโก เปรู และโคลัมเบีย โดยพิจารณาจากพฤติกรรมการกินและสภาพร่างกายของพวกเขา งานวิจัยนี้ได้ใช้เทคนิคการทำเหมืองข้อมูลที่หลากหลายเพื่อพยากรณ์ และเปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล ซึ่งเทคนิคที่จะนำมาใช้ 5 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าสุ่ม (Random Forest) เทคนิคในกลุ่มโครงข่ายประสาทเทียม การถดถอยเชิงเส้น (linear regression) และเทคนิคนาอีฟเบย์ (Naïve Bayes) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละแบบจำลอง



ภาพที่ 7 : ขั้นตอนการสร้างแบบจำลอง และการเพิ่มประสิทธิภาพให้กับแบบจำลอง โดยใช้โปรแกรม RapidMiner Studio

หมายเหตุ.จาก https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/functions/function_fitting.html

5. การประเมินผล (Evaluation)

ผู้วิจัยทำการแบ่งชุดข้อมูลจำนวน 2 กลุ่ม ด้วยวิธีการ 10-fold cross validation โดยแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน เพื่อใช้สำหรับเป็นข้อมูลในการสอน และข้อมูลที่ใช้สำหรับการทดสอบแบบจำลอง และทำการทดสอบ ประสิทธิภาพของแบบจำลองด้วยค่าความแม่น (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) ค่าความไว (Sensitivity) ค่าจำเพาะ (Specificity) และค่าทำนายผลบวก (Positive predictive value) ดังสมการที่ 1-5 ดังนี้

5.1 ค่าความถูกต้อง (Accuracy) คือ ค่าที่แบบจำลองสามารถพยากรณ์ผู้ป่วยที่จะเกิดโรค และไม่เกิดโรคของ ข้อมูลทั้งหมดอย่างถูกต้อง ดังสมการที่ 1

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

5.2 ค่าความถ่วงดุล (F-measure) คือ ค่าที่กำเนิดจากการเปรียบเทียบโดย ค่า Precision และ ค่า Recall ในคลาสเป้าหมาย ดังสมการที่ 2-3

$$\text{F} - \text{measure คลาสเป้าหมาย YES} = \frac{(2*\text{Precision(YES)}*\text{Recall(Yes)})}{\text{Precision(YES)+Recall(Yes)}} \quad (2)$$

$$\text{F} - \text{measure คลาสเป้าหมาย NO} = \frac{(2*\text{Precision(YES)}*\text{Recall(Yes)})}{\text{Precision(YES)+Recall(Yes)}} \quad (3)$$

5.3 ค่าความระลึก (Recall) คือ ค่าที่แบบจำลองที่สามารถนำไปพยากรณ์ข้อมูลของผู้ป่วยที่เกิดโรคได้ถูกต้องต่อผู้ป่วยที่เกิดเป็นโรคจริง ดังสมการที่ 4

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

5.4 ค่าจำเพาะ (Specificity) คือ ค่าที่แบบจำลองที่สามารถพยากรณ์ข้อมูลของผู้ป่วยที่ยังไม่เกิดโรคได้ถูกต้องต่อผู้ป่วยที่พยากรณ์สาเหตุเกิดโรค ดังสมการที่ 5

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (5)$$

5.5 ค่าความแม่นยำ (Precision หรือ Positive predictive value) คือ ค่าของแบบจำลองที่ทำนายให้ถูกต้อง คำนวณ จากจำนวนข้อมูลที่ทำนายถูกในคลาสนั้น จำนวนข้อมูลทั้งหมดที่การทำนายให้ผลลัพธ์เดียวกันในคลาสนั้น ดังสมการ ที่ 6-7

$$\text{PPV ของคลาสเป้าหมาย YES} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\text{PPV ของคลาสเป้าหมาย NO} = \frac{\text{TN}}{\text{TN+FP}} \quad (7)$$

โดยที่ True Positive (TP) คือ ค่าคลาสของเป้าหมายคือ Yes และแบบพยากรณ์ว่า Yes
False Negatives (FN) คือ ค่าคลาสของเป้าหมายคือ Yes และแบบพยากรณ์ว่า No
True Negatives (TN) คือ ค่าคลาสของเป้าหมายคือ No และแบบพยากรณ์ว่า No
False Positive (FP) คือ ค่าคลาสของเป้าหมายคือ No และแบบพยากรณ์ว่า Yes

สำหรับการวิเคราะห์ข้อมูลครั้งนี้ ผู้วิจัยได้ทำการเพิ่มประสิทธิภาพของแบบจำลองการพยากรณ์โอกาสการเกิดโรคอ้วนด้วยวิธีการหาค่าที่เหมาะสมที่สุด (Optimization) ด้วยวิธีด้านวิวัฒนาการ (Evolutionary Algorithms) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละแบบจำลอง

6. การนำไปใช้งาน (Deployment)

เมื่อทำการวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) ผลการวิเคราะห์ข้อมูลพบว่าเทคนิคที่มีความเหมาะสมมากที่สุดในการสร้างแบบจำลองการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย คือ เทคนิคป่าสุ่ม (Random Forest) ซึ่งจากผลลัพธ์ของการสร้างแบบจำลองนี้สามารถนำไปใช้สำหรับการพยากรณ์โอกาสการโรคอ้วน เพื่อช่วยในการคัดกรองผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์ และการวางแผนรักษาเบื้องต้นจากแพทย์ผู้เชี่ยวชาญได้

ผลการดำเนินงาน

สำหรับผลการทดลองได้ใช้จำนวน data set จำนวนข้อมูลทั้งหมด 2111 แถว 17 คุณลักษณะ ซึ่งผู้วิจัยได้ใช้เทคนิคการทำเหมืองข้อมูลที่หลากหลายเพื่อพยากรณ์ และ เปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล โดยเทคนิคที่จะนำมาใช้ 4 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าสุ่ม (Random Forest) เทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) และเทคนิคนาอีฟเบย์ (Naïve Bayes) แบ่งระดับความอ้วน (Class) เป็น 7 ระดับ คือInsufficient Weight, Normal Weight Obesity Type I, Obesity Type II, Obesity Type III, Overweight Level I และ Overweight Level II ผลการทำนายโมเดลตัวแบบแต่ละอัลกอริทึม แสดงดังตารางที่ 2-7

ตารางที่ *2:* เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

Class	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II	Class Precision
Insufficient Weight	266	6	0	0	0	0	0	97.40%
Normal Weight	7	244	0	0	0	34	2	90.00%
Obesity Type I	0	0	336	5	0	25	8	94.40%
Obesity Type II	0	0	10	284	1	0	2	97.90%
Obesity Type III	0	0	0	1	323	0	0	99.70%
Overweight Level I	0	20	0	0	0	258	12	84.60%
Overweight Level II	0	1	10	0	0	11	268	91.80%
Class Recall	97.80%	85.00%	95.70%	95.60%	99.97%	89.00%	98.40%	

จากตารางที่ 2 ผลการทำนาย จากการใช้อัลกอริทึมเทคนิคต้นไม้ตัดสินใจ (Decision Tree) การทำนายผลตัวแบบแต่ละระดับ (Class) พบว่า ความแม่นยำ (Precision) ที่เกิดขึ้นในแต่ละระดับ (Class) คือ Insufficient Weight 97.40%, Normal Weight 90.00%, Obesity Type I 94.40%, Obesity Type II 97.90%, Obesity Type III 99.70%, Overweight Level I 84.60% และ Overweight Level II 91.80% ค่าความระลึกของข้อมูล (recall) คือ Insufficient Weight 97.80%, Normal Weight 85.00%, Obesity Type I 95.70%, Obesity Type II 95.60%, Obesity Type III 99.97%, Overweight Level I 89.00% และ Overweight Level II 98.40%

ตารางที่ 3 : เทคนิคป่าสุ่ม (Random Forest)

Class	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II	Class Precision
Insufficient Weight	258	14	0	0	0	0	0	99.60%
Normal Weight	1	271	0	0	0	13	2	84.40%
Obesity Type I	0	3	373	1	0	0	9	98.50%
Obesity Type II	0	1	2	293	1	0	0	99.30%
Obesity Type III	0	0	0	1	323	0	0	99.70%
Overweight Level I	0	26	0	0	0	261	3	93.20%
Overweight Level II	0	6	3	0	0	5	276	95.20%
Class Recall	94.90%	94.40%	96.00%	98.70%	98.70%	90.00%	95.20%	

จากตารางที่ 3 ผลการทำนาย จากการใช้อัลกอริทึมวิธีเทคนิคป่าสุ่ม (Random Forest) การทำนายผลตัวแบบแต่ละระดับ (Class) พบว่า ความแม่นยำ (Precision) ที่เกิดขึ้นในแต่ละระดับ (Class) คือ Insufficient Weight 99.60%, Normal Weight 84.40%, Obesity Type I 98.50%, Obesity Type II 99.30%, Obesity Type III 99.70%, Overweight Level I 93.20% และ Overweight Level II 95.20% ค่าความระลึกของข้อมูล (recall) คือ Insufficient Weight 94.90%, Normal Weight 94.40%, Obesity Type I 96.00%, Obesity Type II 98.70%, Obesity Type III 98.70%, Overweight Level I 90.00% และ Overweight Level II 95.20%

ตารางที่ 4 *:* เทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

Model	ค่าความถูกต้อง (Accuracy) %	ค่าความแม่นยำ (Precision) %	ค่าความระลึก (Recall) %	ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE)
17:3:7	87.67	88.80	87.70	19.51
17:4:7	82.46	88.10	82.50	19.19
17:5:7	81.04	90.10%	81.00	17.74
17:6:7	95.26	95.20	95.30	11.36
17:7:7	87.67	88.50	87.70	17.15
17:8:7	89.09	90.00	89.10	15.63
17:9:7	93.83	93.90	93.80	13.47
17:10:7	93.83	94.60	93.80	11.94

จากตารางที่ 4 ผลการสร้างแบบจำลองจากเทคนิคโครงข่ายประสาทเทียมเลือกใช้อัลกอริทึม Multilayer Perceptron ใช้การแบ่งข้อมูลฝึกสอน (Training set) ร้อยละ 90 และชุดข้อมูลทดสอบ (Testing set) ร้อยละ 10 ทำการกำหนดอัตราการเรียนรู้ (Learning Rate) ที่ 0.3 และปรับจำนวนโหนดของชั้นซ่อน (Hidden Layer) จาก 3 โหนด ไปจนถึง 10 โหนด เพื่อเปรียบเทียบค่าความถูกต้องของแบบจำลอง มีรายละเอียด ดังนี้

ตัวแบบ 17:3:7 มีค่า Accuracy ที่ 87.67%, Precision ที่ 88.80%, Recall ที่ 87.70%, และ RMSE ที่ 19.51

ตัวแบบ 17:4:7 มีค่า Accuracy ที่ 82.46%, Precision ที่ 88.10%, Recall ที่ 82.50%, และ RMSE ที่ 19.19

ตัวแบบ 17:5:7 มีค่า Accuracy ที่ 81.04%, Precision ที่ 90.10%, Recall ที่ 81.00%, และ RMSE ที่ 17.74

ตัวแบบ 17:6:7 มีค่า Accuracy ที่ 95.26%, Precision ที่ 95.20%, Recall ที่ 95.30%, และ RMSE ที่ 11.36

ตัวแบบ 17:7:7 มีค่า Accuracy ที่ 87.67%, Precision ที่ 88.50%, Recall ที่ 87.70%, และ RMSE ที่ 17.15

ตัวแบบ 17:8:7 มีค่า Accuracy ที่ 89.09%, Precision ที่ 90.00%, Recall ที่ 89.10%, และ RMSE ที่ 15.63

ตัวแบบ 17:9:7 มีค่า Accuracy ที่ 93.83%, Precision ที่ 93.90%, Recall ที่ 93.80%, และ RMSE ที่ 13.47

ตัวแบบ 17:10:7 มีค่า Accuracy ที่ 93.83%, Precision ที่ 94.60%, Recall ที่ 93.80%, และ RMSE ที่ 11.94

พบว่า ตัวแบบ 17:6:7 ค่าความถูกต้องสูงที่สุด (95.26%) คือ นำเข้าจำนวน 17 โหนด ชั้นซ่อนจำนวน 6 โหนด และขั้นแสดงผลจำนวน 8 โหนด ซึ่งใช้อัตราการเรียนรู้เท่ากับ 0.3 ได้โครงข่ายประสาทเทียม

ตารางที่ 5 : เทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ตัวแบบ 17:6:7

Class	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II	Class Precision
Insufficient Weight	19	3	0	0	0	0	0	86.40 %
Normal Weight	2	31	0	0	0	2	0	88.60 %
Obesity Type I	0	0	39	0	0	0	0	100.00 %
Obesity Type II	0	0	0	33	0	0	0	100.00 %
Obesity Type III	0	0	0	0	38	0	0	100.00 %
Overweight Level I	1	1	0	0	0	23	1	92.00 %
Overweight Level II	0	0	0	0	0	0	18	94.70 %
Class Recall	86.40 %	88.60 %	100.00 %	100.00 %	100.00 %	88.50 %	100.00 %	

จากตารางที่ 5 ผลการทำนาย จากการใช้อัลกอริทึม Artificial Neural Network: ANN) Hidden layer 6 การทำนายผลโมเดลแต่ละระดับ (Class) พบว่า ความแม่นยำ (Precision) ที่เกิดขึ้นในแต่ละระดับ (Class) คือ Insufficient Weight 86.40%, Normal Weight 88.60%, Obesity Type I 100.00 %, Obesity Type II 100.00%, Obesity Type III 100.00%, Overweight Level I 92.00% และ Overweight Level II 94.70% ส่วนค่าความระลึกของข้อมูล (recall) คือ Insufficient Weight 86.40%, Normal Weight 88.60%, Obesity Type I 100.00%, Obesity Type II 100.00%, Obesity Type III 100.00%, Overweight Level I 88.50% และ Overweight Level II 100.00%

ตารางที่ 6 *เทคนิคนาอีฟเบย์ (Naïve Bayes)*

Class	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweigh t Level II	Class Precision
Insufficient Weight	244	23	0	0	0	5	0	73.70%
Normal Weight	77	139	0	0	17	38	16	57.70%
Obesity Type I	0	1	181	62	15	25	67	55.50%
Obesity Type II	0	1	44	247	5	0	0	78.40%
Obesity Type III	0	0	0	0	324	0	0	86.60%
Overweight Level I	10	34	28	0	4	148	66	63.00%
Overweight Level II	0	43	73	6	9	19	140	48.40%
Class Recall	89.70%	48.40%	51.60%	83.20%	100%	63.00%	48.40%	

จากตารางที่ 6 ผลการทำนาย จากการใช้อัลกอริทึมเทคนิคนาอีฟเบย์ (Naïve Bayes) การทำนายผลตัวแบบแต่ละระดับ (Class) พบว่า ความแม่นยำ (Precision) ที่เกิดขึ้นในแต่ละระดับ (Class) คือ Insufficient Weight 73.70%, Normal Weight 57.70%, Obesity Type I 55.50%, Obesity Type II 78.40%, Obesity Type III 86.60%, Overweight Level I 63.00% และ Overweight Level II 48.40% ส่วนค่าความระลึกของข้อมูล (recall) คือ Insufficient Weight 89.70%, Normal Weight 48.40%, Obesity Type I 51.60%, Obesity Type II 83.20%, Obesity Type III 100%, Overweight Level I 63.00% และ Overweight Level II 48.40%

ผลการวิจัย

ตารางที่ 7 *เปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล*

ตัวแบบ (Model)	ค่าความถูกต้อง (ACCURRACY)	ค่าความแม่นยำ (PRECISION)	ค่าความระลึก (RECALL)	ค่าความถ่วงดุล (F – MEASURE)	ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE)
เทคนิคต้นไม้ตัดสินใจ (Decision Tree)	93.75 %	93.80%	93.70 %	93.80 %	12.80 %
เทคนิคป่าสุ่ม (Random Forest)	95.64 %	95.90 %	95.60%	95.70%	11.48 %
เทคนิคในกลุ่มโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)	95.26 %	95.20 %	95.30 %	95.20 %	11.36 %
เทคนิคนาอีฟเบย์ (Naïve Bayes)	67.41%	66.2%	67.40%	66.50 %	25.30 %

จากตารางที่ 7 พบว่า
ค่าความถูกต้อง (Accuracy):
คำนวณจากจำนวนทั้งหมดของการทำนายที่ถูกต้อง (True Positive + True Negative) หารด้วยจำนวนทั้งหมดของข้อมูลทดสอบ
สัดส่วนของการทำนายที่ถูกต้องทั้งหมด
สูงสุดใน Random Forest (95.64%)

ค่าความแม่นยำ (Precision):
คำนวณจากจำนวน True Positive หารด้วยผลรวมของ True Positive และ False Positive
สัดส่วนของการทำนายที่ถูกต้องที่มีความแม่นยำ
สูงสุดใน Random Forest (95.90%)

ค่าความระลึก (Recall):
คำนวณจากจำนวน True Positive หารด้วยผลรวมของ True Positive และ False Negative
สัดส่วนของข้อมูลที่ถูกต้องที่ถูกต้องทั้งหมด
สูงสุดใน Random Forest (95.60%)

ค่าความถ่วงดุล (F-Measure):
คำนวณจากค่าความแม่นยำ และความระลึก
สัดส่วนผสมระหว่างความแม่นยำ และความระลึก
สูงสุดใน Random Forest (95.70%)

ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE):
คำนวณจากความต่างระหว่างค่าทำนาย และค่าจริงของข้อมูลทดสอบ
สามารถใช้ในการวัดความคลาดเคลื่อนทั้งหมดของการทำนาย
ต่ำสุดใน Naïve Bayes (11.36%)

จากการทดสอบประสิทธิภาพเทคนิคการทำเหมืองข้อมูลทั้ง 4 อัลกอริทึม พบว่า เทคนิคป่าสุ่ม (Random Forest) ให้ค่าความถูกต้อง (Accuracy) มากที่สุด ที่ 95.64% ค่าความแม่นยำ (Precision) ที่ 95.90% ค่าความระลึก (Recall) ที่ 95.60% ค่าความถ่วงดุล (F–measure) ที่ 95.70% และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error : RMSE) 11.48 % จึงสรุปได้ว่าวิธีเทคนิคป่าสุ่ม (Random Forest) มีประสิทธิภาพในการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกายดีที่สุด ดังตารางที่ 7

สรุปและอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกายโดยใช้อัลกอริทึมเหมืองข้อมูล 4 เทคนิค ได้แก่ เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network), เทคนิคป่าสุ่ม (Random Forest), เทคนิคนาอีฟเบย์ (Naïve Bayes), และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เพื่อจับกลุ่มระดับของโรคอ้วน และ เปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation เพื่อหาตัวแบบที่เหมาะสมที่สุดในการทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกาย โดยใช้ชุดข้อมูลสำหรับการประมาณระดับโรคอ้วนในบุคคลจากประเทศเม็กซิโก เปรู และโคลัมเบีย โดยพิจารณาจากพฤติกรรมการกินและสภาพร่างกายของพวกเขา จากผลการวิจัยพบว่าเทคนิคป่าสุ่ม (Random Forest) สามารถทำนายระดับความอ้วนจากพฤติกรรมการกิน และสภาพร่างกายได้ดีที่สุด ซึ่งสอดคล้องกับงานวิจัยของ กฤตกนก และกิตติพล (2566) ที่ศึกษาเรื่องการพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล โดยปัจจัยที่ใช้ในการพยากรณ์โรคเบาหวานคล้ายคลึงกับปัจจัยในการทำนายโรคอ้วน เพราะโรคอ้วนเป็นสาเหตุทำให้เกิดอาการความผิดปกติของระบบ เมตาบอลิกในร่างกาย (Metabolic Syndrome) ซึ่งก่อให้เกิดโรคไม่ติดต่อเรื้อรัง เช่น โรคหัวใจและหลอดเลือด โรคเบาหวาน ชนิดที่ 2 เป็นต้น (สถาบันพัฒนาสุขภาวะเขตเมือง, ม.ป.ป, “องค์ความรู้การแก้ไขปัญหาโรคอ้วน”, ย่อหน้าที่ 1) มีปัจจัยดังนี้ เพศ อายุ น้ำหนัก รอบเอว ดัชนีมวลกาย ประวัติโรคประจำตัว ประวัติโรคจากครอบครัว พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา พฤติกรรมการกินอาหาร

แคลอรีสูง และจำนวนมื้ออาหารในแต่ละวัน พบว่าเทคนิคป่าสุ่ม (Random Forest) ให้ผลลัพธ์ในการพยากรณ์โรคเบาหวานที่ดีที่สุด

ข้อเสนอแนะ/งานวิจัยในอนาคต

1. **ข้อเสนอแนะในการนำผลวิจัยไปใช้**

   เนื่องจากประสิทธิภาพของแบบจำลองพยากรณ์ที่ได้จากงานวิจัยนี้มีค่าความแม่นยำ 95.90% และค่าความคลาดเคลื่อนเฉลี่ย 11.48% ดังนั้นการนำไปใช้ต้องพิจารณาถึงความคลาดเคลื่อนที่อาจเกิดขึ้น

2. **ข้อเสนอแนะในการวิจัยครั้งต่อไป**

   2.1 ควรมีการศึกษาข้อมูลปัจจัยที่มีผลต่อการการพยากรณ์ระดับความอ้วน จากข้อมูลพื้นฐานของผู้ป่วย พฤติกรรมการกิน และสภาพร่างกายเพิ่มเติม เช่น โรคประจำตัว ปัจจัยด้านความเครียด ปัจจัยด้านตัวกระตุ้นการรับประทานอาหาร (สารประสาทหลายประการ, ฮอร์โมน) ปัจจัยด้านสภาพสิ่งแวดล้อม ปัจจัยด้านสภาพเศรษฐกิจ เป็นต้น เพื่อให้ครอบคลุมกับการวินิจฉัยทางการแพทย์

   2.2 ควรใช้จำนวนข้อมูลให้มีปริมาณมากขึ้นซึ่งจะส่งผลต่อประสิทธิภาพของแบบจำลอง

เอกสารอ้างอิง

ณัฐพล แสนคำ และคณะ. (2560). ระบบสนับสนุนทางการแพทย์สำหรับคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใชเทคนิคเหมืองข้อมูล. **วารสารวิชาการโรงเรียนนายร้อยพระจุลจอมเกล้า**, ปีที่ 15, 161-170

กฤตกนก ศรีพิมพ์สอ และคณะ. (2566). การพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล**. วารสารวิชาการ “การจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม”**, ปีที่ 10 ฉบับที่ 1, 52-63

เบญจภัค จงหมื่นไวย์. (2558). การเปรียบเทียบปัจจัยโรคประจำตัวผู้สูงอายุโดยใช้อัลกอริทึมการจัดกลุ่ม J48 และ Naïve Bayes: กรณีศึกษาสาธารณสุขโพธิ์กลางนครราชสีมา. **โครงงานวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ**, ปีที่ 1 ฉบับที่ 1, 43-51

สุรวัชร ศรีเปารยะ และคณะ. (2560). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. **วิทยาศาสตร์และเทคโนโลยี**, ปีที่ 25 ฉบับที่ 5, 840-853

นงเยาว์ ในอรุณ. (2564). การเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล. **ไม่ปรากฏชื่อวารสาร**, 138-147

นพรัตน์ นนท์ศิริ และคณะ. (2564). การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล. **วิชาการพระจอมเกล้าพระนครเหนือ**, ปีที่ 33 ฉบับที่ 2, 538-547

มหาวิทยาลัยวลัยลักษณ์. (2559). **งานประชุมวิชาการระดับชาติทาง ด้านเทคโนโลยีสารสนเทศ ครั้งที่ 8**. กระบี่

bedee-expert. (2566).**โรคอ้วน สัญญาณอันตรายจุดเริ่มต้นของโรคร้ายอื่น ๆ**.สืบค้นเมื่อวันที่ 9 มกราคม

2567, BeDeebyBDMS: https://www.bedee.com/articles/gen-med/obesity

ไทยรัฐ ออนไลน์. (2563).**เมื่อ "โรคอ้วน" คุกคามคนทั้งโลก!**.สืบค้นเมื่อวันที่ 9 มกราคม 2567, ไทยรัฐ: https://www.thairath.co.th/lifestyle/woman/health/1971096

สุรีย์ ศิลาวงษ์. (2565).**"โรคอ้วน" กระทบเศรษฐกิจ 13.2% ของงบประมาณสาธารณสุขทั่วโลก**.สืบค้นเมื่อวันที่ 9 มกราคม 2567, กรุงเทพธุรกิจ: https://www.bangkokbiznews.com/social/991651

Reuters. (2566).**ภายในปี 2035 คนครึ่งโลกจะมีภาวะน้ำหนักเกิน-โรคอ้วน**.สืบค้นเมื่อวันที่ 9 มกราคม 2567, PPTV Online: https://www.pptvhd36.com/health/news/2924

สำนักสื่อสารความเสี่ยงฯ กรมควบคุมโรค. (2566).**กรมควบคุมโรค ผลักดันคนไทยใส่ใจสุขภาพ ปรับเปลี่ยนมุมมองลด "โรคอ้วน"**.สืบค้นเมื่อวันที่ 9 มกราคม 2567, สำนักสื่อสารความเสี่ยงและพัฒนาพฤติกรรมสุขภาพ: https://ddc.moph.go.th/brc/news.php?news=32470&deptcode=brc

วงการแพทย์. (2563).**ตายเพราะอ้วนมากกว่าที่คิด**.สืบค้นเมื่อวันที่ 9 มกราคม 2567,วงการแพทย์พลัสมีเดีย: https://www.wongkarnpat.com/viewya.php?id=405

สถาบันพัฒนาสุขภาวะเขตเมือง. (ม.ป.ป). **องค์ความรู้การแก้ไขปัญหาโรคอ้วน**.สืบค้นเมื่อวันที่ 3 กุมภาพันธ์ 2567, สถาบันพัฒนาสุขภาวะเขตเมือง:https://mwi.anamai.moph.go.th/webupload/migrated/files/mwi/n2930_fa4f7473c27e667e778bb12b69fe8676_article_20200508152011.pdf