การจำแนกผู้ป่วยโรคเบาหวานด้วยใช้วิธีการเรียนรู้ด้วยเครื่อง โดยการหาปัจจัยการคัดเลือกคุณลักษณะ

The classification of diabetic patients using machine learning method by feature selection

สายัณห์ เทพแดง[1*]

Sayan Tepdang[1*]

บทคัดย่อ

การจำแนกผู้ป่วยโรคเบาหวาน เป็นเรื่องที่ยากในการจำแนกเนื่องจากไม่มีปัจจัยที่แน่นอนและต้องใช้หลากหลายปัจจัยในการวินิจฉัยโรคเบาหวาน บทความวิจัยนี้มีวัตถุประสงค์ในการจำแนกผู้ป่วยว่าเป็นโรคเบาหวาน หรือไม่เป็นโรคเบาหวาน ด้วยการใช้วิธีการเรียนรู้ด้วยเครื่อง (machine learning) โดยการหาปัจจัยการคัดเลือกคุณลักษณะ (feature selection) และได้ใช้ข้อมูลการจำแนกโรคเบาหวานจากเว็บไซต์ www.kaggle.com จำนวน 536 คน ซึ่งมีการเก็บข้อมูลทั้งหมด จำนวน 8 ปัจจัย ได้แก่ จำนวนครั้งที่ตั้งครรภ์ (pregnant), กลูโคสในเลือด (glucose), การวัดความดันโลหิต (blood pressure), ความหนาของผิวหนัง (skin thickness), ระดับอินซูลินในเลือด (insulin) และดัชนีมวลกาย (BMI) พันธุกรรมที่เป็นเปอร์เซ็นต์การเป็นโรคเบาหวาน (diabetes pedigree function) และอายุ (age) ในการเกิดโรคเบาหวาน โดยใช้การเรียนรู้และการทดสอบในอัตรา 90:10, 80:20, 70:30, 60:40, 50:50 เปอร์เซ็นต์ และใช้วิธีแบ่งข้อมูลแบบ 10-fold cross validation ผลการวิจัยพบว่าวิธีที่ดีที่สุด คือ gradient boosted trees มีประสิทธิภาพอยู่ที่ 87.14 เปอร์เซ็นต์ และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 0.80 และพบว่าการหาปัจจัยการคัดเลือกคุณลักษณะ ด้วยวิธีการคัดเลือกปัจจัยโดยใช้ค่าน้ำหนักของข้อมูล (filter approach) ตามวิธีโครงสร้างต้นไม้ (decision tree) ใช้ปัจจัยจำนวนเพียง 4 ปัจจัย ได้แก่ ระดับค่ากลูโคสในเลือด อายุ จำนวนครั้งที่ตั้งครรภ์ และสุดท้ายระดับค่าอินซูลิน ซึ่งถ้ามีการจำแนกผู้ป่วยโรคเบาหวานได้อย่างดีมีประสิทธิภาพ ก็สามารถรักษาได้อย่างรวดเร็ว และทำให้ผู้ป่วยโรคเบาหวานหายป่วยและมีอายุยืนยาวมากยิ่งขึ้น

คำสำคัญ: การจำแนกโรคเบาหวาน การเรียนรู้ด้วยเครื่อง การคัดเลือกคุณลักษณะ

Abstract

The classification of types of diabetic patients is difficult because there are not only variant features, but many features needed to diagnose the symptom of diabetes. This research proposes to classify the types of patients, whether or not they are diabetic, using machine learning to find the factor for feature selection. The study was utilized

[1] คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก

[1] Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok

[*] Corresponding author. E-mail address: Sayan_te@rmutto.ac.th

Received: January 30, 2023; Revised: March 20, 2023; Accepted: April 10, 2023

data of 536 people from the website https://www.kaggle.com, that collected on 8 features causing diabetes as following; Pregnancies, glucose in blood, blood pressure, skin thickness, insulin in blood, body mass index, diabetes pedigree function, and age. By training and testing ratio of 90:10%, 80:20%, 70:30%, 60:40%, 50:50% and splitting a data set for 10-fold cross-validation, the result was showed that the optimizing method, Gradient Boosted Trees, has an efficiency at 87.14% and standard deviation at 0.80 with the best efficacy of feature selection by Filter-based factor selection method with Decision Tree of only 4 factors: glucose in blood, age, frequency of pregnancies and insulin in blood. According those of factors, the efficacy of diabetic classification would heal and cure diabetic with a speedy recovery and longer life.

Keywords: the classification of diabetic, machine learning, feature selection

บทนำ

โรคเบาหวาน (diabetes mellitus: DM) เป็นภาวะที่ร่างกายมีน้ำตาล (กลูโคส) ในเลือดสูงกว่าปกติ เนื่องจากการขาดฮอร์โมนอินซูลิน (insulin) ที่สร้างมาจากตับอ่อนไม่สามารถทำหน้าที่ในการนำน้ำตาลในเลือดเข้าไปในเซลล์ต่าง ๆ ของร่างกาย เพื่อนำไปใช้เป็นพลังงาน หากปล่อยให้ร่างกายอยู่ในสภาวะนี้เป็นเวลานานจะทำให้อวัยวะต่าง ๆ เสื่อม เกิดโรคและอาการแทรกซ้อนขึ้น (Thatoom Hospital, 2014) โดยสถานการณ์โรคเบาหวานทั่วโลกในปี 2022 มีผู้ป่วยจำนวน 537 ล้านคน และคาดว่าในปี 2030 จะมีผู้ป่วยเบาหวานเพิ่มขึ้นเป็น 643 ล้านคน และโรคเบาหวานมีส่วนทำให้เสียชีวิต สูงถึง 6.70 ล้านคน หรือเสียชีวิต 1 ราย ในทุก ๆ 5 วินาที จากรายงานสถิติสาธารณสุข กระทรวงสาธารณสุขในส่วนของประเทศไทยพบอุบัติการณ์โรคเบาหวานมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง มีผู้ป่วยรายใหม่เพิ่มขึ้น 3 แสนคนต่อปี และมีผู้ป่วยโรคเบาหวานอยู่ในระบบทะเบียน 3.30 ล้านคน ในปี 2021 มีผู้เสียชีวิตจากโรคเบาหวานทั้งหมด 16,388 คน (อัตราตาย 25.1 ต่อประชากรแสนคน) ค่าใช้จ่ายด้านสาธารณสุขในการรักษาโรคเบาหวานเฉลี่ยสูงถึง 47,596 ล้านบาทต่อปี นอกจากนี้โรคเบาหวานยังคงเป็นสาเหตุหลักที่ก่อให้เกิดโรคอื่น ๆ ในกลุ่มโรค NCDs เช่น โรคหัวใจ โรคหลอดเลือดสมอง โรคความดันโลหิตสูง และโรคไตวายเรื้อรัง (Bureau of Information Office of the Permanent Secretary, 2020)

สาเหตุของการเป็นโรคเบาหวานมีหลายปัจจัยร่วมกัน เช่น อายุ การตั้งครรภ์ ความอ้วน ความเครียด เรื้อรัง และการได้รับยาบางชนิด เป็นต้น ในส่วนคนที่มีความเสี่ยงโรคเบาหวานมีหลายปัจจัย เช่น อายุมากกว่า 45 ปีขึ้นไป มีประวัติคนในครอบครัวเป็นโรคเบาหวาน เป็นผู้มีน้ำหนักเกินหรือมีดัชนีมวลกายที่สูง ความดันโลหิตสูง (ความดันโลหิตตั้งแต่ 140/90 มิลลิเมตรปรอท) ไม่ออกกำลังกาย ดื่มสุรา สูบบุหรี่ และเคยตรวจพบระดับน้ำตาลสะสมมากกว่าหรือเท่ากับ 5.70 เปอร์เซ็นต์ เป็นต้น โดยภาวะก่อนเป็นเบาหวาน (impaired fasting glucose) จะมีระดับน้ำตาลในกระแสเลือดระหว่าง 100-125 มิลลิกรัมต่อเดซิลิตร ไขมันชนิดชนิดเอชดีแอล (HDL) น้อยกว่า 35 มิลลิกรัมต่อเดซิลิตร หรือไตรกลีเซอไรด์ (TG) มากกว่า 250 มิลลิกรัมต่อเดซิลิตร เป็นต้น (Phuket Hospital, 2022)

ในส่วนของงานวิจัยในปัจจุบันได้นำวิธีการปัญญาประดิษฐ์ (artificial intelligence: AI) มาใช้ในการวิจัย โดยพบว่าการเรียนรู้ด้วยเครื่อง (machine learning) เป็นส่วนหนึ่งของปัญญาประดิษฐ์ การเรียนรู้ด้วยเครื่องสามารถแบ่งได้ 3 ประเภทหลัก คือ supervised learning เป็นการสอนให้เรียนรู้จากการแบ่งชุดข้อมูล การระบุ input และ output ไว้อย่างชัดเจน เพื่อสร้างโมเดลต่าง ๆ เป็นต้น ในส่วนของ unsupervised learning เป็นการสอนให้เรียนรู้จากชุดข้อมูลที่ไม่มีการแบ่งกลุ่ม หรือระบุความสัมพันธ์ของข้อมูลไว้ชัดเจน และสุดท้าย reinforcement learning เป็นการสอนให้เรียนรู้จากการทดลองและพยายามค้นหาคำตอบของปัญหา เพื่อให้ได้คำตอบที่มีประสิทธิภาพดีที่สุด

ในส่วนการนำวิธีการเรียนรู้ด้วยเครื่องไปใช้ในงานวิจัยในด้านต่าง ๆ เช่น ด้านภาษาศาสตร์ (linguistics) โดยได้นำมาใช้ในการประมวลผลภาษาธรรมชาติ (natural language processing) คือ การทำให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์มากขึ้น เช่น การนำไปวิเคราะห์ข้อมูลทางด้านการแปลภาษา ข้อความ และความรู้สึกต่าง ๆ ในโซเชียลมีเดีย เป็นต้น ในส่วนของงานวิจัยที่นำไปใช้ในด้านวิศวกรรมศาสตร์ (engineering) ได้นำไปใช้ในหาค้นหาองค์ความรู้ต่าง ๆ เช่น การสร้างหุ่นยนต์ การวิเคราะห์ข้อมูลขนาดใหญ่ (big data) หรือจะเป็นลักษณะงานที่เป็น internet of things (IoT) ที่สามารถตอบสนองความต้องการของผู้ใช้งานได้ เป็นต้น และในส่วนงานวิจัยที่นำไปใช้ทางด้าน การแพทย์ หรือวิทยาศาสตร์สุขภาพ ได้นำไปใช้ในการประมวลผลภาพทางการแพทย์ (medical image processing) เช่น การนำภาพ MRI มาไปจำแนกการเกิดโรคต่าง ๆ และยังสามารถนำไปใช้ในการพยากรณ์อาการของผู้ป่วย ตลอดจนการวินิจฉัยโรคต่าง ๆ ได้อย่างถูกต้อง และแม่นยำมากขึ้น นอกจากนี้สามารถนำวิธีการเรียนรู้ด้วยเครื่องไปใช้ในงานวิจัยทางด้านอื่น ๆ ในหลากหลายด้านได้อีกด้วย

งานวิจัยที่เกี่ยวข้องการจำแนกผู้ป่วยโรคเบาหวานแบ่งได้สองประเภท ได้แก่ การเรียนรู้ด้วยเครื่องเพียงอย่างเดียว และการเรียนรู้ด้วยเครื่องโดยการหาปัจจัยการคัดเลือกคุณลักษณะ (feature selection)

ในส่วนของงานวิจัยประเภทแรกการเรียนรู้ด้วยเครื่องเพียงอย่างเดียว คือ การเรียนรู้ด้วยเครื่องในการจำแนกตามที่ได้กำหนดไว้โดยจะให้คำตอบเป็น label / class เท่านั้น งานวิจัยในกลุ่มนี้ เช่น การจำแนกประเภทโรคเบาหวานด้วยการเรียนรู้ของเครื่องและการคาดการณ์สำหรับการใช้งานด้านการดูแลสุขภาพ (Butt et al., 2021) การสำรวจเครื่องจักรอย่างครอบคลุมการเรียนรู้เทคนิคการตรวจหาโรคเบาหวาน (Sidong, Xuejiao, & Chunyan, 2018) การจำแนกประเภทของโรคเบาหวานโดยใช้ซอฟต์แวร์คอมพิวเตอร์และการเรียนรู้ด้วยเครื่อง (Saxena, Sharma, Gupta, & Sampada, 2022) และระบบการจำแนกผู้ป่วยเบาหวานด้วยเทคนิคการเรียนรู้ของเครื่อง (Rawat, & Suryakant, 2019) จากงานวิจัยประเภทแรกพบว่ามีการใช้วิธีการเรียนรู้ด้วยเครื่องในการจำแนกหลายวิธี เช่น logit boost , Naïve Bayes, deep learning, logistic regression และ support vector machine (SVM) เป็นต้น ซึ่งพบว่าข้อดีของงานวิจัยประเภทนี้ คือสามารถทำการวิจัยได้อย่างรวดเร็ว แต่ก็พบว่ามีประสิทธิภาพที่อาจจะไม่ได้ดีที่สุด

ในส่วนงานวิจัยประเภทที่สอง คือ การเรียนรู้ด้วยเครื่องโดยการหาปัจจัยการคัดเลือกคุณลักษณะเป็นการจำแนกตามที่ได้กำหนดไว้โดยจะให้คำตอบเป็น label / class เช่นกัน แต่จะพบว่ามีการหาปัจจัยการคัดเลือกคุณลักษณะด้วย ตัวอย่างเช่น ปัจจัย (feature) ที่มีจำนวนมากในการจำแนกประเภท ข้อความทัศนคติ (sentiment) ออกเป็นเชิงบวก (positive) หรือเชิงลบ (negative) นั้นจะมีคำในข้อความต่าง ๆ ที่ใช้เป็นปัจจัยจำนวนมาก ปัจจัยเหล่านี้บางอันก็ไม่ได้มีความสำคัญในการแบ่งแยกคลาส (class) ออกเป็นเชิงบวกหรือเชิงลบได้ ดังนั้น จึงจำเป็นต้องทำการคัดเลือกปัจจัยที่สำคัญหรือมีน้ำหนักมากที่สุดมาใช้งาน ขั้นตอนนี้เรียกว่าการคัดเลือกปัจจัย หรือ feature selection ซึ่งสามารถแบ่งได้เป็น 2 กลุ่มใหญ่ดังนี้ ได้แก่

filter approach เป็นการหาปัจจัยการคัดเลือกคุณลักษณะโดยใช้วิธีการคำนวณหาค่าน้ำหนักซึ่งอาจจะเป็นค่าความสัมพันธ์ระหว่างแต่ละปัจจัยในและคลาสต่าง ๆ แล้วจะเลือกปัจจัยที่มีค่าน้ำหนัก เช่น การหาค่าสหสมพันธ์ (correlation coefficient) และ การหาค่าน้ำหนักตามโครงสร้างต้นไม้ (decision tree) เป็นต้น

wrapper approach เป็นการคัดเลือกปัจจัยด้วยการสร้างโมเดล (classification model) ขึ้นมาจากเซตของปัจจัยที่กำหนดไว้และวัดประสิทธิภาพการทำงานของโมเดล และเลือกเซตของปัจจัยที่ทำให้โมเดลมีประสิทธิภาพมากที่สุดมาใช้งาน เช่น โมเดลที่ให้ค่าความถูกต้อง (accuracy) มากที่สุด โดยการคัดเลือกปัจจัยด้วยวิธีการนี้แบ่งย่อยได้เป็น 2 แบบใหญ่ ๆ คือ forward selection เป็นการสร้างโมเดลโดยการเพิ่มปัจจัยทีละ 1 ปัจจัย ถ้าปัจจัยที่ใส่เพิ่มทำให้ประสิทธิภาพที่ดี ก็จะเก็บไว้และเลือกปัจจัยอื่น ๆ มาเพิ่มต่อไปจนประสิทธิภาพของโมเดลไม่ได้ดีขึ้นก็จะหยุดทำงาน ในส่วนของ backward elimination เป็นการสร้างโมเดลที่เริ่มจากการใช้ปัจจัยทั้งหมดก่อนและตัด (eliminate) ปัจจัยที่ไม่สำคัญทิ้งไปทีละปัจจัยถ้าประสิทธิภาพดีขึ้นก็ตัดปัจจัยอื่น ๆ ต่อไป (Th.Linkedin, 2023)

งานวิจัยประเภทที่สอง การเรียนรู้ด้วยเครื่องโดยการหาปัจจัยการคัดเลือกคุณลักษณะ เช่น การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิควิธีแบบร่วมกันตัดสินใจและวิธีเลือกคุณลักษณะเด่นไปข้างหน้า (Nonsiri, Chaichitwanidchakol, & Somkantha, 2022) แนวทางใหม่สำหรับการเลือกคุณลักษณะและการจำแนกประเภทโรคเบาหวานวิธีการเรียนรู้ของเครื่อง (Saxena, Sharma, Gupta, & Sampada, 2022) และการเลือกคุณสมบัติที่สำคัญและการเปรียบเทียบความแม่นยำที่แตกต่างกันโมเดลการเรียนรู้ด้วยเครื่องสำหรับการตรวจหาเบาหวานในระยะเริ่มต้น (Rubaiat, Rahman, & Hasan, 2018) ซึ่งพบว่าข้อดีของงานวิจัยในกลุ่มนี้ คือ มีการหาปัจจัยที่มีประสิทธิภาพมากที่สุด แต่อาจจะใช้เวลานานในการหาปัจจัยที่ดีที่สุด

จากงานวิจัยส่วนใหญ่พบว่าการจำแนกผู้ป่วยโรคเบาหวาน เป็นไปในลักษณะการเรียนรู้ด้วยเครื่องเพียงอย่างเดียว เนื่องจากสามารถทำได้รวดเร็วและไม่มีความซับซ้อน โดยพบว่าข้อจำกัดของประเภทนี้ คือ ผลที่ได้อาจจะมีประสิทธิภาพที่ไม่ดีที่สุด เพราะไม่ได้หาปัจจัยที่สำคัญในการเรียนรู้ ตัวอย่างเช่น ถ้ามีคุณสมบัติหรือปัจจัยจำนวนมากในการเรียนรู้ จะทำให้เกิดการเรียนรู้ที่มากขึ้น ซึ่งพบว่าบางปัจจัยอาจจะไม่สำคัญและทำให้เกิดความปัญหาในการจำแนกและทำให้ผลประสิทธิภาพ

ลดลงได้ เป็นต้น จากข้อจำกัดดังกล่าวจากการเรียนรู้ด้วยเครื่องเพียงอย่างเดียว ผู้วิจัยได้มีวัตถุประสงค์การจำแนกผู้ป่วยว่าเป็นโรคเบาหวานหรือไม่เป็นโรคเบาหวาน ด้วยการใช้วิธีการเรียนรู้ด้วยเครื่อง โดยการหาปัจจัยการคัดเลือกคุณลักษณะเพื่อสามารถหาผลการวิจัยที่มีประสิทธิภาพดีที่สุดและลดจำนวนปัจจัยจำนวนมากในการเรียนรู้ด้วยเครื่อง โดยใช้การเรียนรู้ด้วยเครื่องได้ใช้วิธี แบบ supervised learning เป็นการสอนให้เรียนรู้จากการแบ่งชุดข้อมูลการระบุ input และ output ไว้ เพื่อสร้างโมเดลต่าง ๆ โดยจะใช้ 5 วิธี ได้แก่ การเรียนรู้โดยการใช้หลักการความน่าจะเป็น (Naïve Bayes) การเรียนรู้แบบ binary classification (logistic regression) การเรียนรู้ด้วยเครื่องแบบโครงข่ายประสาท (deep learning) การเรียนรู้แบบโครงสร้างต้นไม้ (gradient boosted trees) และสุดท้ายการเรียนรู้ด้วยเครื่องโดยการหาสัมประสิทธิ์ของสมการ (support vector machine) เนื่องจากแต่ละวิธีของการเรียนรู้ด้วยเครื่องมีข้อดีข้อเสียแตกต่างกัน และพบว่าทั้ง 5 วิธีการนี้ได้นำไปใช้ในงานวิจัยด้านต่าง ๆ ในการจำแนกและพบว่าได้ผลดีมีประสิทธิภาพในงานวิจัย ดังนั้นจึงได้นำไปใช้ในการจำแนกผู้ป่วยโรคเบาหวานต่อไป ในส่วนของการหาปัจจัยการคัดเลือกคุณลักษณะจะใช้วิธีการหาค่าน้ำหนักตามโครงสร้างต้นไม้ การวิจัยจึงเป็นเรื่องที่น่าสนใจและสามารถหาปัจจัยที่สำคัญในการจำแนกโรคเบาหวานได้อย่างมีประสิทธิภาพมากกว่าการเรียนรู้ด้วยเครื่องเพียงอย่างเดียว

วิธีการศึกษา

วิธีการดำเนินการวิจัย จะทำจากการศึกษาวิธีการและทฤษฎีต่าง ๆ ที่เกี่ยวข้องกับการจำแนกผู้ป่วยว่าเป็นโรคเบาหวาน หรือไม่เป็นโรคเบาหวาน จะประกอบไปด้วย 5 ขั้นตอน ได้แก่ dataset (ข้อมูล) data preparation (การเตรียมข้อมูล) feature selection (การหาปัจจัยการคัดเลือกคุณลักษณะ) classification (การจำแนก) และสุดท้าย (result) ผลการทดลอง ดัง (Figure 1)

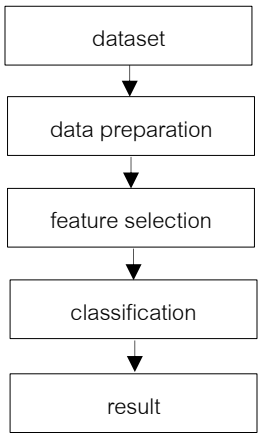

Figure 1 The research process.

ข้อมูล (dataset) คือ ข้อมูลที่มีการรวบรวมเป็นทุติยภูมิ (secondary data) ซึ่งพบว่ามีการดาวน์โหลดข้อมูลจำนวน 33,667 ครั้ง ในระยะเวลา 6 เดือนที่ผ่านมา โดยข้อมูลนำมาจากสถาบันโรคเบาหวานและระบบทางเดินอาหารและโรคไตแห่งชาติของประเทศอินเดีย โดยเฉพาะอย่างยิ่งการเก็บข้อมูลจะเป็นผู้หญิงทั้งหมดมีอายุไม่ต่ำกว่า 21 ปี มีการนำใช้ จำนวน 536 คน ข้อมูลโดยแบ่งเป็นคนที่เป็นโรคเบาหวาน จำนวน 268 คน และคนที่ไม่เป็นโรคเบาหวาน จำนวน 268 คน มีการจัดเก็บปัจจัยในการเกิดโรคมี 8 ปัจจัย ซึ่งข้อมูลมาจากเว็บไซต์ www.kaggle.com (Dataset, 2023) ดัง (Figure 2)

จาก (Figure 2) ตัวอย่างข้อมูลเป็นในลักษณะแถว (row) และ คอลัมน์ (column) และอยู่ในรูปของ CSV ไฟล์ โดยจะอธิบายการเก็บข้อมูลดัง (Table 1)

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1

Figure 2 The sample of the diabetes dataset.

Table 1 The detail of data structure.

features	detail
pregnancies	the frequency of pregnancies
glucose	the glucose level in blood
blood pressure	the blood pressure measurement
skin thickness	the thickness of the skin
insulin	the insulin level in blood
BMI	the body mass index
diabetes pedigree function	the diabetes percentage
age	the age
outcome	the final result 1 is yes and 0 is no

การเตรียมข้อมูล (data preparation) จัดเตรียมข้อมูลให้อยู่ในรูปแบบเพื่อนำไปทำการ train และ test โดยจะข้อมูลที่เป็นแบบ 10-fold cross validation

การหาปัจจัยการคัดเลือกคุณลักษณะ (feature selection) ในส่วนการหาปัจจัยการคัดเลือกคุณลักษณะเป็นแบบ filter approach เป็นการคัดเลือกปัจจัยจะใช้โดยการค่าน้ำหนักของข้อมูล ซึ่งจะใช้แบบ decision tree เป็นการหาน้ำหนักของโครงสร้างต้นไม้การตัดสินใจจากบนลงล่าง ดังสูตร (1) และ สูตร (2) การหาค่าน้ำหนักของต้นไม้

ลักษณะเรียกว่าเกนความรู้ (information gain) ตามสูตรที่ (2) จะต้องนิยามค่าหนึ่งที่ใช้บอกความไม่บริสุทธิ์ของข้อมูลก่อน เรียกว่าเอนโทรปี (entropy) โดยนิยามเอนโทรปีของต้นไม้การตัดสินใจในตัวในเซตของตัวอย่าง S คือ E (S) ตามสูตร (1) ตามด้านล่าง (Th.wikipedia, 2021)

$$E(s) = -\sum_{j=1}^{n} ps(j)\, log_{\,2\, Ps(j)} \quad (1)$$

s คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลาย ๆ กรณี

$Ps(j)$ คือ อัตราส่วนของกรณีใน S ที่ตัวแปรตามหรือผลลัพธ์มีค่า j ในส่วนของการหาค่า (information gain) และสูตร (2)

$$Gain(S, A) = E(s) - \sum_{v=value(A)} \frac{|sv|}{|s|} E(S_v) \quad (2)$$

s คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลาย ๆ กรณี

E คือ เอนโทรปีของตัวอย่าง A คือ ตัวแปรต้นที่พิจารณา

$value(A)$ คือ เซตของค่าของ A ที่เป็นไปได้ S_v คือ ตัวอย่างที่ A มีค่า v ทั้งหมด

หลังจากนั้นเมื่อได้ค่าต่าง ๆ และได้ค่าน้ำหนักของข้อมูลที่มาจากการหาโดย decision tree ซึ่งเป็นการ filter approach โดยการหาปัจจัยการคัดเลือกคุณลักษณะโดยใช้วิธีการคำนวณหาค่าน้ำหนัก หลังจากนั้นนำไปใช้ในส่วนของการหาปัจจัยแบบ wrapper approach เป็นการคัดเลือกปัจจัยในการสร้างโมเดลเป็นแบบ backward elimination โดยการใช้ค่าน้ำหนักที่หาได้นำใช้ในการตัดปัจจัยที่ไม่สำคัญออกไป

การจำแนก (classification) การจำแนกโรคเบาหวาน จากข้อมูลพบว่ามีความหลากหลาย เช่น เป็นจำนวนเต็ม เป็นจุดทศนิยม และมีหลากหลายปัจจัยในการเก็บข้อมูล และงานวิจัยก่อนหน้านี้ได้ใช้วิธีการจำแนกหลายวิธี เช่น Naïve Bayes, gradient boosted trees, deep learning, logistic regression และ support vector machine (SVM) พบว่ามีประสิทธิภาพในการจำแนกและได้ผลดีแตกต่างกัน ประกอบกับผู้วิจัยต้องการวิธีการเรียนรู้ด้วยเครื่องในการจำแนกที่แตกต่างกันด้วย ดังนั้นจึงได้ใช้ 5 วิธีในการวิเคราะห์ข้อมูลและจำแนกข้อมูลและใช้สูตรที่ 3 ในการประเมินประสิทธิภาพของโมเดล ซึ่งได้แก่ 1) Naïve Bayes เป็นเรื่องใช้หลักการของเรื่องความน่าจะเป็น (probability) ในการจำแนกแบ่งกลุ่ม (Medium, 2023) 2) logistic regression เป็นเทคนิคการจำแนกโดยอาศัยหลักการความน่าจะเป็นในการเกิดเหตุการณ์หรือจะไม่เกิดเหตุการณ์นั้น ซึ่งสามารถจัดอยู่ในประเภท binary classification ตัวอย่างการใช้งาน logistic regression ส่วนใหญ่นำไปใช้ในการจำแนกโอกาสในการเกิดโรคหรือไม่เกิดโรค (Kasetsart University, 2018) 3) deep learning คือวิธีการเรียนรู้ด้วยเครื่องแบบโครงข่าย

ประสาท โดยนำระบบโครงข่ายประสาท (neural network) มาซ้อนกันหลายชั้น (layer) และทำการเรียนรู้จากข้อมูลในการจำแนก (ABB, 2023) 4) gradient boosted trees คือโครงสร้างแบบต้นไม้ จะทำการประเมินผลแต่ละจำลองให้มีค่าประสิทธิภาพที่ดีที่สุด โดยการพยายามให้ classifier instance ที่มาใหม่แต่ละตัว มีความแม่นยำขึ้นเรื่อย ๆ โดยเรียนรู้จากค่าความคลาดเคลื่อนสะสมที่เกิดจากการจำแนกก่อนหน้า (Dhurakij Pundit University, 2023) 5) support vector machine (SVM) เป็นอัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูล (Glurgeek, 2023)

ในส่วนของโปรแกรมที่นำมาใช้ในการประมวลผล คือ โปรแกรม rapidminer โดยเลือกการตั้งค่าวิธีต่าง ๆ เป็นแบบ optimization เพื่อให้ค่าที่เหมาะที่สุดในการประมวลผล

ผลการศึกษา

การดำเนินการวิจัยในขั้นตอนแรกใช้วิธีการเรียนรู้ด้วยเครื่องจะใช้ปัจจัยทั้งหมดก่อน ซึ่งใช้จำนวน 8 ปัจจัยในการจำแนกโรคเบาหวาน ได้แก่ จำนวนการตั้งครรภ์ (pregnancies) ระดับกลูโคสในเลือด การวัดความดันโลหิต (blood pressure) ความหนาของผิวหนัง (skin thickness) ระดับอินซูลินในเลือด (insulin) ดัชนีมวลกาย (BMI) พันธุกรรมที่เป็นเปอร์เซ็นต์การเป็นโรคเบาหวาน (diabetes pedigree function) อายุ (age) ซึ่งจะใช้จำนวนเรียนรู้และการทดสอบใน อัตรา 90:10 เปอร์เซ็นต์ และใช้วิธีแบบ 10-fold cross validation โดยสูตรในการประเมินประสิทธิภาพ ตามสูตร (3) และผลการทดลอง ดัง (Table 2) โดยผลการวิจัยจาก 8 ปัจจัยทั้งหมดในการจำแนกโรคเบาหวาน

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \qquad (3)$$

ผลบวกจริง (true positive) ผู้ป่วยตรวจพบว่ามีโรคอย่างถูกต้อง

ผลบวกปลอม (false positive) คนปกติตรวจพบว่ามีโรคอย่างไม่ถูกต้อง

ผลลบจริง (true negative) คนปกติตรวจพบว่าไม่มีโรคอย่างถูกต้อง

ผลลบปลอม (false negative) คนป่วยตรวจพบว่าไม่มีโรคอย่างไม่ถูกต้อง

จาก (Table 2) ผลการวิจัยจากปัจจัยทั้งหมด 8 ปัจจัยในการจำแนกโรคเบาหวาน พบว่า gradient boosted trees มีประสิทธิภาพดีที่สุด (Acc) อยู่ที่ 86.79 เปอร์เซ็นต์ และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 0.98 หลังจากนั้นจะทำการหาปัจจัยการคัดเลือกคุณลักษณะ (feature selection) ซึ่งจะเป็นวิธีการ filter approach โดยผลการคำนวณหาค่าน้ำหนักตาม decision tree มีผลดัง (Table 3)

เมื่อสร้างโมเดลที่เริ่มจากการใช้ปัจจัยทั้งหมดจำนวน 8 ปัจจัยแล้ว จาก (Table 3) การหาค่าน้ำหนักตามโครงสร้างของต้นไม้ decision tree โดยเรียงน้ำหนักที่มากสุดตามลำดับบนลงล่าง และใช้ค่าน้ำหนักจาก (Table 3) ไปหา feature selection ซึ่งจะได้จำนวน 7 ปัจจัย, 6 ปัจจัย และลดลงเรื่อยจนถึง 3 ปัจจัย ตามลำดับ ซึ่งแสดงจำนวนปัจจัย ดัง (Table 4-5) เป็นผลการจำแนกจากการหา feature selection

Table 2 The result of all features.

classification	accuracy (Acc)	standard deviation (SD)
Naïve Bayes	77.14	10.05
logistic regression	82.14	11.01
deep learning	76.79	9.62
gradient boosted trees	86.79	0.98
support vector machine	80.00	11.18

Table 3 The feature selection for weight the decision tree.

feature selection	decision tree
glucose	0.33
age	0.11
pregnancies	0.07
insulin	0.05
diabetes pedigree function	0.05
skin thickness	0.04
BMI	0.02
blood pressure	0.01

Table 4 The result of finding the feature selection.

number of features	feature selection	it cannot be used because minimum weight (Table 3)
7	- glucose - age - pregnancies - insulin - diabetes pedigree function - skin thickness - BMI	- blood pressure

Table 4 The result of finding the feature selection (continue).

number of features	feature selection	it cannot be used because minimum weight (Table 3)
6	- glucose - age - pregnancies - insulin - diabetes pedigree function - skin thickness	- blood pressure - BMI
5	- glucose - age - pregnancies - insulin - diabetes pedigree function	- blood pressure - BMI - skin thickness
4	- glucose - age - pregnancies - insulin	- blood pressure - BMI - skin thickness - diabetes pedigree function
3	- glucose - age - pregnancies	- blood pressure - BMI - skin thickness - diabetes pedigree function

Table 5 The result of finding the classification by feature selection.

feature selection	Naïve Bayes		logistic regression		deep learning		gradient boosted trees		SVM	
	Acc	SD	Acc	SD	Acc	SD	Acc	SD	Acc	SD
7	77.14	10.05	82.14	11.01	79.29	10.61	81.79	6.24	82.14	6.56
6	82.14	6.56	81.79	6.24	82.14	6.56	84.29	5.27	81.79	6.24
5	82.14	6.56	84.29	5.27	84.64	13.63	86.79	0.98	84.29	5.27
4	82.14	6.56	84.29	5.27	81.79	6.24	87.14	0.80	76.79	6.19
3	77.14	10.05	74.29	1.60	82.14	11.01	84.64	5.45	80.00	6.85

จาก (Table 5) ผลการวิจัยพบว่าประสิทธิภาพ (Acc) มีค่าสูงสุดคือ gradient boosted trees อยู่ที่ 87.14 เปอร์เซ็นต์ และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 0.80 ในส่วนการหาปัจจัย (feature selection) ตามค่าน้ำหนักตามโครงสร้างของต้นไม้ decision tree พบว่าการใช้ปัจจัย 4 ปัจจัย ได้แก่ ระดับกลูโคสในเลือด (glucose) อายุ (age) จำนวนการตั้งครรภ์ (pregnancies) ระดับอินซูลินในเลือด (insulin) หลังจากนั้นก็นำไปดำเนินการวิจัยในรูปแบบข้อมูลเพื่อการเรียนรู้และการทดสอบในขนาด 90:10 เปอร์เซ็นต์, 80:20 เปอร์เซ็นต์, 70:30 เปอร์เซ็นต์, 60:40 เปอร์เซ็นต์ และ 50:50 เปอร์เซ็นต์ ดัง (Table 6)

Table 6 The result of the classification training: test.

train: test	Naïve Bayes		logistic regression		deep Learning		gradient boosted trees		SVM	
	Acc	SD	Acc	SD	Acc	SD	Acc	SD	Acc	SD
90:10%	82.14	6.56	84.29	5.27	81.79	6.24	87.14	0.80	76.79	6.19
80:20%	72.67	5.61	75.00	2.89	72.33	3.25	77.92	3.51	70.17	3.01
70:30%	73.72	5.99	77.19	1.88	73.04	3.64	76.32	2.24	71.90	2.80
60:40%	76.37	6.05	74.04	3.05	77.31	5.89	76.49	2.44	71.20	5.55
50:50%	68.57	4.29	71.20	2.65	71.20	2.65	71.71	3.55	60.73	3.25

จาก (Table 6) พบว่าวิธีการที่ดีที่สุดในระดับขนาดการเรียนรู้และการทดสอบ 90:10 เปอร์เซ็นต์ คือ gradient boosted trees มีค่าประสิทธิภาพ (Acc) อยู่ที่ 87.14 และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 0.80 และขนาดการเรียนรู้และการทดสอบ 80:20 เปอร์เซ็นต์ คือ gradient boosted trees มีค่าประสิทธิภาพ (Acc) อยู่ที่ 77.92 และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 3.51 ซึ่งเป็นวิธีการที่ดีที่สุด

รองลงมาคือ logistic regression ขนาดการเรียนรู้และการทดสอบ 90:10 เปอร์เซ็นต์ มีค่าประสิทธิภาพ (Acc) อยู่ที่ 84.29 และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 5.27 และขนาดการเรียนรู้และการทดสอบ 80:20 เปอร์เซ็นต์ อยู่ที่ logistic regression มีค่าประสิทธิภาพ (Acc) อยู่ที่ 75.00 และค่าเบี่ยงเบนมาตรฐาน (SD) อยู่ที่ 2.89 เป็นต้น และสุดท้ายเป็น (Table 7) การจำแนกผิดพลาด

Table 7 The errors in classification.

number	glucose	age	pregnancies	insulin	result (0)	result (1)	result for system	real
1	123	22	3	240	0.41	0.59	1	0
2	127	51	11	0	0.20	0.80	1	0
3	100	46	14	184	0.53	0.47	0	1
4	102	36	6	0	0.64	0.36	0	1

จาก (Table 7) การจำแนกการผิดพลาด พบว่าเกิดปัจจัยค่าน้ำหนักของโครงสร้างแบบต้นไม้ ดัง (Table 3) พบว่า ระดับกลูโคสในเลือด อยู่ที่ 0.33 อายุ อยู่ที่ 0.11 จำนวนการตั้งครรภ์ อยู่ที่ 0.07 ระดับอินซูลินในเลือด อยู่ที่ 0.05 จึงพบว่ามีรายการการผิดพลาดดังนี้

รายการที่ 1 มีการจำแนกผิดพลาด ซึ่งจากการจำแนกบอกว่า ไม่เป็นโรคเบาหวาน (result (0)) อยู่ที่ 0.41 เปอร์เซ็นต์ และบอกว่าเป็นโรคเบาหวาน (result (1)) อยู่ที่ 0.59 เปอร์เซ็นต์ ผลการจำแนกจากระบบ result for system บอกว่าเป็นโรคเบาหวาน (1) แต่ผลที่ถูกต้อง (real) คือ ไม่เป็นโรคเบาหวาน (0) ในส่วนที่มีผลทำให้เกิดความผิดพลาดในการจำแนก คือ มีระดับกลูโคสในเลือด 123 อยู่ในระดับเกือบสูง (ไม่ควรเกิน 125 มิลลิกรัมต่อเดซิลิตร) ระดับค่าอินซูลินในเลือด 240 อยู่ในระดับสูง แต่อายุเพียง 22 ปี ทำให้เกิดการผิดพลาดในจำแนกโรคเบาหวาน ซึ่งพบว่าข้อมูลแปรปรวนในระดับค่าอินซูลินในเลือด และอายุ เป็นต้น

รายการที่ 2 เกิดการผิดพลาดในจำแนกโรค พบว่าเกิดจาก ระดับกลูโคสในเลือด อยู่ในระดับสูงถึง 127 ระดับอายุก็สูงด้วย ทำให้เกิดการจำแนกที่ผิดพลาดว่าเป็นโรคเบาหวาน แต่โดยความเป็นจริงระดับกลูโคสในเลือด สูงมีโอกาศเป็นโรคเบาหวานสูงด้วย แต่ด้วยผลที่ถูกต้องกลับพบว่าไม่เป็นโรคเบาหวาน

รายการที่ 3 และ 4 พบว่าค่าน้ำตาลอยู่ในระดับต่ำ แต่มีปัจจัยอื่น ๆ มีความแปรปวน เช่น อายุสูง จำนวนการตั้งครรภ์มีจำนวนมาก เป็นต้น ซึ่งจะทำให้เกิดการผิดพลาดในการจำแนกโรคเบาหวาน โดยการวิจัยพบว่าจะการวินิจฉัยการเป็นโรคเบาหวานหรือไม่เป็นโรคเบาหวานขึ้นอยู่กับหลายปัจจัยในการจำแนกตลอดจนขึ้นอยู่กับปัจเจกแต่ละบุคคลและแปรปรวนของข้อมูลด้วย

อภิปรายผล

จากการวิจัยพบว่าการวินิจฉัยการเป็นโรคเบาหวานหรือไม่เป็นโรคเบาหวานขึ้นอยู่กับหลายปัจจัยในการจำแนก และการวิจัยก็พบว่าการวินิจฉัยโรคเบาหวานมีลักษณะแบบกึ่งโครงสร้าง เช่น การมีระดับค่าน้ำตาลในกระแสเลือดเกิน 125 มิลลิกรัมต่อเดซิลิตร โดย

ส่วนใหญ่พบว่าถ้ามีค่าน้ำตาลเกินก็จะเป็นโรคเบาหวาน หรืออาจจะมีปัจจัยอื่น ๆ ร่วมด้วย เช่น การมีภาวะอ้วน การมีพันธุกรรมการเป็นโรคเบาหวาน และคนที่มีอายุมากกว่า 45 ปีขึ้นไป เป็นต้น

นอกจากนี้การวิจัยพบว่าวิธีการแบบ Naïve Bayes เหมาะกับข้อมูลที่เป็นแบบเป็นข้อมูลต่อเนื่อง (continuous data) (Medium, 2023) วิธีการแบบ deep learning เหมาะสมกับการใช้ข้อมูลจำนวนมากในการเรียนรู้ (train) ซึ่งพบว่าจากข้อมูลในการจำแนกโรคเบาหวานมีข้อมูลจำนวนน้อยในการเรียนรู้ จึงทำให้ประสิทธิภาพไม่สูงมากนัก (Thai Programmer Association, 2023) และวิธีการแบบ support vector machine เหมาะสมกับข้อมูลที่เป็นลักษณะแบบอนุกรมเวลา (time series) เป็นต้น จากวิธีการวิจัยทั้งสามวิธีพบว่ามีประสิทธิภาพไม่สูงมากนักเนื่องจากพบว่าข้อมูลไม่เหมาะสมกับวิธีการในการเรียนรู้ประเภทนี้ (Tepdang, & Ponprasert, 2022) และในส่วนของ logistic regression analysis เป็นวิธีการที่ประสิทธิภาพรองลงมาจาก gradient boosted trees เนื่อง logistic regression เหมาะสมกับข้อมูลเพื่อจำแนกว่าเป็นโรคหรือไม่เป็นโรค ซึ่งถือได้ว่ามีประสิทธิภาพดีในระดับหนึ่ง (Kasetsart University, 2018) และสุดท้ายวิธีการ gradient boosted trees มีประสิทธิภาพดีที่สุด เนื่องจากพบว่าข้อมูลการเป็นโรคเบาหวานมีลักษณะแบบกึ่งโครงสร้างซึ่งพบว่าเหมาะสมกับ วิธีการ gradient boosted trees มีลักษณะการเรียนรู้แบบโครงสร้างแบบต้นไม้ โดยจากระบบการประมวลผลได้สร้างโครงสร้างแบบต้นไม้ ดัง (Figure 3) ซึ่งมีการสร้างโครงสร้างแบบต้นไม้จำนวนมากในการจำแนกโรค ด้วยวิธี gradient boosted trees (Dhurakij Pundit University, 2023)

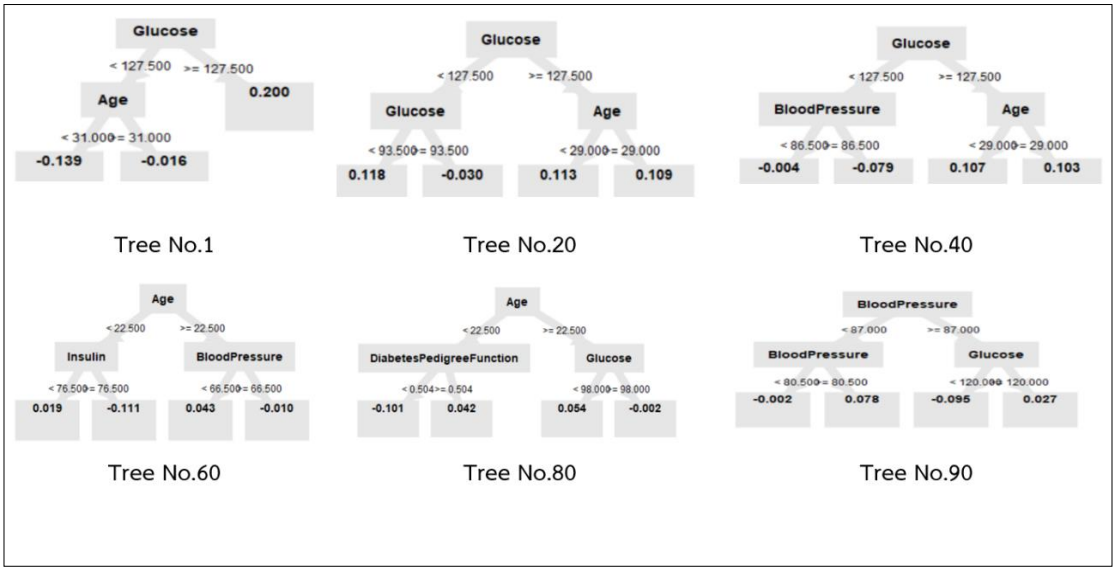

Figure 3 The tree of gradient boosted trees.

จากการวิจัยพบว่าได้สอดคล้องกับงานวิจัยในการพยากรณ์หรือการจำแนกผู้ป่วยโรคเบาหวานด้วยใช้วิธีการเรียนรู้ด้วยเครื่อง โดยพบว่าการใช้วิธีแบบ gradient boosted trees มีประสิทธิภาพสูงที่สุดเมื่อเปรียบเทียบกับวิธีการอื่น ๆ ที่เรียนรู้ด้วยเครื่อง (Lai, Huang, Keshavjee, Guergachi, & Gao, 2019) และ (Selvi, & Muthulakshmi, 2021) และในส่วนการจำแนกผู้ป่วยโรคเบาหวานด้วยใช้วิธีการเรียนรู้ด้วยเครื่องโดยการหาปัจจัยการคัดเลือกคุณลักษณะ พบว่าได้สอดคล้องกับงานวิจัย (Nagaraj, Deepalakshmi, Mansour, & Almazroa, 2021) ซึ่งพบว่าการใช้วิธีแบบ gradient boosted trees มีประสิทธิภาพดีที่สุดเช่นกัน แต่ในส่วนที่แตกต่างกันของการวิจัย คือ วิธีการหาปัจจัยการในคัดเลือกคุณลักษณะจะใช้วิธีการหาแตกต่างกัน

สรุป

จากผลการวิจัยในครั้งนี้ การจำแนกผู้ป่วยโรคเบาหวานด้วยใช้วิธีการเรียนรู้ด้วยเครื่อง โดยการหาปัจจัยการคัดเลือกคุณลักษณะพบว่าทำการเปรียบเทียบการเรียนรู้ด้วยเครื่องในการจำแนก 5 วิธี ได้แก่ Naïve Bayes, logistic regression, deep learning, gradient boosted trees, SVM โดยวิธีที่ดีที่สุดคือ gradient boosted trees และการวิจัยได้แสดงให้เห็นการจำแนกในการหาปัจจัยการคัดเลือกคุณลักษณะ (feature selection) โดยการค่าน้ำหนักของข้อมูล ซึ่งจะใช้แบบ decision tree เพื่อการหาปัจจัยซึ่งเรียงลำดับจาก 8 ปัจจัย, 7 ปัจจัย, 6 ปัจจัย และจนไปถึง 3 ปัจจัย ตามลำดับ จากผลการวิจัยพบว่าใช้เพียง 4 ปัจจัยในการจำแนกมีประสิทธิภาพดีที่สุด ได้แก่ ระดับค่ากลูโคสในเลือด อายุ จำนวนครั้งที่ตั้งครรภ์ และสุดท้ายระดับค่าอินซูลิน ซึ่งการวิจัยครั้งนี้พบว่ามีประสิทธิภาพในการจำแนกผู้ป่วยโรคเบาหวานได้ดี และยังพบอีกว่า การจำแนกผู้ป่วยโรคเบาหวานด้วยใช้วิธีการเรียนรู้ด้วยเครื่อง โดยการหาปัจจัยการคัดเลือกคุณลักษณะประสิทธิภาพดีกว่าการใช้วิธีการเรียนรู้ด้วยเครื่องเพียงอย่างเดียว อยู่ที่ 0.35 เปอร์เซ็นต์ ในส่วนของการจำแนกผิดพลาดพบว่าไม่มีกฎเกณฑ์แน่นอน และขึ้นอยู่กับปัจเจกแต่ละบุคคล ส่งผลทำให้การวินิจฉัยโรคเบาหวานผิดพลาด ดังนั้นแนวโน้มการวิจัยในอนาคต อาจจะหาวิธีการใหม่สำหรับการเรียนด้วยเครื่องและการหาปัจจัยการคัดเลือกคุณลักษณะเพื่อช่วยเพิ่มประสิทธิภาพในการจำแนกโรคให้ดีขึ้น ผลที่ได้จากการวิจัยสามารถนำไปใช้ในกลุ่มเสี่ยงเพื่อการจำแนกผู้ป่วยโรคเบาหวานได้ และถ้าสามารถจำแนกผู้ป่วยโรคเบาหวานได้อย่างรวดเร็วก็จะสามารถวางแผนการรักษา จะทำให้ผู้ป่วยสามารถรักษาหายและมีอายุยืนยาวมากขึ้น

เอกสารอ้างอิง

ABB. (2023). *Deep learning*. Retrieved 14 January 2023, from https://new.abb.com/news/detail/58004/deep-learning (in Thai)

Bureau of Information Office of the Permanent Secretary. (2020). *Diabetes mellitus*. Retrieved 14 December 2022, from https://pr.moph.go.th/?url=pr/detail/2/02/181256/ (in Thai)

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, *2021,* 9930985.

Dataset. (2023). *Ddiabetes mellitus*. Retrieved 14 December 2022, from https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

Dhurakij Pundit University. (2023). *Gradient boosted trees*. Retrieved 14 January 2023, from https://grad.dpu.ac.th/upload/content/ files/year9-3/9-30.pdf (in Thai)

Glurgeek. (2023). *Support vector machine (SVM)*. Retrieved 14 January 2023, from https://www.glurgeek.com/education/support-vector-machine/ (in Thai)

Kasetsart University. (2018). *Logistic regression*. Retrieved 14 January 2023, from https://forest-admin.forest.ku.ac.th/304xxx/?q=system/files/book/5%282018%29%20Logistic%20Regression.pdf (in Thai)

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders*, *19*, 101.

Medium. (2023). *Naïve Bayes classification*. Retrieved 14 January 2023, from https://peachapong-poolpol.medium.com/na%C3%AFve-bayes-classification-cb6cf905505d (in Thai)

Nagaraj, P., Deepalakshmi, P., Mansour, R. F., & Almazroa, A. (2021). Artificial flora algorithm-based feature selection with gradient boosted tree model for diabetes classification. *Diabetes, Metabolic Syndrome and Obesity*, *14*, 2789-2806.

Nonsiri, N., Chaichitwanidchakol, P., & Somkantha, K. (2022). Data classification for diabetes risk diagnosis using majority voting ensemble method and forward feature selection method. *Udon thani Rajabhat University Journal of Sciences and Technoogy, 10*(2), 107-122.

Phuket Hospital. (2022). *Diabetes mellitus*. Retrieved 14 December 2022, from https://www.phukethospital.com/th/news-events/diabetes/ (in Thai)

Rawat, V., & Suryakant, S. (2019). A classification system for diabetic patients with machine learning techniques. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(3), 729-744.

Rubaiat, S. Y., Rahman, M. M., & Hasan, M. K. (2018). Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. *International Conference on Innovation in Engineering and Technology* (pp.1-6). Dhaka, Bangladesh: IEEE.

Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Computational Intelligence and Neuroscience*, *2022*(Special issue), 3820360.

Selvi, R. T., & Muthulakshmi, I. (2021). Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 1717-1730.

Sidong, W., Xuejiao, Z., & Chunyan, M. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)* (pp. 291-295). Singapore: IEEE.

Tepdang, S., & Ponprasert, R. (2022). Forecasting and clustering of cassava price by machine learning (A study of Cassava prices in Thailand). *Indonesian Journal of Electrical Engineering and Informatics*, *10*(4), 825-836.

Thai Programmer Association. (2023). *Deep learning* Retrieved 14 December 2022, from https://www.thaiprogrammer.org/2018/12/deep-learning

Thatoom Hospital. (2014). *Diabetes mellitus*. Retrieved 14 December 2022, from http://www.thatoomhsp.com/ (in Thai)

Th.LinkedIn. (2023). *Feature selection*. Retrieved 14 January 2023, from https://th.linkedin.com/pulse/ (in Thai)

Th.Wikipedia. (2021). *Decision tree*. Retrieved 14 December 2022, from https://th.wikipedia.org/wiki/