



การประมาณค่า HbA1c ในผู้ป่วยเบาหวานชนิดที่ 2 ด้วยเทคนิคการเรียนรู้ด้วยเครื่อง: กรณีศึกษา
โรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยอง



โดย
นางสาวผ่องเพ็ญ คำพิว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรเกศาสตรมหาบัณฑิต
สาขาวิชาสารสนเทศศาสตร์ทางสุขภาพ แผน ก แบบ ก 2 ระดับปริญญาโทมหาบัณฑิต

มหาวิทยาลัยศิลปากร

ปีการศึกษา 2566

ลิขสิทธิ์ของมหาวิทยาลัยศิลปากร

การประมาณค่า HbA1c ในผู้ป่วยเบาหวานชนิดที่ 2 ด้วยเทคนิคการเรียนรู้ด้วยเครื่อง:
กรณีศึกษาโรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราช
กุมารี ระยอง



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรเภสัชศาสตรมหาบัณฑิต
สาขาวิชาสารสนเทศศาสตร์ทางสุขภาพ แผน ก แบบ ก 2 ระดับปริญญาโทมหาบัณฑิต
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566
ลิขสิทธิ์ของมหาวิทยาลัยศิลปากร

ESTIMATION OF HBA1C IN TYPE 2 DIABETES PATIENTS USING MACHINE
LEARNING : A CASE STUDY OF RAYONG HOSPITAL IN HONOR OF HER
ROYAL HIGHNESS PRINCESS MAHA CHAKRI SIRINDHORN



A Thesis Submitted in Partial Fulfillment of the Requirements
for Master of Pharmacy (HEALTH INFORMATICS)

Academic Year 2023

Copyright of Silpakorn University

หัวข้อ การประมาณค่า HbA1c ในผู้ป่วยเบาหวานชนิดที่ 2 ด้วยเทคนิค
การเรียนรู้ด้วยเครื่อง: กรณีศึกษาโรงพยาบาลเฉลิมพระเกียรติ
สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยอง
โดย นางสาวผ่องเพ็ญ คำผิว
สาขาวิชา สารสนเทศศาสตร์ทางสุขภาพ แผนก ก แบบ ก 2 ระดับปริญญา
มหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก สรวง รุ่งประกายพรรณ

คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรเภสัชศาสตรมหาบัณฑิต

..... คณะบดีคณะเภสัชศาสตร์
()

พิจารณาเห็นชอบโดย

..... ประธานกรรมการ
(วิรัช เลิศนที)

..... อาจารย์ที่ปรึกษาหลัก
(สรวง รุ่งประกายพรรณ)

..... ผู้ทรงคุณวุฒิภายนอก
(อัมรินทร์ พันธุ์ รอดทุกข์)

620820022 : สารสนเทศศาสตร์ทางสุขภาพ แผน ก แบบ ก 2 ระดับปริญญาโทฉบับที่ ๓

คำสำคัญ : HbA1c, machine learning, predictive model

นางสาว ผ่องเพ็ญ คำพิว: การประมาณค่า HbA1c ในผู้ป่วยเบาหวานชนิดที่ 2 ด้วยเทคนิคการเรียนรู้ด้วยเครื่อง: ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร
 ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร
 อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : สรวง รุ่งประกายพรรณ

ค่า HbA1c เป็นตัวแปรสำคัญในการติดตามผลของการใช้ยาและความร่วมมือในการรักษาของผู้ป่วยเบาหวานชนิดที่ 2 ซึ่งโดยปกติผู้ป่วยจะได้รับการตรวจปีละ 1 - 2 ครั้ง แต่ในผู้ป่วยบางกลุ่มบางรายอาจไม่สามารถมารับการตรวจตามนัดหมายได้ ทำให้ข้อมูลของค่า HbA1c ขาดหายไป ดังนั้นการศึกษานี้จึงมุ่งเน้นการประมาณค่า HbA1c ผ่านแบบจำลองที่ได้จากเทคนิคการเรียนรู้ด้วยเครื่อง โดยรวบรวมข้อมูลย้อนหลังจากเวชระเบียนอิเล็กทรอนิกส์ของโรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ราชอง ซึ่งมีผู้ป่วยทั้งหมด 5067 รายและตัวแปรที่นำมาศึกษา 23 ตัวแปร เมื่อวิเคราะห์ความสัมพันธ์ด้วยวิธี Pearson สำหรับตัวแปรชนิด Interval ขึ้นไป และ Spearman สำหรับตัวแปรชนิด Ordinal ผลการศึกษาด้วยวิธีของ Pearson พบว่าตัวแปรที่มีความสัมพันธ์กับค่า HbA1c มากที่สุดคือ ค่า FBS รองลงมาคือค่า Creatinine และ eGFR ส่วนผลการวิเคราะห์ด้วยวิธีของ Spearman พบว่าตัวแปรการได้รับยา insulin, จำนวนชนิดของยาลดระดับน้ำตาลในเลือดที่ได้รับ และเพศ เป็นตัวแปรที่มีความสัมพันธ์มากที่สุดตามลำดับ ตัวแปรที่ไม่มีความสัมพันธ์กับค่า HbA1c ที่ระดับนัยสำคัญ 0.05 คือ ค่า SBP, ค่า BUN, ระดับการศึกษา และการได้รับยา Metformin เมื่อสร้างแบบจำลองสำหรับการประมาณค่า ด้วยอัลกอริทึม 6 ชนิด คือ Multiple linear regression (MLR), Support vector machine (SVM), Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), Deep neural network (DNN), Extreme gradient boost (XG boost) และเปรียบเทียบกลุ่มตัวอย่างหลายๆ แบบพบว่าแบบจำลองชนิด Random Forest (R squared = 0.3726) และ Extreme Gradient boost (R squared = 0.3560) มีประสิทธิภาพสูงที่สุด ส่วนแบบจำลองที่มีประสิทธิภาพน้อยที่สุดคือแบบจำลอง MARS ในกลุ่มตัวอย่างทั้งหมดและแบบจำลอง DNN ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD และการกำหนดกลุ่มตัวอย่างจะช่วยลดความแปรปรวนของการประมาณค่า HbA1c ได้ โดยเมื่อทดสอบในกลุ่มผู้ป่วยที่ไม่มีโรคร่วม CKD และมีค่า HbA1c อยู่ในช่วง 6.5 - 9.9 % จะได้แบบจำลองที่มีค่าความแปรปรวนน้อยที่สุดคือค่า MAPE 8 %



620820022 : Major (HEALTH INFORMATICS)

Keyword : HbA1c, machine learning, predictive model

MISS Phongpen KAMPEW : Estimation of HbA1c in type 2 diabetes patients using machine learning : a case study of Rayong Hospital in Honor of Her Royal Highness Princess Maha Chakri Sirindhorn Thesis advisor : Suang Rungpragayphan

HbA1c is an important variable for follow up medication and adherence of type 2 diabetes mellitus patient, But sometimes patient can lose their HbA1c check up that appoint 1-2 times a year. This study focus on estimate HbA1c by using machine learning technic. Recruit 5,067 sample and 23 variables from Electronic Health Record of Rayong Hospital in Honor of Her Royal Highness Princess Maha Chakri Sirindhorn. The relationships were analyzed using Pearson's correlation for interval and higher-level variables, and Spearman's correlation for ordinal variables. The results of the Pearson's analysis showed that the variable with the strongest correlation to HbA1c was FBS, followed by Creatinine and eGFR. The Spearman's analysis indicated that the variables with the highest correlations were insulin treatment, the number of blood glucose-lowering medication types received, and gender, respectively. Variables that did not exhibit a significant correlation with HbA1c at the 0.05 level were SBP, BUN, education level, and Metformin treatment. Create predictive model using 6 algorithms, Multiple Linear Regression (MLR), Support Vector Machine (SVM), Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), Deep Neural Network (DNN) and Extreme Gradient Boost (XG boost) with many type of sample group. The highest performance was found to be Random Forest (R squared = 0.3726) and Extreme Gradient Boost (R squared = 0.3560), while the models with the lowest performance were MARS in all-type patient sample group and DNN in non-ckd patient sample group. Specification sample group can decrease prediction error, The lowest error found in non-ckd and HbA1c 6.5 - 9.9 % sample group with MAPE 8%

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.ภก. สรวง รุ่งประกายพรรณ อาจารย์ที่ปรึกษาโดยให้คำแนะนำในการดำเนินการวิจัยอย่างสม่ำเสมอ อาจารย์ช่วยกำกับประเด็นในการทดลอง ตลอดจนสนับสนุนข้อมูลทางวิชาการสำหรับงานวิจัยนี้

ขอขอบคุณ เจ้าหน้าที่จากหน่วยงานสารสนเทศ และเพื่อนร่วมงาน หน่วยงานเภสัชกรรม โรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยอง สำหรับฐานข้อมูลและความช่วยเหลือในการเรียบเรียงข้อมูลในการวิจัยนี้

ขอขอบคุณ คณะกรรมการสอบวิทยานิพนธ์ที่ได้ให้คำแนะนำและข้อเสนอแนะในการปรับปรุงการนำเสนอและเรียบเรียงใ้้งานวิจัยมีความชัดเจนมากยิ่งขึ้น

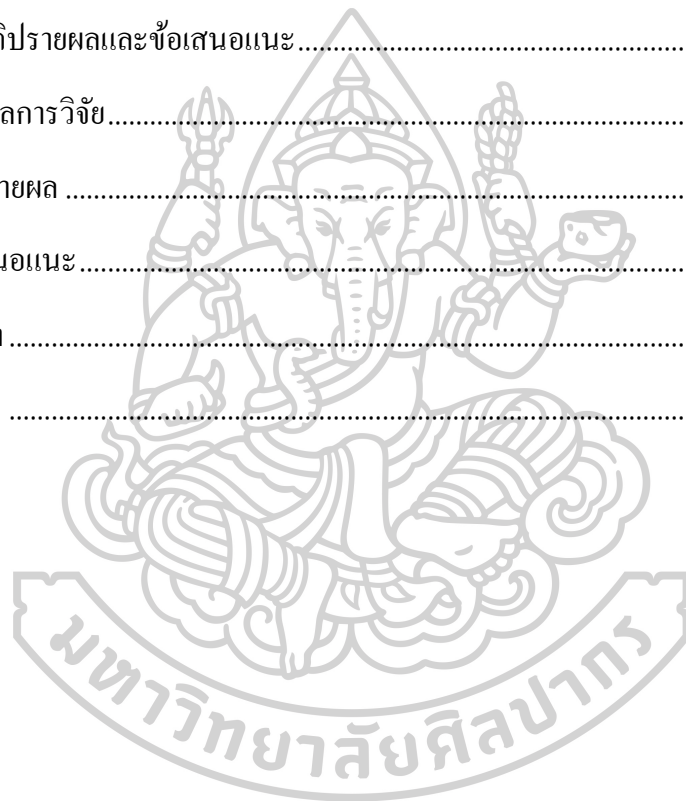
ผ่องเพ็ญ คำพิว



สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	ฉ
กิตติกรรมประกาศ.....	ช
สารบัญ.....	ช
สารบัญตาราง	ญ
สารบัญรูปภาพ	ฉ
บทที่ 1	1
บทนำ.....	1
ที่มาและความสำคัญ.....	1
วัตถุประสงค์.....	4
สมมติฐานของการศึกษาและกรอบแนวคิดในการวิจัย.....	4
ขอบเขตของการวิจัย.....	4
ประโยชน์ที่คาดว่าจะได้รับ.....	4
บทที่ 2	5
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง.....	5
งานวิจัยที่เกี่ยวข้อง.....	6
บทที่ 3	12
วิธีดำเนินการวิจัย	12
รูปแบบการวิจัย.....	12
ประชากรและกลุ่มตัวอย่าง	12
การเตรียมข้อมูล.....	12

การคัดเลือกตัวแปร.....	15
บทที่ 4	19
ผลการวิจัย.....	19
ผลการทดลอง.....	26
พัฒนาแบบจำลอง.....	28
บทที่ 5	34
สรุปผล อภิปรายผลและข้อเสนอแนะ.....	34
สรุปผลการวิจัย.....	34
อภิปรายผล	41
ข้อเสนอแนะ.....	42
รายการอ้างอิง	44
ประวัติผู้เขียน	49



สารบัญตาราง

หน้า

ตารางที่ 1 ตารางแสดง สรุปชุดข้อมูลตัวแปรที่นำมาใช้	14
ตารางที่ 2 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างทั้งหมด...	20
ตารางที่ 3 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD	21
ตารางที่ 4 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดโดยใช้ตัวแปร ทั้งหมด	26
ตารางที่ 5 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดโดยใช้ตัวแปรจากวิธี Stepwise	27
ตารางที่ 6 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c < 6.5 %	29
ตารางที่ 7 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c ≥ 10 %	30
ตารางที่ 8 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c 6.5 – 9.9 %	31
ตารางที่ 9 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดในกลุ่มตัวอย่างที่ไม่มี ผู้ป่วย CKD ในช่วงข้อมูล HbA1c 6.5 – 9.9 %	32
ตารางที่ 10 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดในกลุ่มตัวอย่างที่ไม่ มีผู้ป่วย CKD ซึ่งแบ่งเป็นกลุ่มย่อย 2 กลุ่ม ในช่วงข้อมูล HbA1C 6.5 – 9.9 %	33

สารบัญรูปภาพ

	หน้า
รูปภาพที่ 1 Flow chart แสดงกระบวนการดำเนินงานวิจัย.....	18
รูปภาพที่ 2 แผนภูมิแสดงระดับความสัมพันธ์ด้วยวิธี Pearson	22
รูปภาพที่ 3 แผนภูมิระดับความสัมพันธ์ด้วยวิธี Spearman	23
รูปภาพที่ 4 แผนภูมิแสดงค่า R squared ของแบบจำลองที่ใช้ตัวแปรทั้งหมด.....	35
รูปภาพที่ 5 แผนภูมิแสดงค่า RMSE ของแบบจำลองที่ใช้ตัวแปรทั้งหมด	36
รูปภาพที่ 6 แผนภูมิแสดงค่า MAPE ของแบบจำลองที่ใช้ตัวแปรทั้งหมด.....	37
รูปภาพที่ 7 แผนภูมิแสดงค่า R squared ของแบบจำลองที่ใช้ตัวแปร Stepwise.....	38
รูปภาพที่ 8 แผนภูมิเปรียบเทียบค่า R squared ของทั้ง 4 กลุ่มทดลอง.....	39
รูปภาพที่ 9 แผนภูมิเปรียบเทียบค่า MAPE ของทั้ง 4 กลุ่มทดลอง.....	40
รูปภาพที่ 10 แผนภูมิแสดงค่า R squared ของทั้ง 2 กลุ่มทดลอง	41



บทที่ 1

บทนำ

ที่มาและความสำคัญ

ค่าน้ำตาลเฉลี่ยสะสมหรือค่า HbA1c เป็นค่าที่แสดงถึงปริมาณน้ำตาลที่เข้ามาเกาะเคลือบที่ฮีโมโกลบิน ซึ่งเป็นกระบวนการที่ไม่สามารถย้อนกลับได้โดยจะมีค่าคงที่ตลอดอายุขัยของเม็ดเลือดแดง (ประมาณ 120 วัน) ทำให้ค่า HbA1c ถูกนำมาใช้ในการติดตามระดับน้ำตาลในเลือดในช่วงระยะเวลา 2 – 3 เดือนที่ผ่านมา โดยจะสะท้อนถึงผลการรักษาและความร่วมมือของผู้ป่วย นอกจากนี้ค่า HbA1c จะเป็นตัวแปรสำคัญในการทำนายการดำเนินไปของโรคเบาหวาน และการเกิดโรคแทรกซ้อนสำคัญได้แก่ nephropathy, retinopathy, coronary artery disease, cerebrovascular disease, infections และ lower limb amputation (K.C., 2018) สำหรับแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน พ.ศ. 2560 และข้อกำหนดของ American Diabetes Association (ADA) ได้กำหนดการวินิจฉัยโรคเบาหวานโดยใช้ค่า HbA1c เป็นหนึ่งในวิธีการที่ได้รับการรับรอง ในผู้ป่วยที่มีค่า HbA1c มากกว่าหรือเท่ากับ 6.5 % จะถูกวินิจฉัยเป็นเบาหวานและเริ่มให้การรักษา(ADA, 2022)

นอกจากนี้ค่า HbA1c ยังมีความสัมพันธ์ในเชิงบวกกับค่าน้ำตาลในเลือดแบบอื่นๆ เช่น ค่าประมาณของน้ำตาลเฉลี่ย (estimate Average Glucose: eAG) ผ่านสมการความสัมพันธ์ที่ได้รับการรับรองคือ $eAG (mg/dl) = (28.7 \times HbA1c) - 46.7, R^2=0.84$ (ADA, 2022) ค่าระดับน้ำตาลในเลือดหลังอดอาหาร (Fasting Plasma Glucose : FPG) สามารถนำมาใช้ในการคำนวณหาค่า HbA1c ผ่านสมการ $calculated HbA1c = 2.6 + 0.03 \times FPG (mg/dl)$ (Temsch et al., 2008) ค่าน้ำตาลเฉลี่ยจากการเจาะเลือดด้วยตนเอง (Self-monitored mean blood glucose) ก็มีความสัมพันธ์ในเชิงบวกกับค่า HbA1c โดยมีค่าความสัมพันธ์ (R^2) = 0.69 (Nkoana & Khine, 2020) จากความสัมพันธ์ดังกล่าวทำให้ค่า HbA1c ถูกใช้เป็นตัวแปรสำคัญในงานวิจัยโรคเบาหวาน

การเจาะเลือดหาค่า HbA1c เมื่อเทียบกับค่า FBS แล้วมีข้อดีกว่าคือผู้ป่วยไม่จำเป็นต้องอดอาหารก่อนเจาะเลือด ทำให้มีความสะดวกสำหรับผู้ป่วยและบุคลากรทางการแพทย์ แต่เมื่อเทียบค่าใช้จ่ายแล้วการตรวจหาค่า HbA1c จากห้องปฏิบัติการมีค่าใช้จ่ายมากกว่าการตรวจค่า FBS นอกจากนี้การตรวจหาค่า HbA1c ที่มีความแม่นยำสูงจำเป็นต้องตรวจในห้องปฏิบัติการที่ได้รับการรับรองมาตรฐานจาก National Glycohemoglobin Standardization Program (NGSP) และ Diabetes

Control and Complications Trial (DCCT) (สมาคมโรคเบาหวานแห่งประเทศไทย, 2017) ทั้งนี้ เพราะความแม่นยำของค่า HbA1c ขึ้นกับปัจจัยที่หลากหลาย จากงานวิจัยของ ทศนีย์ สิริธัญญ์สกุล ได้ศึกษาเปรียบเทียบค่า HbA1c ในกลุ่มผู้ป่วยเบาหวานปกติกับกลุ่มที่มีภาวะโลหิตจางร่วมด้วย พบว่าในกลุ่มผู้ป่วยเบาหวานที่มีภาวะโลหิตจาง ไม่สามารถใช้ค่า HbA1c ในการประเมินระดับน้ำตาลในเลือดที่ถูกต้องได้ (Sirithansakul, 2021) จากการศึกษาของ Michael S. Radin เกี่ยวกับ ปัจจัยที่ทำให้ค่า HbA1c จากห้องปฏิบัติการมีความคลาดเคลื่อน พบว่าค่า HbA1c จะเพิ่มสูงมากกว่าปกติในกรณีที่มีผู้ป่วยมีภาวะ Anemia (iron deficiency, B-12 or folate deficiency anemia), Asplenism, severe hypertriglyceridemia, Hyperbilirubinemia, Uremia, Chronic alcoholism, Chronic salicylate ingestion, Chronic opioid poisoning, lead poisoning และจะมีค่าลดลงกว่าปกติ ในกรณีที่มีผู้ป่วยมีภาวะ Hemolytic anemia, Blood loss, Splenomegaly, Red blood cell Transfusion, Pregnancy, Vitamin E ingestion, Vitamin C ingestion, Ribavirin or interferon-alpha use (Radin, 2014)

จากสถานการณ์การแพร่ระบาดของโคโรนาไวรัสปัจจุบัน ทำให้ทุกโรงพยาบาลต้องทำการปรับเปลี่ยนแผนการรักษาเพื่อให้สอดคล้องกับสถานการณ์ โดยอ้างอิงจาก “ระบบบริการสุขภาพวิถีใหม่” ที่แนะนำโดยกระทรวงสาธารณสุข มีการใช้ เทคโนโลยีคอมพิวเตอร์เข้ามาช่วยในหลายๆ ด้าน อาทิเช่น ระบบนัดหมายผ่านช่องทางออนไลน์ ระบบให้คำปรึกษาการแพทย์ทางไกล (Teleconsultation) โครงการรับยาร้านยาใกล้บ้าน หรือการจัดส่งยาทางไปรษณีย์ โรงพยาบาลบางแห่งได้จัดให้มีการจัดบริการหมอครอบครัวออนไลน์สำหรับผู้ป่วยโรคเรื้อรังที่ควบคุมโรคได้ดี โดยก่อนวันนัดหมาย 1-3 วัน ผู้ป่วยจะต้องวัดค่าความดันโลหิตหรือค่าระดับน้ำตาลปลายนิ้ว (DTX) ของตนเองซึ่งสามารถทำเองได้ที่บ้าน หรือโดยอาสาสมัครสาธารณสุขประจำหมู่บ้านหรือที่ร้านยาอบอุ่นที่เข้าร่วมโครงการ หลังจากนั้นผู้ป่วยจะได้ปรึกษาแพทย์ผ่านระบบออนไลน์ตามนัดหมาย (กรมควบคุมโรค, 2021) ในกรณีนี้หากสามารถทำนายค่า HbA1c ที่เป็นตัวแปรในการติดตามสภาวะของโรคเบาหวานได้ ก็จะมีประโยชน์ในการนำมาพิจารณาประกอบการรักษาของแพทย์ได้ดีมากยิ่งขึ้น

ในปัจจุบันเทคโนโลยีเกี่ยวกับปัญญาประดิษฐ์ (Artificial intelligence) และการเรียนรู้ด้วยเครื่อง (Machine learning) ได้เข้ามามีบทบาทสำคัญในงานวิจัยทางการแพทย์เนื่องจากสามารถประมวลผลข้อมูลของผู้ป่วยได้ลึกซึ้ง เหมาะสมกับข้อมูลของผู้ป่วยที่มีความซับซ้อนสูง ข้อมูลมี

หลายมิติและมีคุณลักษณะที่แตกต่างกัน รวมทั้งสามารถจัดการกับข้อมูลที่มีปริมาณสูงมากได้ โดยเฉพาะงานที่วิจัยที่เกี่ยวข้องกับโรคเบาหวาน ปัญญาประดิษฐ์ทำให้เกิดการเข้าใจพื้นฐานของโรค การดำเนินไปของโรค มีความเข้าใจในปัจจัยเสี่ยงของโรคเบาหวาน ช่วยพัฒนาการวินิจฉัยโรคและ พัฒนาการรักษาผ่านการสร้างแบบจำลองสำหรับทำนาย(Gautier et al., 2021)

การสร้างแบบจำลองสำหรับทำนายหรือการประมาณค่า (Prediction Model) มีรากฐานมาจากความสัมพันธ์ของตัวแปรต่างๆ ที่นำมาผ่านสมการทางคณิตศาสตร์และประเมินประสิทธิภาพแล้ว หากมีความแม่นยำสูงก็สามารถนำมาใช้ในการดูแลผู้ป่วยได้ จากการทบทวนวรรณกรรมของ Hema Sekhar Reddy R.และคณะในการเปรียบเทียบระหว่าง วิธีการแบบเดิม (Conventional statistical method) และการเรียนรู้ด้วยเครื่อง (Machine learning) วิธีการแบบเดิมมีข้อดีในเรื่องของการแปลผล เนื่องจากสามารถเข้าใจในความสัมพันธ์ระหว่างตัวแปรต่างๆ เป็นอย่างดี แต่เทคนิคการเรียนรู้ด้วยเครื่องเข้ามามีบทบาทมากขึ้น โดยแสดงถึงข้อดีที่เหนือกว่าวิธีการแบบเดิม อาทิเช่น สามารถประมวลผลจากข้อมูลขนาดใหญ่ได้อย่างรวดเร็วประหยัดเวลาและทรัพยากร สามารถประมวลผลข้อมูลที่จัดการได้ยากเช่นรูปภาพต่างๆ โครงสร้างยีน โครงสร้างของยา อีกทั้งข้อมูลที่ใช้ไม่จำเป็นต้องหาความสัมพันธ์ สามารถนำมาประยุกต์ใช้ในการรักษาแบบเฉพาะบุคคลได้ (Rajula et al., 2020)สำหรับการศึกษาหาค่า HbA1c จากวิธีทางคณิตศาสตร์ปัจจุบันพบว่าค่าที่คำนวณได้ (calculated HbA1c) ยังไม่มีความแม่นยำเพียงพอและยังไม่มีสมการที่เหมาะสม จากความสำคัญและประโยชน์ของค่า HbA1c ผู้วิจัยซึ่งเป็นเภสัชกรประจำโรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยองจึงมีความสนใจนำเทคนิคการเรียนรู้ด้วยเครื่องมาใช้ในการสร้างแบบจำลองสำหรับประมาณค่า HbA1c ของผู้ป่วยเบาหวานชนิดที่ 2 เพื่อนำไปประยุกต์ใช้ในการติดตามและดูแลผู้ป่วย

จากการทบทวนข้อมูลที่ผ่านมาทำให้ทราบว่าตัวแปรที่มาตรฐานสำคัญ (gold standard) ของโรคเบาหวานคือค่า HbA1c เมื่อนำมาประมวลผลผ่านการเรียนรู้ด้วยเครื่อง ทำให้เกิดความสะดวก รวดเร็ว และแม่นยำในการทำนายผลการทดลองเกี่ยวกับโรคเบาหวาน อาทิ การศึกษาผลในการควบคุมระดับน้ำตาลจากยาสูตรผสม Empagliflozin + Linagliptin เทียบกับการรักษาด้วยยาชนิดเดียวซึ่งใช้ค่า HbA1c และค่า FBS เป็นตัวแปร โดยใช้แบบจำลองสำหรับทำนายจากการเรียนรู้ด้วยเครื่องชนิด Random forest (Del Parigi et al., 2019)การทำนายการเกิด Hypoglycemia โดยใช้แบบจำลองสำหรับทำนายจากการเรียนรู้ด้วยเครื่อง พบว่ามีแบบจำลองที่ได้มีความถูกต้องและ

แม่นยำสูงในการทำนายการเกิดภาวะ Hypoglycemia (Kodama et al., 2021) นอกจากนี้ยังมี การศึกษาสร้างแบบจำลองสำหรับทำนายการเกิดโรคแทรกซ้อนต่างๆ ได้แก่ Diabetes retinopathy, neuropathy, nephropathy (Dagliati et al., 2018) metabolic syndrome, dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity, retinopathy (Jian et al., 2021)

วัตถุประสงค์

เพื่อพัฒนาแบบจำลองจากเทคนิคการเรียนรู้ด้วยเครื่องสำหรับประมาณค่า HbA1c ในผู้ป่วย เบาหวาน ชนิดที่ 2 ณ โรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราช กุมารี ระยอง

สมมติฐานของการศึกษาและกรอบแนวคิดในการวิจัย

แบบจำลองสำหรับประมาณค่า HbA1c ที่พัฒนาขึ้นมีประสิทธิภาพสูง

ขอบเขตของการวิจัย

การวิจัยนี้เป็นการพัฒนาแบบจำลองสำหรับประมาณค่า HbA1c โดยใช้การเรียนรู้ด้วย เครื่องแบบ Supervised โดยใช้ข้อมูลจากเวชระเบียนของผู้ป่วยเบาหวานชนิดที่ 2 ที่มารับบริการที่ คลินิกเบาหวาน ณ โรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยอง ตั้งแต่ตุลาคม 2558 - กันยายน 2564

ประโยชน์ที่คาดว่าจะได้รับ

ได้แบบจำลองสำหรับการประมาณค่า HbA1c ที่น่าเชื่อถือ สามารถนำมาใช้ประโยชน์ได้ จริง สามารถนำมาประกอบการตัดสินใจในการรักษาของแพทย์ได้

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การเรียนรู้ด้วยเครื่อง (Machine Learning) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence) สามารถสร้างอัลกอริทึมในการเรียนรู้ข้อมูลและทำนายผลข้อมูลได้ โดยการใส่ข้อมูลตัวแปรที่สนใจ จากนั้นอัลกอริทึมจะทำงานสร้างแบบจำลองขึ้นมา ยิ่งข้อมูลตัวแปรมีความเกี่ยวข้องกันมากแบบจำลองที่ได้ก็จะสามารถทำนายผลได้อย่างมีประสิทธิภาพ โดยทั่วไปการเรียนรู้ด้วยเครื่องแบ่งออกเป็น 3 ประเภท คือ

1. Supervised Learning (การเรียนรู้แบบมีผู้สอน) เป็นการให้คอมพิวเตอร์ประมวลผลหาคำตอบจากข้อมูลตัวแปรที่กำหนดให้เป็นชุดสอน (Training set data) ซึ่งประกอบด้วยข้อมูลตัวแปรอิสระหรือตัวแปรต้นและข้อมูลตัวแปรตาม (ผลลัพธ์ที่ต้องการ) จากนั้นคอมพิวเตอร์จะสร้างแบบจำลองตามอัลกอริทึมที่กำหนดไว้ และจะเรียนรู้ไปเรื่อย ๆ จนกว่าจะได้แบบจำลองที่มีความแม่นยำมากที่สุด
2. Unsupervised Learning (การเรียนรู้แบบไม่มีผู้สอน) ตรงข้ามกับ Supervised Learning โดยที่จะใส่ข้อมูลตัวแปรที่สนใจทั้งหมด ไม่มีการกำหนดกลุ่มเป้าหมายในการทำนาย เพราะคอมพิวเตอร์จะประมวลผล เน้นที่การจำแนกข้อมูล ตามความคล้ายคลึงของตัวแปร
3. Reinforcement Learning (การเรียนรู้แบบเสริมกำลัง) เป็นการเรียนรู้ที่เครื่องสามารถประมวลผลและตัดสินใจเองได้ โดยอาศัยสิ่งแวดล้อม หรือสิ่งรอบๆตัวที่เปลี่ยนไปทุกๆการตัดสินใจ โดยมีการลองผิดลองถูก เก็บประสบการณ์ แล้วหาทางออกที่ดีที่สุด

การวิเคราะห์ข้อมูลด้วยสมการถดถอย (Regression Analysis) เป็นการศึกษาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยทั่วไปสามารถแบ่งสมการถดถอยออกเป็น 2 ประเภทคือ Linear regression หมายถึงความสัมพันธ์ของตัวแปรมีลักษณะเป็นเชิงเส้นและ Non-linear regression หมายถึงตัวแปรที่ศึกษามีความสัมพันธ์กันแต่ไม่สามารถแสดงออกมาในรูปแบบเชิงเส้น

ตัวชี้วัดคุณภาพของแบบจำลองชนิด Regression ที่นำมาใช้ในงานวิจัยนี้ ได้แก่

1. ค่าเฉลี่ยของรากที่สองของกำลังสองของความคลาดเคลื่อน Root Mean Squared Error (RMSE)

2. ค่าเฉลี่ยของค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน Mean Absolute Percentage Error (MAPE)

3. ค่าสัมประสิทธิ์สหสัมพันธ์ R-Squared

งานวิจัยที่เกี่ยวข้อง

จากบทความของ Kushan De Silva และคณะ ได้ทบทวนงานวิจัยที่ใช้แบบจำลองสำหรับทำนายจากการเรียนรู้ด้วยเครื่องในกลุ่มโรคเบาหวานชนิดที่ 2 จำนวน 23 งานวิจัยจัดเป็นแบบจำลองสำหรับทำนายการเกิดโรคเบาหวาน 15 งานวิจัย และแบบจำลองสำหรับทำนายการดำเนินไปของโรค 8 งานวิจัย ในงานวิจัยที่มีการนำค่า HbA1c มาประกอบการวิเคราะห์พบว่า ค่า HbA1c จะเป็นปัจจัยสำคัญสำหรับแบบจำลองเสมอ (Silva et al., 2020)

งานวิจัยของ Che Ngufor และคณะ ได้ศึกษาการเปลี่ยนแปลงของค่า HbA1c ในผู้ป่วยเบาหวานที่ควบคุมเบาหวานได้ดี เพื่อทำนายความเสี่ยงที่ผู้ป่วยจะไม่สามารถควบคุมเบาหวานได้อีก พบว่าการใช้ algorithm แบบ non-linear จะมีความเหมาะสมกับข้อมูลมากที่สุด (Ngufor et al., 2019)

Tadao Ooka และคณะ ได้ศึกษาการสร้างแบบจำลองสำหรับทำนายจากการเปลี่ยนแปลงของค่า HbA1c ในกลุ่มตัวอย่างที่มีค่า HbA1c < 6.5% จัดอยู่ในภาวะ Prediabetes จำนวน 42,908 คน และมีตัวแปรทั้งหมด 51 ตัวแปร โดยเปรียบเทียบระหว่างการเรียนรู้ด้วยเครื่องชนิด Random forest และ multivariate logistic regression (MLR) พบว่า Random forest ให้ค่าการทำนายที่มากกว่า MLR ซึ่งแบบจำลองสำหรับทำนายที่ได้จากการเรียนรู้ด้วยเครื่องมีความแม่นยำสูงและสามารถนำมาปรับใช้กับผู้ป่วยเบาหวานได้ (Ooka et al., 2021)

Maria C Kjellsson และคณะ ได้ศึกษาการสร้างแบบจำลองสำหรับทำนายค่า HbA1c ที่เปลี่ยนแปลงจากผลของยาโดยใช้ข้อมูลเดิมของผู้ป่วย โดยเก็บข้อมูลของระดับกลูโคสและอินซูลินในร่างกายเป็นระยะเวลา 1 สัปดาห์เพื่อนำมาสร้างแบบจำลองในการทำนายค่า HbA1c ในอีก 12 สัปดาห์ต่อมา โดยใช้เทคนิค semi-mechanistic model-based พบว่าในสัปดาห์ที่ 8 และ 12 มีค่ามากกว่าค่าที่ได้จากการสังเกต (Kjellsson et al., 2013)

การศึกษาของ Rui Lu และคณะ ในการประยุกต์ใช้ multivariate adaptive regression splines (MARS) models ในการหาตัวแปรที่มีผลกระทบและสามารถทำนายแนวโน้มของค่า HbA1c ได้ โดยในการศึกษานี้ใช้กลุ่มผู้ป่วย Type 2 Diabetes Mellitus ทั้งหมด 6,462 รายและตัวแปรทั้งหมด

13 ตัวแปร ได้แก่ อายุ, เพศ, ประวัติการเป็นเบาหวานของคนในครอบครัว, การออกกำลังกาย, น้ำหนัก, ส่วนสูง, ระยะเวลาป่วยเป็นเบาหวาน, ขนาดของยาอินซูลิน, จำนวนยาเบาหวานที่ใช้, ภาวะโรคร่วม, ค่า HbA1c, ค่าน้ำตาล 2hPBG และ โรคความดันโลหิตสูง โดยศึกษาเปรียบเทียบระหว่าง MARS model และ Logistic regression model ในการทำนายแนวโน้มของค่า HbA1cว่าจะอยู่ในกลุ่มที่มีพัฒนาการ ($HbA1c < 7\%$) หรือกลุ่มที่ไม่มีพัฒนาการ ($HbA1c \geq 7\%$) โดยใช้ค่า Accuracy, Youden's index, Recall rate, G-mean และ AUC พบว่า MARS model สามารถทำนายได้ถูกต้องมากกว่า(Lu et al., 2021)

การศึกษาของ Sunil B. Nagaraj และคณะ ได้ใช้เทคนิคการเรียนรู้ด้วยเครื่องมาสร้างสมการสำหรับทำนายการตอบสนองต่อการรักษาผ่านค่า HbA1c หลังจากได้รับการรักษาด้วยอินซูลิน ทั้งในระยะสั้น (6 เดือน) และระยะยาว (24 เดือน) โดยผู้ป่วยที่ตอบสนองดีคือ ค่า HbA1c ลดลง $\geq 5\%$ หรือ ค่า HbA1c $\leq 53\text{mmol/mol}$ หากไม่เข้าเกณฑ์จะจัดว่าเป็นผู้ป่วยที่ตอบสนองไม่ดี ใช้ข้อมูลจากฐานข้อมูล GIANNT โดยเป็นผู้ป่วยที่เริ่มใช้อินซูลินครั้งแรก และได้รับการตรวจหาค่า HbA1c อย่างน้อยปีละ 1 ครั้ง ใช้ตัวแปรทั้งหมด 24 ตัวแปร จากนั้นนำมาผ่านสมการ 3 แบบ คือ 1) Elastic net regularization-based generalized linear model 2) Support vector machine 3) Random forests ทำการตรวจสอบความถูกต้องด้วยเทคนิค leave-one-subject-out (LOSO) cross-validation technique พบว่า Elastic net regularization-based generalized linear model สามารถทำนายการตอบสนองต่อการรักษาได้ถูกต้องแม่นยำมากที่สุด โดยเฉพาะเมื่อใช้ตัวแปร eGFR ร่วมด้วยจะสามารถทำนายได้ถูกต้องทั้งในระยะสั้น (AUCs = 0.80) และระยะยาว (AUCs = 0.81)(Nagaraj et al., 2019)

จากการศึกษาของ Yuting Fan และคณะในการใช้ การเรียนรู้ด้วยเครื่องในการทำนายการเกิดภาวะแทรกซ้อนและระดับน้ำตาลในเลือดของผู้ป่วยเบาหวานที่ไม่ให้ความร่วมมือในการรักษา โดยคัดเลือกจากผู้ป่วยที่ไม่ได้รับการตรวจวัดระดับน้ำตาลในเลือดและไม่มีการเปลี่ยนแปลงการรักษา จำนวน 165 คนมีตัวแปรต้น 9 ชนิดที่นำมาใช้คือ อายุ, ระยะเวลาการเป็นเบาหวานมานานกว่า 1 ปี, ระยะเวลาที่ไม่ได้รับการปรับยาเบาหวานมานานกว่า 1 ปี, จำนวนชนิดของยาอินซูลิน, ราคาของยาเบาหวาน, จำนวนชนิดของยาเบาหวาน, เพศ, ประวัติทางพันธุกรรมของโรคเบาหวานในครอบครัวและภาวะไขมันในเลือด ตัวแปรตาม 5 ชนิดคือ ระยะเวลาในการเกิดโรค Diabetes peripheral neuropathy, Diabetes angiopathy, diabetes eye disease, Diabetes nephropathy และการ

ควบคุมระดับน้ำตาลในเลือด (ค่า HbA1c) การทำนายใช้ machine learning algorithm 8 ชนิด สร้างเป็น model ทั้งหมด 18 model พบว่ามี 16 model ที่เหมาะสมกับการทำนายการเกิดโรคแทรกซ้อน และ 2 ชนิดเหมาะสำหรับการทำนายค่า HbA1c นั่นคือ Artificial neural network (AUC = 0.604) และ Bayesian network (AUC = 0.825)(Fan et al., 2021)

Machine learning หรือการเรียนรู้ด้วยเครื่องจัดเป็นเครื่องมือในการประมวลผลข้อมูล ทั้งในด้านจัดทำเหมืองข้อมูล การวิเคราะห์ทางสถิติ และการสร้างแบบจำลองต่างๆ จากบทความของ Luis Fregoso-Aparicio และคณะมุ่งเน้นหาแนวทางในการแก้ไขจุดอ่อนของการสร้างแบบจำลองสำหรับทำนายจากการเรียนรู้ด้วยเครื่อง ประการแรกคือการหาเทคนิคที่เหมาะสมในการสร้าง model และประการที่สองคือการขาดความเข้าใจเกี่ยวกับตัวแปรทำให้แปลผลได้ไม่ดี ซึ่งสามารถหาพิจารณาได้จากคำถามสำคัญ 3 อย่าง คือ

คำถามแรกคือ มีข้อมูลลักษณะไหนบ้างที่นำมาใช้ในการสร้างแบบจำลองพบว่า ข้อมูลสามารถเลือกได้กว้างจะมากหรือน้อยขึ้นอยู่กับเทคนิคการสร้างและความซับซ้อน นอกจากนี้ยังพบว่าข้อมูลเกี่ยวกับ lifestyle, socioeconomic และ diagnostic สามารถนำมาใช้เพื่อเพิ่มประสิทธิภาพของแบบจำลองได้ จึงไม่ควรใช้แค่เวชระเบียนเพียงอย่างเดียว

คำถามที่สอง คือเทคนิคแบบไหนที่เหมาะสมสำหรับการสร้างแบบจำลองสำหรับทำนายของผู้ป่วยเบาหวานชนิดที่ 2 โดยการเปรียบเทียบค่า AUC และ Accuracy model ที่มีค่าทั้งสองอย่างสูงสามารถนำมาพิจารณาใช้ได้

คำถามที่สามคือการ validate ที่เหมาะสมควรมีอะไรบ้าง ซึ่งโดยทั่วไปจะใช้ค่า Accuracy, Specificity, Sensitivity และ AUC เป็นหลักแต่สามารถพิจารณาค่า F1-score, precision และ MCC score มาเพิ่มในการทดสอบประสิทธิภาพได้

โดยสรุปแบบจำลองที่มีประสิทธิภาพมากที่สุด คือแบบจำลองที่ได้จากการเรียนรู้ด้วยเครื่องชนิด Deep Neural Network ซึ่งสามารถจัดการกับข้อมูลขนาดใหญ่และไม่เป็นระเบียบได้อย่างดี นอกจากนี้การใช้เทคนิค Balancing data และ Feature selection ยังสามารถเพิ่มประสิทธิภาพของแบบจำลองได้(Fregoso-Aparicio et al., 2021)

จากบทความของ Kushan De Silva และคณะ ในการเปรียบเทียบประสิทธิภาพของแบบจำลองสำหรับทำนายโดยมุ่งเน้นผลวิเคราะห์ขั้นต้น ได้แก่ Model discrimination ability

(ความสามารถในการจำแนกของแบบจำลอง) โดยใช้ค่า c-index พบว่ามีงานวิจัยที่อยู่ในเกณฑ์ดี มาก 6 model (c-index 0.9 – 1), เกณฑ์ดี 20 model (c-index 0.8 – 0.9), เกณฑ์ปานกลาง 10 model (c-index 0.7 – 0.8) ซึ่งค่า discrimination ability ที่สูงแสดงถึงความเป็นไปได้ในการนำแบบจำลอง มาใช้และง่ายต่อการแปลผลลัพธ์ Model calibration (การสอบเทียบแบบจำลอง) ในการทบทวนวรรณกรรมนี้พบเพียง 7 งานวิจัยเท่านั้นที่มีการสอบเทียบโดยใช้วิธีทางสถิติ Classification metrics (การทดสอบประสิทธิภาพของแบบจำลอง) พบว่าในงานวิจัยส่วนมากจะใช้ค่า Accuracy, Specificity และ Sensitivity ส่วนการสอบเทียบแบบจำลองพบว่าการสอบเทียบใช้เพียงข้อมูลภายในงานวิจัยของตนเองเท่านั้น และแบบจำลองจากงานวิจัยต่างๆ จะนำมาใช้เพื่อการคัดกรองหาผู้ป่วยเบาหวาน และมีเพียง 4 งานวิจัยเท่านั้นที่มีแนวโน้มที่ดีที่จะนำมาใช้ประกอบการตัดสินใจในการรักษา การทบทวนวรรณกรรมชั้นนี้ยังมีการประยุกต์ใช้เครื่องมือในการจำแนก Risk of bias ในงานวิจัยต่างๆ โดยใช้เกณฑ์ของ PROBAST จำแนกเป็น low, moderate, high risk of bias งานวิจัยที่ได้ผลเป็น low risk of bias จะได้แบบจำลองสำหรับทำนายที่มีประสิทธิภาพสูง(Silva et al., 2020)

จากบทความของ Jyotisma Chaki และคณะ ได้รวบรวมงานวิจัยที่มีการสร้างแบบจำลองจากการเรียนรู้ด้วยเครื่องในกลุ่มผู้ป่วยเบาหวานจาก Scopus และ Pubmed โดยนำมาวิเคราะห์และสามารถนำมาใช้อ้างอิงในการสร้างแบบจำลองได้ โดยต้องพิจารณาเกี่ยวกับ Dataset ในงานวิจัยส่วนมากนำมาใช้คือ Self-created dataset ทั้งนี้เพื่อให้เหมาะสมกับกลุ่มเป้าหมายที่ต้องการแต่แบบจำลองที่ได้อาจไม่สามารถนำมาปรับใช้ในฐานข้อมูลที่ใหญ่ขึ้น ดังนั้นควรมีการรวบรวมข้อมูลจากหลายๆแห่งให้มีข้อมูลที่มากและหลากหลาย แบบจำลองที่ได้ก็就会有ความน่าเชื่อถือมากขึ้น ขั้นตอนในกระบวนการ Pre-processing methods ที่ควรนำมาพิจารณาในการเตรียมข้อมูลได้แก่ Replace the missing values and irrelevant values, Green Channel Extraction, Adaptive Thresholding, Discrete Wavelet Transform, Outlier Detection, Illumination Correction, Resize, Augmentation, Data Normalization, Grayscale Conversion, ROI selection, Histogram equalization ควรมีการจัดทำ Feature Selection and Analysis เพราะข้อมูลที่ถูกจัดการดี ทำให้ได้ผลลัพธ์ที่แม่นยำมากขึ้นจึงควรพิจารณาในด้านของ overfitting, accuracy และ processing time ควรตรวจสอบคุณลักษณะของข้อมูลที่ส่งผลต่อการจำแนกเพื่อให้การจัดกลุ่มมีความแม่นยำมากยิ่งขึ้น ในขั้นตอนการสร้างแบบจำลองควรใช้แบบ supervised เพราะมีความน่าเชื่อถือมากกว่า unsupervised จากการทบทวนพบว่าในการสร้างแบบจำลอง การใช้ algorithm ชนิด Deep Neural Network และ Support Vector Machine จะให้ผลลัพธ์ในการจำแนกที่ดีกว่า รองลงมาคือ Random

Forest และ Ensemble Classifier จากนั้นควรทดสอบประสิทธิภาพของแบบจำลองโดยหาค่า Accuracy, Specificity, Sensitivity และค่า AUC ข้อเสนอแนะอื่นๆ ประการแรกในการสร้างแบบจำลองอาจทำให้เกิดการเรียนรู้โครงสร้างของแบบจำลองในเชิงลึกและเมื่อเข้าใจอย่างลึกซึ้งสามารถนำมาประยุกต์ในการปรับแบบจำลอง ให้มีความเฉพาะกับข้อมูลที่มีมากยิ่งขึ้น ประการที่สองการใช้ข้อมูลที่เพียงพอทำให้ model มีประสิทธิภาพสูงขึ้น ประการที่สามสามารถพิจารณาในการใช้ deep learning predictive model, telemedicine และ cloud computing สำหรับให้บริการผู้ป่วยที่อยู่ในพื้นที่ห่างไกลได้(Chaki J., 2020)

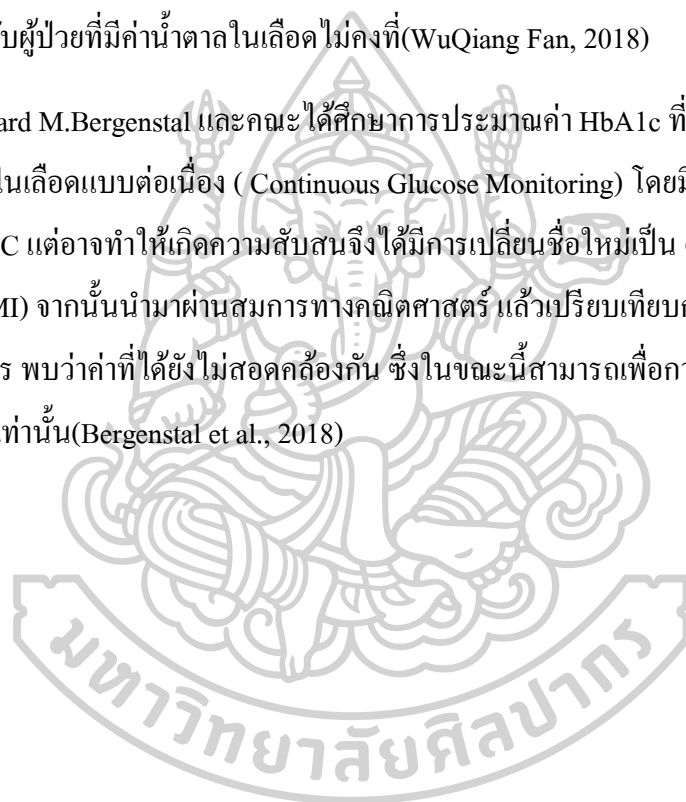
Shruthi Soundara Rajan และคณะได้ศึกษาเปรียบเทียบค่า HbA1c จากวิธี Nephelometry เทียบกับ calculated HbA1c จากสูตร $HbA1c = 2.6 + 0.03 \times FBS \text{ (mg/dl)}$ ในผู้ป่วยเบาหวานและคนปกติโดยใช้สถิติ paired t-test พบว่าไม่มีความแตกต่างอย่างมีนัยสำคัญแต่เมื่อนำมาหาความเอนเอียงในการทดลองด้วย Bland Altman plot พบว่ามี significant bias ดังนั้นสมการนี้จึงยังไม่สามารถนำมาใช้ได้(Shruthi Soundara Rajan, 2020)

Temsch W และคณะได้ศึกษาหาค่า calculated HbA1c จากค่า HbA1c ในอดีต และค่า Self-measured blood glucose (SMBG) โดยใช้ model ทางคณิตศาสตร์ (truncated Fourier series) พบว่าสามารถทำนายได้ในกลุ่มผู้ป่วยที่ควบคุมระดับน้ำตาลได้ดีเท่านั้นหากเป็นกลุ่มผู้ป่วยที่ควบคุมระดับน้ำตาลได้ไม่ดีอาจทำให้แปลผลผิดได้(Temsch et al., 2008)

Yongjin Xu และคณะได้ทำการทดสอบความแม่นยำในการทำนายค่า HbA1c โดยใช้แบบจำลองทาง kinetic ชนิด Linear regression ที่มีตัวแปรต้นคือ ค่า RBC turnover rate และ The apparent hemoglobin glycation rate ซึ่งค่าที่ได้เรียกว่า calculated HbA1c โดยศึกษาในกลุ่มผู้ป่วย Type 1 Diabetes ที่ได้รับการรักษาด้วยอินซูลิน ผ่านเครื่อง sensor-augmented pump มีกลุ่มตัวอย่าง 51 ราย ซึ่งมีค่า HbA1c จากห้องปฏิบัติการ อยู่ในช่วง 6.6 – 7.5 ทำการเปรียบเทียบผลจากการทำนาย 3 แบบคือ 1) ค่า calculated HbA1c 2) ค่า estimated HbA1c (ได้จากการคำนวณด้วยวิธี the A1c-Derived Average Glucose) 3) ค่า Glucose management indicator (GMI) กับค่า HbA1c จากห้องปฏิบัติการพบว่าค่า calculated HbA1c มีเปอร์เซ็นต์ในการทำนายที่แม่นยำกว่าแบบอื่น เมื่อเปรียบเทียบค่า Mean Absolute Deviation (MAD, MAE) calculated HbA1c ได้ 0.11% , estimated HbA1c ได้ 0.54% และ GMI ได้ 0.47% (Xu et al., 2021)

WuQiang fan และคณะ ได้ศึกษาการประมาณค่า HbA1c จากค่า Self-Monitored Blood Glucose values (SMBG) ผ่านสมการทางคณิตศาสตร์ โดยทำการเก็บค่าน้ำตาลก่อนอาหาร (Pre-meal) และค่าน้ำตาลหลังอาหาร (Post-meal) จากกลุ่มตัวอย่างทั้งหมด 547 คน ซึ่งประกอบไปด้วยผู้ป่วยเบาหวานชนิดที่ 1 ผู้ป่วยเบาหวานชนิดที่ 2 และผู้ที่ไม่เป็นเบาหวาน พบว่าสมการที่ได้เป็นความสัมพันธ์แบบ linear regression โดยมีตัวแปรสำคัญคือ Pre-dinner SMBG และ Pre-breakfast SMBG โดยสามารถใช้ตัวแปรแค่สองตัวก็สามารถทดแทนตัวแปรทั้งหมด 6 ครั้งได้ แต่การศึกษานี้มีข้อจำกัดคือกลุ่มตัวอย่างเป็นกลุ่มผู้ป่วยที่มีค่า HbA1c คงที่ตลอดการศึกษาซึ่งอาจไม่เหมาะสมในการนำมาใช้กับผู้ป่วยที่มีค่าน้ำตาลในเลือดไม่คงที่ (WuQiang Fan, 2018)

Richard M. Bergenstal และคณะ ได้ศึกษาการประมาณค่า HbA1c ที่ได้จากเครื่องตรวจวัดระดับน้ำตาลในเลือดแบบต่อเนื่อง (Continuous Glucose Monitoring) โดยมีชื่อเรียกในขั้นต้นคือ Estimated A1C แต่อาจทำให้เกิดความสับสนจึงได้มีการเปลี่ยนชื่อใหม่เป็น Glucose Management Indicator (GMI) จากนั้นนำมาผ่านสมการทางคณิตศาสตร์ แล้วเปรียบเทียบกับค่า HbA1c ที่ได้จากห้องปฏิบัติการ พบว่าค่าที่ได้ยังไม่สอดคล้องกัน ซึ่งในขณะนี้สามารถเพื่อการศึกษาและการจัดการดูแลผู้ป่วยได้เท่านั้น (Bergenstal et al., 2018)



บทที่ 3

วิธีดำเนินการวิจัย

รูปแบบการวิจัย

การวิจัยศึกษาวิจัยแบบกึ่งทดลอง (Quasi-experimental study research) โดยวิธีการ (Methodology) จัดการข้อมูลด้วยกระบวนการเหมืองข้อมูล (Data mining) และสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ด้วยเครื่อง (Machine learning)

ประชากรและกลุ่มตัวอย่าง

ประชากร คือข้อมูลจากเวชระเบียนอิเล็กทรอนิกส์ของผู้ป่วยนอกที่มารับบริการที่คลินิกเบาหวาน ณ โรงพยาบาลเฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยะเวลาโดยมีคุณสมบัติคือ ได้รับการวินิจฉัยเป็นโรคเบาหวานชนิดที่ 2 ได้รับการรักษาด้วยยาลดระดับน้ำตาลในเลือดอย่างน้อย 1 ชนิด

กลุ่มตัวอย่าง คือประชากรที่มารับบริการในช่วงระยะเวลาตั้งแต่ ตุลาคม 2558 – กันยายน 2564

การคัดเลือกเข้า (Inclusion criteria) คือกลุ่มตัวอย่างที่ได้รับการวินิจฉัยเป็นโรคเบาหวานชนิดที่ 2 และได้รับการรักษาด้วยยาลดระดับน้ำตาลในเลือดอย่างน้อย 1 ชนิดและมีข้อมูลตัวแปรครบถ้วนมากกว่าหรือเท่ากับ 60%

การคัดออก (Exclusion criteria) คือกลุ่มตัวอย่างที่ไม่เข้าเกณฑ์การคัดเลือกเข้าหรือไม่มีข้อมูลตัวแปรตาม (ค่า HbA1c) หรือมีข้อมูลตัวแปรน้อยกว่า 60%

การเตรียมข้อมูล

ใช้กระบวนการ Data mining มีลำดับดังนี้

1. Data Selection เป็นขั้นตอนสำหรับการคัดเลือกข้อมูลที่เกี่ยวข้อง เมื่อกำหนดให้ตัวแปรตามคือค่า HbA1c เป็นเป้าหมายสำคัญสำหรับงานวิจัยนี้ ดังนั้นข้อมูลที่เกี่ยวข้องที่สามารถนำมาพิจารณาสร้างแบบจำลองจึงกำหนดผ่านการทบทวนวรรณกรรม โดยตัวแปรต้นได้เป็นกลุ่มข้อมูลดังนี้
 - ข้อมูลทางประชากรประกอบด้วย (อายุกำหนดเป็นจำนวนเต็มของปี), เพศ(ชายและหญิง), ค่า BMI (ได้จากค่าน้ำหนักและส่วนสูง), ประวัติครอบครัวของการป่วยเบาหวาน (ผู้ป่วยที่จะถูกระบุว่า

มีประวัติครอบครัวคือมีญาติสายตรงที่ได้รับการวินิจฉัยว่าเป็นเบาหวาน), ระยะเวลาของการป่วยเบาหวาน (คือระยะเวลาที่ได้รับการวินิจฉัยว่าเป็นเบาหวานชนิดที่ 2 หรือได้รับการขึ้นทะเบียนผู้ป่วยเบาหวานชนิดที่สอง จนถึงเวลาที่มารับบริการ), การสูบบุหรี่, สภาวะทางอารมณ์ (หมายถึงการได้รับผลบวกจากการทดสอบภาวะความเครียดจากการเจ็บป่วยของคลินิกเบาหวาน)

- ข้อมูลการห้องปฏิบัติการ (ค่าน้ำตาล FBS, ค่าไขมัน Total cholesterol, ค่าไขมัน Triglycerides, ค่าไขมัน HDL, ค่าไขมัน LDL, ค่า systolic blood pressure [SBP], ค่า albumin-to-creatinine ratio [ACR], ค่า estimated glomerular filtration rate [eGFR] (ได้จากบันทึกในเวชระเบียนอิเล็กทรอนิกส์หรือจากการคำนวณโดยใช้สูตร CKD-EPI formula, ค่า micro/macro-albuminuria)
- ข้อมูลโรคร่วม ได้แก่ โรค Hypertension (ผู้ป่วยที่ได้รับการวินิจฉัยและระบุในเวชระเบียนอิเล็กทรอนิกส์), Dyslipidemia (ผู้ป่วยที่ได้รับการวินิจฉัยและระบุในเวชระเบียนอิเล็กทรอนิกส์), Chronic kidney disease (ผู้ป่วยที่มีค่า eGFR < 60)
- ข้อมูลด้านยาลดระดับน้ำตาลในเลือดที่ได้รับ โดยพิจารณาจากกลุ่มยาที่ได้รับ ได้แก่ (Metformin, Sulphonyl urea, Thiazolidinediones, Insulin) และจำนวนชนิดของยาลดระดับน้ำตาลในเลือดที่ได้รับ
- ข้อมูลทางเศรษฐศาสตร์ (ระดับการศึกษา, ระดับรายได้)

จากข้อมูลของตัวแปรต้นที่กำหนดไว้พบว่าตัวแปรบางอย่างไม่สามารถนำมาใช้ได้ เนื่องจากมีการบันทึกไว้จำนวนน้อยมาก ไม่ครอบคลุมกลุ่มตัวอย่าง และไม่สามารถใช้ค่าเฉลี่ยเพื่อทดแทนค่าจริงได้ จึงต้องกำจัดตัวแปรเหล่านั้นไปได้แก่ ข้อมูลทางประชากร คือ ตัวแปรระยะเวลาของการป่วยเบาหวาน, ข้อมูลการห้องปฏิบัติการ คือ ตัวแปรค่า albumin-to-creatinine ratio (ACR) และค่า micro/macro-albuminuria ,ข้อมูลโรคร่วม คือ ตัวแปรโรคไขมันในเลือดสูง (Dyslipidemia) และข้อมูลทางเศรษฐศาสตร์ คือ ตัวแปรระดับรายได้

จากนั้นตัดข้อมูลตัวแปรที่ไม่จำเป็นและสามารถระบุข้อมูลย้อนกลับถึงกลุ่มตัวอย่างได้ตามแนวทางจริยธรรมสำหรับการวิจัย ได้แก่ ข้อมูล HN, VN, ชื่อ-นามสกุล

โดยสรุปข้อมูลที่นำเข้ามาครั้งแรกมีกลุ่มตัวอย่างทั้งหมด 46,727 ราย เมื่อนำมาผ่านการพิจารณาผ่านเกณฑ์การคัดเข้า-คัดออก จะเหลือกลุ่มตัวอย่างทั้งหมด 7,247 ราย

2. Data Cleaning เป็นขั้นตอนการกลั่นกรองข้อมูล มีการลบข้อมูลที่ซ้ำซ้อน แก้ไขข้อมูลให้ถูกต้อง จัดการกับข้อมูลที่หายไปหรือข้อมูลที่ผิดปกติ เช่นการพบการพิมพ์ข้อความผสมในข้อมูลชนิดตัวเลข ให้ลบข้อความนั้นออกเหลือแต่ตัวเลขเท่านั้น และมีการพบการพิมพ์จุดทศนิยมที่ผิดทำให้ค่าของข้อมูลมากหรือน้อยกว่าปกติ ในกรณีของ Missing value ให้ใช้ค่าเฉลี่ยของตัวแปรนั้นๆ ทดแทนค่าที่หายไป จากนั้นกำจัดค่าที่สูงมากหรือต่ำมากผิดปกติ หรือ Outlier โดยใช้วิธีพิจารณาค่าพิสัยระหว่างควอไทล์ (interquartile range, IQR) เนื่องจากไม่ต้องพิจารณาถึงการกระจายและผลที่ได้มีความน่าเชื่อถือ เมื่อนำข้อมูลมาผ่านกระบวนการนี้จะคงเหลือข้อมูลกลุ่มตัวอย่างเท่ากับ 5,067 ราย

3. Data Transformation เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการสร้างแบบจำลอง เช่น การเปลี่ยนแปลงข้อมูลจาก nominal data เป็นตัวเลข (ordinal data) เพื่อให้ง่ายต่อการประมวลผล อาทิ เพศชาย = 1 เพศหญิง = 0, ไม่สูบบุหรี่ = 0 สูบบุหรี่ = 1 ซึ่งสามารถแสดงได้ตามตารางที่ 1 แต่แบบจำลองบางชนิดไม่รองรับข้อมูลชนิด ordinal data ดังนั้นจึงมีการแปลงข้อมูลอีกครั้งให้เหมาะสมคือแปลงเป็นตัวแปรหุ่น (Dummy variable) อีกครั้งก่อนนำมาสร้างแบบจำลอง

ตารางที่ 1 ตารางแสดง สรุปชุดข้อมูลตัวแปรที่นำมาใช้

ตัวแปร	คำอธิบาย	ช่วงข้อมูล
HbA1c (mg%)	ค่าน้ำตาลเฉลี่ยสะสม	3.2 – 11.2
Age (years)	อายุของผู้ป่วย	25 - 89
Sex (male=1, female=0)	เพศ	1,0
BMI (kg/m ²)	ค่าดัชนีมวลกาย	13.43 – 38.83
SBP (mmHg)	ค่าความดันโลหิตตัวบน	95 - 167
FBS (mg/dL)	ค่าน้ำตาลในเลือดหลังอดอาหาร	40.5 – 252.5
BUN (mg/dL)	ค่าปริมาณไนโตรเจนในเลือด	0.4 - 30
Creatinine (mg/dL)	ค่าปริมาณ creatinine	0.005 – 1.805
eGFR (mL/min/1.73m ²)	ค่าปริมาณเลือดที่ไหลผ่านตัวกรองของไตในหนึ่งนาที	-1.11 – 165.53
HDL (mg/dL)	ค่าไขมันชนิด HDL	15 - 79
LDL (mg/dL)	ค่าไขมันชนิด LDL	3.5 – 215.5

Cholesterol (mg/dL)	ค่าไขมันชนิด Cholesterol	74 - 306
Triglyceride (mg/dL)	ค่าไขมันชนิด Triglyceride	-44.5 - 351
Smoking (yes=1, no=0)	การสูบบุหรี่	1,0
Education	ระดับการศึกษา	0 - 9
HT combination (yes=1, no=0)	มีโรคร่วมเป็นความดันโลหิตสูง	1,0
Stress (yes=1, no=0)	มีภาวะเครียดเนื่องมาจากการเป็นเบาหวาน	1,0
Family history (yes=1, no=0)	มีประวัติญาติสายตรงเป็นเบาหวาน	1,0
Metformin used (yes=1, no=0)	ได้รับยากลุ่ม Biguanides คือยา Metformin	1,0
Sulfonyl urea used (yes=1, no=0)	ได้รับยากลุ่ม Sulfonylureas	1,0
Thiazo used (yes=1, no=0)	ได้รับยากลุ่ม Thiazolidinediones	1,0
Insulin used (yes=1, no=0)	ได้รับยาอินซูลินชนิดฉีด	1,0
Medical used	จำนวนชนิดยาที่ได้รับ	1 - 4

การคัดเลือกตัวแปร

โดยจะพิจารณาคัดเลือกตัวแปรจากค่าความสัมพันธ์ด้วยวิธี Pearson และ Spearman เพื่อให้ครอบคลุมตัวแปรที่มีลักษณะแตกต่างกัน เพราะวิธีของ Pearson จะใช้ในกรณีที่เป็นตัวแปรชนิด Interval ขึ้นไป ตัวแปรมีการแจกแจงแบบปกติ และจะมีประสิทธิภาพมากขึ้นหากตัวแปรที่นำมาพิจารณามีความสัมพันธ์กันเป็นเชิงเส้นตรง ส่วนวิธีของ Spearman จะเหมาะสมกับตัวแปรชนิด ordinal และค่าตามสัมพันธ์ที่ได้อาจเป็นเชิงเส้นหรือไม่ก็ได้

เมื่อพิจารณาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างทั้งหมด 5,067 ราย ตามตารางที่ 2 พบว่าตัวแปรต้นทั้งหมดมีความสัมพันธ์กับตัวแปรตามหรือค่า HbA1c ใน

ปริมาณที่ต่ำมาก โดยตัวแปรที่มีความสัมพันธ์สูงสุดคือ ค่า FBS มีความสัมพันธ์เชิงบวก อยู่ที่ 0.4744 ลำดับที่ 2 คือการได้รับยาฉีด Insulin ลำดับที่ 3 คือจำนวนของยาที่ได้รับ ลำดับที่ 4 เป็นความสัมพันธ์เชิงลบนั้นคือค่า Creatinine

เมื่อพิจารณาในเฉพาะกลุ่มตัวอย่างที่ไม่มีโรคร่วม Chronic Kidney Disease (CKD) (กลุ่มตัวอย่าง จำนวน 4287 ราย) ตามตามตารางที่ 3 พบว่าความสัมพันธ์ของตัวแปรต้นและตัวแปรตาม คล้ายคลึงกับกลุ่มตัวอย่างทั้งหมดโดยเฉพาะลำดับแรก แต่ความสัมพันธ์จะมากกว่าเล็กน้อย เช่น ความสัมพันธ์ระหว่างค่า FBS ของกลุ่มตัวอย่างที่ไม่รวมผู้ป่วย CKD เท่ากับ 0.4807 มากกว่า 0.4744 ที่แสดงจากความสัมพันธ์ค่า FBS ของกลุ่มตัวอย่างทั้งหมด ค่าความสัมพันธ์ที่มากขึ้นจะส่งผลต่อประสิทธิภาพในกระบวนการสร้างแบบจำลองในขั้นตอนถัดไป

เมื่อพิจารณาด้านความสัมพันธ์ที่มีนัยสำคัญ โดยกำหนด $P\text{-value} < 0.05$ พบว่าตัวแปรต้นที่ไม่มีไม่มีความสัมพันธ์กับตัวแปรตามได้แก่ ค่า SBP, ค่า BUN, ระดับการศึกษา และการได้รับยา Metformin ดังนั้นในขั้นตอนสร้างแบบจำลองสามารถพิจารณนำตัวแปรที่ไม่มีความสัมพันธ์นี้ออกไปได้

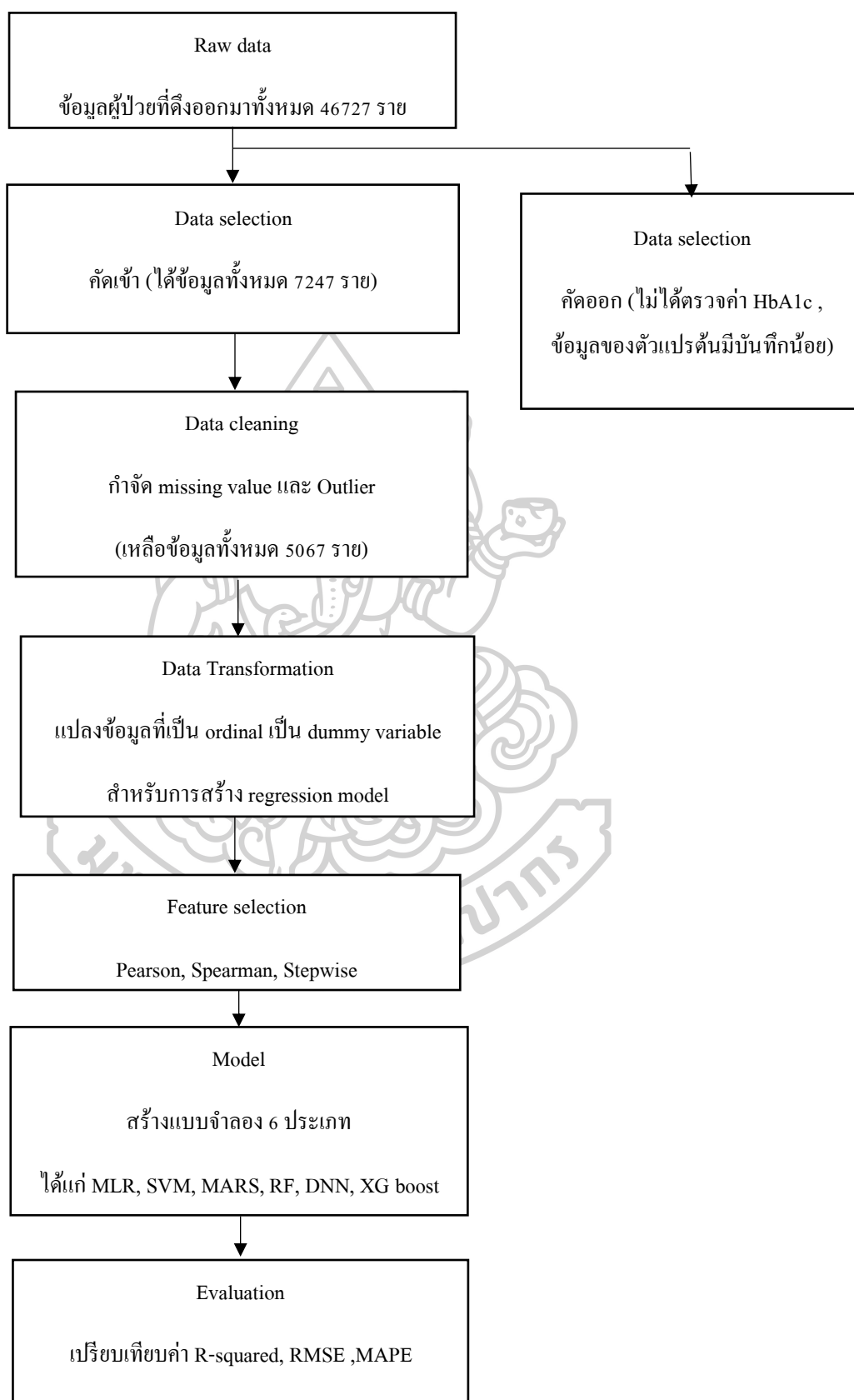
มีการพิจารณาการคัดเลือกตัวแปรอีกหนึ่งวิธีคือวิธี Stepwise ซึ่งเป็นวิธีที่พบบ่อยเมื่อสร้างแบบจำลองชนิด Linear regression โดยแบบจำลองที่กำลังสร้างจะทำการสุ่มเอาตัวแปรต้นเข้ามาเพิ่มในแบบจำลองและสุ่มเอาตัวแปรออก สลับไปเรื่อย ๆ จนได้แบบจำลองที่ดีที่สุด โดยตัวแปรที่พบเมื่อใช้กลุ่มตัวอย่างทั้งหมด คือ Age, FBS, BUN, eGFR, HDL, Cholesterol, Triglyceride, Med count, Sex, Smoking, HT combination, Family history, MFM used, SU used, Thiazolidinediones used และกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ปัจจัยที่พบว่ามีความสัมพันธ์ด้วยวิธี Stepwise ได้แก่ Age, FBS, BUN, Creatinine, eGFR, HDL, LDL, Triglyceride, Med count, Smoking, Stress, HT combination, Family history, MFM used, SU used, Thiazolidinediones used

สรุปการคัดเลือกตัวแปรจากหลายวิธีการพบว่าตัวแปรที่ไม่มีความสัมพันธ์ สามารถตัดออกได้แล้วคือ ค่า SBP และ ระดับการศึกษา ดังนั้นตัวแปรที่นำมาใช้จะเหลือ 20 ตัวแปรและการจัดกลุ่มตัวอย่างใหม่โดยตัดข้อมูลของผู้ป่วย CKD ออกไปโดยเปรียบเทียบกับกลุ่มตัวอย่างทั้งหมด จะสามารถแบ่งกลุ่มตัวอย่างออกเป็น 4 กลุ่มคือ

1. กลุ่มตัวอย่างทั้งหมดใช้ตัวแปรทั้งหมด (5067 ราย 20 ปัจจัย)
2. กลุ่มตัวอย่างไม่รวมผู้ป่วย CKD ใช้ตัวแปรทั้งหมด (4287 ราย 20 ปัจจัย)
3. กลุ่มตัวอย่างทั้งหมด ใช้ตัวแปรจากการคัดเลือกชนิด Stepwise (5067 ราย 15 ปัจจัย)
4. กลุ่มตัวอย่างไม่รวมผู้ป่วย CKD ใช้ตัวแปรจากการคัดเลือกชนิด Stepwise (4287 ราย 16 ปัจจัย)



รูปภาพที่ 1 Flow chart แสดงกระบวนการดำเนินงานวิจัย



บทที่ 4 ผลการวิจัย

งานวิจัยนี้มีจุดมุ่งหมายในการสร้างแบบจำลองสำหรับทำนายด้วยเทคนิคการเรียนรู้ด้วยเครื่อง เพื่อประมาณค่า HbA1c ในผู้ป่วยเบาหวานชนิดที่ 2 ซึ่งใช้ข้อมูลจากเวชระเบียนของ โรงพยาบาล เฉลิมพระเกียรติสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี ระยะเวลาเก็บข้อมูลตั้งแต่ตุลาคม 2558 - กันยายน 2564 จะได้กลุ่มตัวอย่าง 46,727 ราย และตัวแปรที่ศึกษา 23 ตัวแปรที่ผ่านการ คัดเลือกเข้ามาจากการทบทวนวรรณกรรม เมื่อผ่านกระบวนการจัดการข้อมูลด้วยเทคนิค Data mining แล้วจะมีกลุ่มตัวอย่างทั้งหมด 5,067 ราย

มีการจัดกลุ่มตัวอย่างใหม่โดยพิจารณาตัดผู้ป่วยที่มีโรคร่วม CKD ออกไปเนื่องจากมีการศึกษา ว่า การมีภาวะโรค CKD ร่วมด้วยจะทำให้ค่า HbA1c ที่สามารถตรวจได้จากห้องปฏิบัติการมีความ คลาดเคลื่อน(Radin, 2014) เมื่อค่า HbA1c ที่นำมาสร้างแบบจำลองมีความคลาดเคลื่อน แปรปรวน ไม่ตรงกับค่าที่เป็นจริงอาจทำให้การสร้างแบบจำลองมีความผิดพลาดและแบบจำลองที่ได้มี ประสิทธิภาพในการทำนายลดลง ดังนั้นผู้วิจัยจึงได้ทำการตัดกลุ่มตัวอย่างผู้ป่วยที่มีโรคร่วม CKD ออกไปจะได้กลุ่มตัวอย่างใหม่ที่มีผู้ป่วยทั้งหมด 4,287 ราย เมื่อหาความสัมพันธ์ด้วยวิธีของ Pearson ในตัวแปรต้นที่มีลักษณะข้อมูลชนิด interval และ Spearman ในข้อมูลชนิดที่เป็น ordinal พร้อมทั้ง หาค่า p-value ตามตารางที่ 2 ซึ่งแสดงความสัมพันธ์ของกลุ่มตัวอย่างทั้งหมดและตารางที่ 3 แสดง ความสัมพันธ์ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD

ตารางที่ 2 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างทั้งหมด

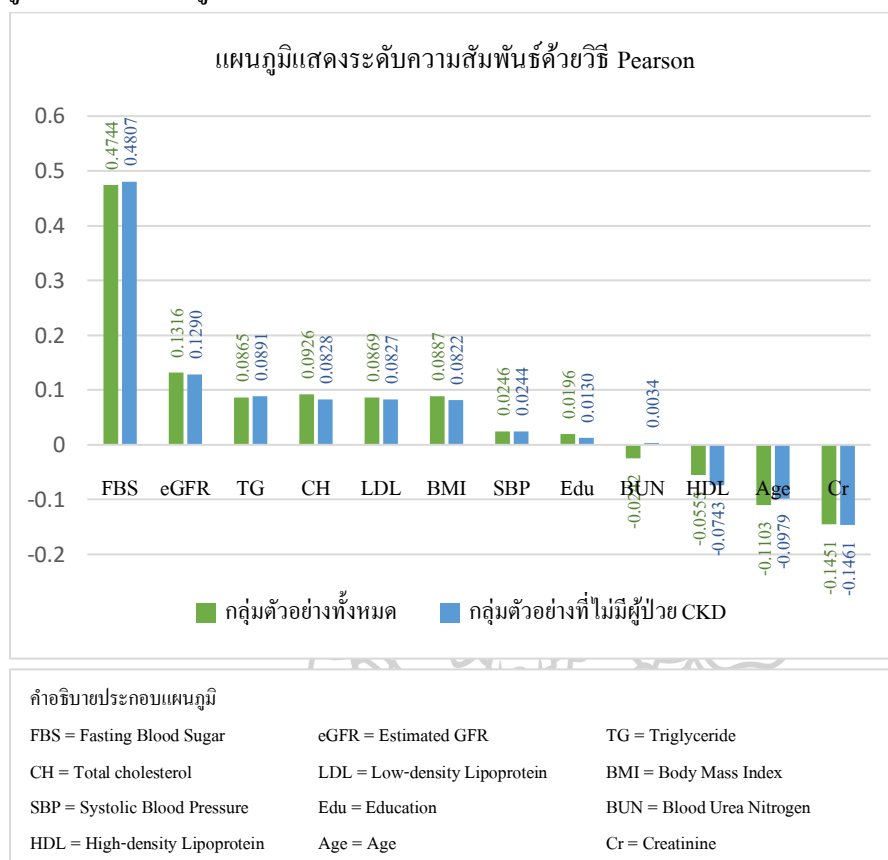
ตัวแปรต้น	วิธี Pearson	วิธี Spearman
Age	-0.1103 (p-value =3.476e-15)	-0.1172 (p-value <2.2e-16)
Sex	-0.1158 (p-value <2.2e-16)	-0.1182 (p-value <2.2e-16)
BMI	0.0887 (p-value =2.544e-10)	0.1063 (p-value =3.227e-14)
SBP	0.0246 (p-value =0.07954)	0.0309 (p-value =0.02792)
FBS	0.4744 (p-value <2.2e-16)	0.4738 (p-value <2.2e-16)
BUN	-0.0242 (p-value =0.08447)	-0.0310 (p-value =0.02715)
Creatinine	-0.1451 (p-value <2.2e-16)	-0.1520 (p-value <2.2e-16)
eGFR	0.1316 (p-value <2.2e-16)	0.1438 (p-value <2.2e-16)
HDL	-0.0555 (p-value = 7.671e-05)	-0.0519 (p-value =0.0002196)
LDL	0.0869 (p-value = 5.829e-10)	0.0839 (p-value = 2.243e-09)
Cholesterol	0.0926 (p-value = 4.01e-11)	0.0909 (p-value =9.021e-11)
Triglyceride	0.0865 (p-value = 7.046e-10)	0.0820 (p-value =5.05e-09)
Smoking	0.0333 (p-value =0.01785)	0.0385 (p-value = 0.006164)
Education	0.0196 (p-value =0.1622)	0.0209 (p-value = 0.1367)
HT combination	-0.0553 (p-value =8.267e-05)	-0.0575 (p-value = 4.272e-05)
Stress	0.0444 (p-value = 0.001561)	0.0440 (p-value = 0.001719)
Family history	0.0907 (p-value = 9.788e-11)	0.0894 (p-value =1.805e-10)
Metformin used	0.0008 (p-value =0.9537)	0.0081 (p-value = 0.5626)
Sulfonyl urea used	0.0481 (p-value = 0.0006157)	0.0647 (p-value = 3.98e-06)
Thiazo used	0.0679 (p-value =1.276e-06)	0.0738 (p-value = 1.458e-07)
Insulin used	0.2947 (p-value < 2.2e-16)	0.2768 (p-value <2.2e-16)
Medical count	0.2369 (p-value <2.2e-16)	0.2423 (p-value < 2.2e-16)

ตารางที่ 3 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย

CKD

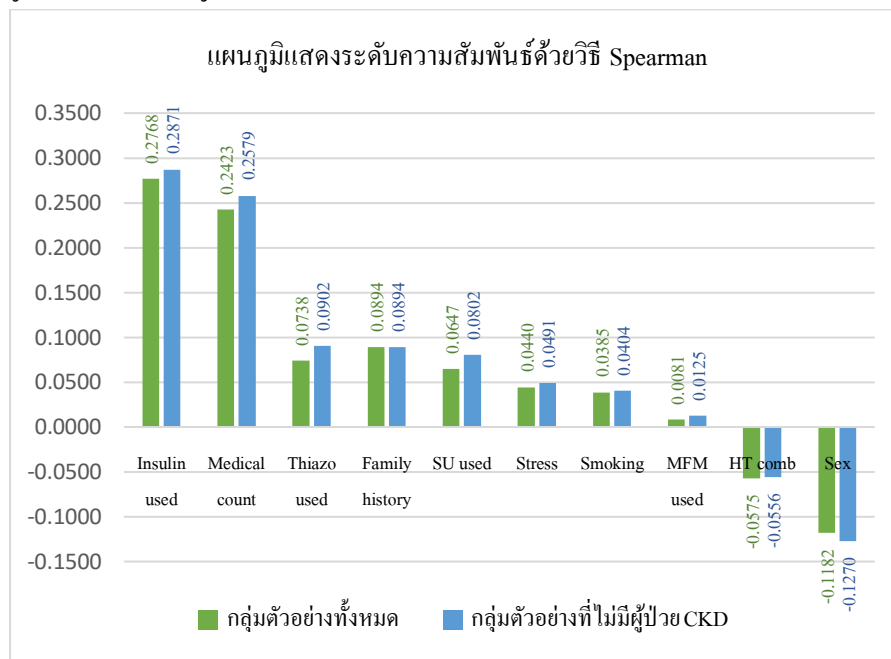
ตัวแปรต้น	วิธี Pearson	วิธี Spearman
Age	-0.0979 (p-value = 1.532e-10)	-0.1069 (p-value = 2.297e-12)
Sex	-0.1227 (p-value = 7.592e-16)	-0.1270 (p-value < 2.2e-16)
BMI	0.0822 (p-value = 7.135e-08)	0.1006 (p-value = 4.13e-11)
SBP	0.0244 (p-value = 0.1108)	0.0277 (p-value = 0.06966)
FBS	0.4807 (p-value < 2.2e-16)	0.4805 (p-value < 2.2e-16)
BUN	0.0034 (p-value = 0.8224)	-0.0072 (p-value = 0.6354)
Creatinine	-0.1461 (p-value < 2.2e-16)	-0.1433 (p-value < 2.2e-16)
eGFR	0.1290 (p-value < 2.2e-16)	0.1379 (p-value < 2.2e-16)
HDL	-0.0743 (p-value = 1.128e-06)	-0.0759 (p-value = 6.349e-07)
LDL	0.0827 (p-value = 5.981e-08)	0.0778 (p-value = 3.364e-07)
Cholesterol	0.0828 (p-value = 5.765e-08)	0.0797 (p-value = 1.728e-07)
Triglyceride	0.0891 (p-value = 5.039e-09)	0.0902 (p-value = 3.344e-09)
Smoking	0.0339 (p-value = 0.02641)	0.0404 (p-value = 0.008084)
Education	0.0130 (p-value = 0.3932)	0.0177 (p-value = 0.2458)
HT combination	-0.0539 (p-value = 0.0004171)	-0.0556 (p-value = 0.0002692)
Stress	0.0535 (p-value = 0.0004563)	0.0491 (p-value = 0.001295)
Family history	0.0905 (p-value = 2.945e-09)	0.0894 (p-value = 4.495e-09)
Metformin used	-0.0001 (p-value = 0.9933)	0.0125 (p-value = 0.4123)
Sulfonyl urea used	0.0646 (p-value = 2.29e-05)	0.0802 (p-value = 1.457e-07)
Thiazo used	0.0857 (p-value = 1.912e-08)	0.0902 (p-value = 3.217e-09)
Insulin used	0.3043 (p-value < 2.2e-16)	0.2871 (p-value < 2.2e-16)
Medical count	0.2516 (p-value < 2.2e-16)	0.2579 (p-value < 2.2e-16)

รูปภาพที่ 2 แผนภูมิแสดงระดับความสัมพันธ์ด้วยวิธี Pearson



จากรูปภาพที่ 2 การศึกษาระดับความสัมพันธ์ด้วยวิธี Pearson ซึ่งใช้ในกรณีที่เป็นตัวแปรชนิด Interval ขึ้นไป ตัวแปรมีการแจกแจงแบบปกติ และจะมีประสิทธิภาพมากขึ้นหากตัวแปรที่นำมาพิจารณามีความสัมพันธ์กันเป็นเชิงเส้นตรง จากการศึกษาค่าความสัมพันธ์ของตัวแปรชนิด Interval กับค่า HbA1c ด้วยวิธี Pearson พบว่าตัวแปรต้นที่มีระดับความสัมพันธ์กับตัวแปรตามสูงที่สุดคือ ค่า FBS มีค่า correlation coefficient เท่ากับ 0.4744 ในกลุ่มตัวอย่างทั้งหมดและ 0.4807 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ตัวแปรที่มีระดับความสัมพันธ์รองลงมาคือค่า Creatinine ซึ่งเป็นความสัมพันธ์เชิงลบ -0.1451 ในกลุ่มตัวอย่างทั้งหมดและ -0.1461 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ลำดับที่ 3 คือค่า eGFR เท่ากับ 0.1316 ในกลุ่มตัวอย่างทั้งหมดและ 0.1290 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ส่วนตัวแปรต้นที่พบว่ามีระดับความสัมพันธ์น้อย 3 ลำดับล่างสุดคือ ค่า SBP, ระดับการศึกษา และค่า BUN ซึ่งแสดงความสัมพันธ์ที่แตกต่างในแต่ละกลุ่มทดลองโดยในกลุ่มตัวอย่างทั้งหมดแสดงความสัมพันธ์เชิงลบแต่ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD กลับแสดงความสัมพันธ์เชิงบวก

รูปภาพที่ 3 แผนภูมิระดับความสัมพันธ์ด้วยวิธี Spearman



คำอธิบายประกอบแผนภูมิ

Insulin used = การได้รับยาอินซูลิน

Thiazio used = การได้รับยาในกลุ่ม thiazolidinediones

SU used การได้รับยาในกลุ่ม sulfonylureas

Smoking = การสูบบุหรี่

HT comb = มีโรคร่วมเป็นโรคความดันโลหิตสูง

Medical count = จำนวนยาลดน้ำตาลที่ได้รับ

Family history = ประวัติครอบครัว

Stress = ภาวะความเครียด

MFM used = การได้รับยา Metformin

Sex = เพศ

จากรูปภาพที่ 3 การศึกษาความสัมพันธ์ด้วยวิธี Spearman ซึ่งเหมาะสมกับตัวแปรชนิด ordinal และค่าตามสัมพันธ์ที่ได้อาจเป็นเชิงเส้นหรือไม่ก็ได้ พบว่าตัวแปรต้นที่มีระดับความสัมพันธ์กับตัวแปรตามสูงที่สุดคือการได้รับยา insulin เท่ากับ 0.2768 ในกลุ่มตัวอย่างทั้งหมดและ 0.2871 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ตัวแปรที่มีระดับความสัมพันธ์รองลงมาคือ จำนวนชนิดของยาลดระดับน้ำตาลในเลือดที่ได้รับเท่ากับ 0.2423 ในกลุ่มตัวอย่างทั้งหมดและ 0.579 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ลำดับที่ 3 คือ เพศ เป็นความสัมพันธ์เชิงลบเท่ากับ -0.1182 ในกลุ่มตัวอย่างทั้งหมดและ -0.1270 ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ส่วนตัวแปรต้นที่พบว่ามีระดับความสัมพันธ์น้อยที่สุดคือการได้รับยา metformin

จากการศึกษาอิทธิพลของขนาดและลักษณะของกลุ่มตัวอย่างที่แตกต่างกัน 2 กลุ่ม คือกลุ่มตัวอย่างทั้งหมดที่มีจำนวนผู้ป่วย 5,067 รายมีทั้งผู้ป่วยปกติและผู้ป่วยที่มีโรคร่วมเป็น CKD เปรียบเทียบกับกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD จำนวน 4,287 ราย ตามรูปภาพที่ 1 และ 2 พบว่าระดับความสัมพันธ์ของทั้ง 2 กลุ่มตัวอย่างมีความใกล้เคียงกันมากเกือบทุกตัวแปร แต่กลุ่มตัวอย่าง

ที่ไม่มีผู้ป่วย CKD พบตัวแปรที่มีระดับความสัมพันธ์ที่มากกว่าจำนวน 12 ตัวแปรจากทั้งหมด 22 ตัวแปร แสดงว่าการคัดสรรกลุ่มตัวอย่างให้มีความคล้ายคลึงกันจะสามารถเพิ่มระดับความสัมพันธ์ระหว่างตัวแปรได้ และความคล้ายคลึงกันของกลุ่มตัวอย่างมีความสำคัญมากกว่าขนาดของกลุ่มตัวอย่าง

การศึกษาความสัมพันธ์ได้กำหนดระดับนัยสำคัญที่ 0.05 พบว่าตัวแปรที่ไม่มีความสัมพันธ์อย่างมีนัยสำคัญของทั้ง 2 กลุ่มตัวอย่างมีความเหมือนกัน นั่นคือพบว่า ค่า SBP, ค่า BUN, ระดับการศึกษา และการได้รับยา Metformin ไม่มีความสัมพันธ์กับตัวแปรตาม (ค่า HbA1c) ที่ระดับนัยสำคัญ (p-value < 0.05)

พิจารณาการคัดเลือกตัวแปรด้วยวิธี Stepwise ซึ่งเป็นกระบวนการที่มักจะพบในแบบจำลองชนิด Linear regression จะได้ตัวแปรที่มีความสัมพันธ์ในแต่ละกลุ่มตัวอย่างดังนี้

1. กลุ่มตัวอย่างทั้งหมด ตัวแปรที่มีความสัมพันธ์ได้แก่ Age, FBS, BUN, eGFR, HDL, Cholesterol, Triglyceride, Med count, Sex, Smoking, HT combination, Family history, MFM used, SU used, Thiazolidinediones used
2. กลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ได้แก่ Age, FBS, BUN, Creatinine, eGFR, HDL, LDL, Triglyceride, Med count, Smoking, Stress, HT combination, Family history, MFM used, SU used, Thiazolidinediones used

จากข้อมูลการคัดเลือกตัวแปรดังกล่าวข้างต้น จะพบว่าตัวแปรต้นที่สามารถตัดออกไปได้เลยคือค่า SBP และ ระดับการศึกษา เพราะไม่พบว่ามีความสัมพันธ์ที่มีนัยสำคัญไม่ว่าจะด้วยวิธีการทดสอบความสัมพันธ์ใดๆ

ดังนั้นจึงสามารถแบ่งกลุ่มตัวอย่างออกเป็น 4 กลุ่มดังนี้

1. กลุ่มตัวอย่างทั้งหมด ใช้ตัวแปรทั้งหมด (5067 ราย 20 ปัจจัย)
2. กลุ่มตัวอย่างไม่รวมผู้ป่วย CKD ใช้ตัวแปรทั้งหมด (4287 ราย 20 ปัจจัย)
3. กลุ่มตัวอย่างทั้งหมด ใช้ตัวแปรจากการคัดเลือกชนิด Stepwise (5067 ราย 15 ปัจจัย)
4. กลุ่มตัวอย่างไม่รวมผู้ป่วย CKD ใช้ตัวแปรจากการคัดเลือกชนิด Stepwise (4287 ราย 16 ปัจจัย)

ในขั้นตอนของการสร้างแบบจำลอง มีการแบ่งกลุ่มตัวอย่างออกเป็นกลุ่ม Training 80 % และกลุ่ม Testing 20 % โดยใช้โปรแกรม RStudio version 4.3.1 ซึ่งจะทดสอบในแบบจำลอง 6 ประเภทที่แตกต่างกันดังนี้

1. Multiple linear regression เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรต้นหลายตัว และตัวแปรตาม เป็นแบบจำลองที่ได้รับความนิยมมากเนื่องจากง่ายต่อการประมวลผล และแสดงว่าตัวแปรนั้นมีความสัมพันธ์ที่เป็นเชิงเส้นตรงจริง ๆ

2. Support vector machine ใช้หลักการของการสร้างเส้นแบ่งแยกข้อมูลออกจากกันโดยมีพื้นฐานมาจากการคำนวณแบบ Linear regression มีหลักการสำคัญคือลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด ลดความผิดพลาดจากการทำนายและเพิ่มระยะการแยกแยะให้มากที่สุด สามารถสร้างแบบจำลองชนิด Classification และ Regression ที่สามารถสร้างสมการถดถอยได้ทั้งแบบ linear และ non-linear

3. Random forest เป็นแบบจำลองที่พัฒนามาจากต้นไม้ตัดสินใจ (Decision tree) โดยจะสร้างต้นไม้ตัดสินใจหลาย ๆ ต้นแต่ละต้นมีข้อมูลตัวแปรที่แตกต่างกันจากนั้นจะทำนายผลของต้นไม้ทั้งหมดและทำการโหวต เลือกผลการทำนายที่ได้ผลโหวตมากที่สุด สามารถใช้ในกรณีแบบจำลองชนิด Classification และ Regression และในการสร้างแบบจำลองไม่ต้องคำนึงว่าจะมีความสัมพันธ์กันเป็นเชิงเส้นหรือไม่

4. Multivariate Adaptive Regression Splines (MARS) สามารถใช้ได้ทั้งในรูปแบบเชิงเส้นและไม่เป็นเชิงเส้น และมีข้อดีในกระบวนการสร้างแบบจำลองจะมีการคัดเลือกตัวแปรให้เหมาะสมที่สุด

5. Deep neural network มีหลักการทำงานคล้ายกับสมองของมนุษย์ ที่ประกอบด้วยเซลล์ประสาท(node) เชื่อมโยงกันเป็นชั้น ๆ(layer) ด้วยเส้นเชื่อม (edge) ซึ่งจะมีการถ่วงน้ำหนักทำให้ได้แบบจำลองที่ดีที่สุด สามารถใช้ได้ความสัมพันธ์แบบ linear และ non-linear นอกจากนี้ยังมีข้อดีคือไม่จำเป็นต้องผ่านกระบวนการ feature selection

6. Extreme gradient boost (XG boost) เป็นแบบจำลองที่พัฒนามาจากต้นไม้ตัดสินใจ (Decision tree) อีกแบบเป็นการสร้างต้นไม้ตัดสินใจต่อกันมีการเรียนรู้ข้อผิดพลาดจากต้นไม้ตัดสินใจก่อนทำซ้ำไปเรื่อย ๆจนได้แบบจำลองที่มีประสิทธิภาพดีที่สุด

ผลการทดลอง

ตารางที่ 4 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิด โดยใช้ตัวแปรทั้งหมด

รายการ	กลุ่มตัวอย่างทั้งหมด			กลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD		
	R squared	RMSE	MAPE	R squared	RMSE	MAPE
MLR	0.2881	1.1565	0.1258	0.3293	1.1164	0.1205
SVM	0.3230	1.1347	0.1205	0.3537	1.1007	0.1160
RF	0.3441	1.1095	0.1226	0.3726	1.0813	0.1180
MARS	0.2834	1.1601	0.1259	0.3545	1.0957	0.1185
DNN	0.3080	1.0733	0.1125	0.3229	1.1011	0.1209
XG boost	0.3401	1.1189	0.1262	0.3560	1.0815	0.1168

ข้อมูลจากตารางที่ 4 แสดงประสิทธิภาพของแบบจำลองโดยใช้ตัวแปรทั้งหมด 20 ตัวแปรที่ผ่านกระบวนการคัดเลือกด้วยวิธี Pearson และ Spearman พิจารณาค่า R-squared เป็นหลักพบว่า

ในกลุ่มตัวอย่างทั้งหมดแบบจำลองที่มีประสิทธิภาพในการทำนายมากที่สุด คือ Random forest 34.41% ซึ่งใกล้เคียงกับแบบจำลองชนิด XG boost 34.01% รองลงมาคือ Support Vector Machine 32.30%, Deep Neural Network 30.80 % และแบบจำลองที่มีประสิทธิภาพน้อยที่สุดคือ MLR 28.81% และ MARS 28.34%

ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD จะพบว่าค่า R squared ที่ได้จะมากกว่ากลุ่มตัวอย่างทั้งหมดในทุกแบบจำลอง แสดงถึงประสิทธิภาพในการทำนายที่ดีว่า โดยแบบจำลองชนิด Random forest มีประสิทธิภาพในการทำนายมากขึ้นเท่ากับ 37.26% และ XG boost 35.60% ดังนั้นการตัดกลุ่มผู้ป่วยที่มีโรคร่วมเป็น CKD จึงมีความสำคัญต่อการสร้างแบบจำลองสำหรับทำนาย

เมื่อพิจารณาค่า RMSE ที่แสดงถึงความผิดพลาดจะพบว่าทั้งในกลุ่มตัวอย่างทั้งหมดและกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD แต่ละแบบจำลองให้ความแตกต่างกันไม่มากนัก โดยมีค่า RMSE เฉลี่ย

อยู่ที่ 1.1 แสดงว่าผลการทำนายค่าHbA1c มีความคลาดเคลื่อนจากค่าจริง +/- 1.1 จุดส่วนค่า MAPE มีค่าเฉลี่ยอยู่ที่ 0.12 แสดงว่าผลการทำนายค่าHbA1c มีความคลาดเคลื่อนจากค่าจริง +/- 12% ทั้งนี้ ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ยังพบว่าค่า RMSE และค่า MAPE ลดลง ยกเว้นแบบจำลองชนิด Deep neural network

ตารางที่ 5 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดโดยใช้ตัวแปรจากวิธี Stepwise

รายการ	กลุ่มตัวอย่างทั้งหมด			กลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD		
	R squared	RMSE	MAPE	R squared	RMSE	MAPE
MLR stepwise	0.2884	1.1563	0.1258	0.3299	1.1160	0.1205
SVM stepwise	0.3356	1.1246	0.1204	0.3528	1.1014	0.1172
RF stepwise	0.3240	1.1277	0.1242	0.3649	1.0909	0.1204
MARS stepwise	0.2753	1.1669	0.1268	0.3500	1.0993	0.1887
DNN stepwise	0.3179	1.0638	0.1112	0.3263	1.0962	0.1210
XG boost stepwise	0.3283	1.1291	0.1225	0.3570	1.0806	0.1177

จากการทดสอบแบบจำลองที่ได้จากการคัดเลือกตัวแปรด้วยวิธี Stepwise จะพบว่าในกลุ่มตัวอย่างทั้งหมด แบบจำลองที่มีประสิทธิภาพมากที่สุดคือ Support Vector Machine 33.56% รองลงมาคือ Random forest 32.40%, XG boost 32.83% แต่ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ยังเป็นแบบจำลองชนิด Random forest 36.49%, XG boost 35.70% ที่มีประสิทธิภาพมากที่สุดคล้ายกับผลการทดลองในการใช้ตัวแปรทั้งหมดในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD

นอกจากนี้ค่า R squared ของกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ยังสูงมากกว่าในกลุ่มตัวอย่างทั้งหมดอีกด้วย

พัฒนาแบบจำลอง

เพื่อเพิ่มประสิทธิภาพในการทำนายค่า HbA1c จะทดลองแบ่งช่วงข้อมูลโดยอ้างอิงจากงานวิจัยของ Abbas Bahrapour และ Saiedeh Haji-Maghsoudi ได้ศึกษาปัจจัยที่มีผลต่อค่า HbA1c ของประชากรชาวอิหร่านด้วยเทคนิค Linear Mixed Quantile Regression โดยจะแบ่งข้อมูลค่า HbA1c เป็นช่วง 5 ช่วงคือ

Quantile 5th มีค่า HbA1c อยู่ในช่วง 5.0 - 5.7

Quantile 25th มีค่า HbA1c อยู่ในช่วง 6.2 - 6.7

Quantile 50th มีค่า HbA1c อยู่ในช่วง 6.8 - 8.0

Quantile 75th มีค่า HbA1c อยู่ในช่วง 7.8 - 9.4

Quantile 95th มีค่า HbA1c อยู่ในช่วง 9.9 - 11.7

จากการศึกษาพบว่าปัจจัยตัวแปรต้นจะมีความสัมพันธ์ต่อตัวแปรตามที่แตกต่างกันในแต่ละช่วง Quantile เช่น ค่า BMI จะพบว่าไม่มีความสัมพันธ์ที่มีนัยสำคัญต่อช่วง Quantile ที่ต่ำ แต่พบความสัมพันธ์เชิงบวกที่มีนัยสำคัญต่อ Quantile ที่สูง 50th ขึ้นไป และ อายุ แสดงความสัมพันธ์เชิงลบที่มีนัยสำคัญต่อ Quantile ที่สูง 50th ขึ้นไป

จากแนวคิดดังกล่าวผู้วิจัยจึงได้แบ่งข้อมูลออกเป็นช่วงของค่า HbA1c ที่แตกต่างกันคือ ค่า HbA1c < 6.5 % (กลุ่มตัวอย่าง 1,449 ราย), HbA1c 6.5 - 9.9 % (กลุ่มตัวอย่าง 2,648 ราย), HbA1c ≥ 10.0 % (กลุ่มตัวอย่าง 190 ราย) โดยแบ่งจากกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD จากนั้นนำมาหาความสัมพันธ์ด้วยวิธี Pearson และ Spearman เมื่อพิจารณาระดับความสัมพันธ์และนัยสำคัญแล้วพบว่าตัวแปรต้นในแต่ละช่วงมีความสัมพันธ์ต่อค่า HbA1c ที่แตกต่างกันออกไป ดังนั้นจึงต้องพิจารณาหาช่วงข้อมูลที่เหมาะสมต่อไป

จากตารางที่ 6 เมื่อพิจารณาถึงความสัมพันธ์ที่มีนัยสำคัญ (p-value < 0.05) จะพบว่ามีเพียงตัวแปรต้นบางตัวเท่านั้นที่มีความสัมพันธ์กับตัวแปรตาม ได้แก่ อายุ, ค่า BMI, FBS, LDL, Cholesterol, Triglyceride, Family history, Medical count ซึ่งตัวแปรที่มีความสัมพันธ์มากที่สุดคือค่า FBS 0.1416 ในกรณีนี้การสร้างแบบจำลองจะได้แบบจำลองที่มีประสิทธิภาพน้อยลงกว่าเดิม

ตารางที่ 6 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c < 6.5 %

ตัวแปรต้น	วิธี Pearson	วิธี Spearman
Age	0.0560 (p-value = 0.03313)	0.0295 (p-value = 0.2617)
Sex	-0.0411 (p-value = 0.1179)	-0.0413 (p-value = 0.1159)
BMI	0.0698 (p-value = 0.007821)	0.0768 (p-value = 0.003421)
FBS	0.1416 (p-value = 6.189e-08)	0.1785 (p-value = 7.745e-12)
BUN	0.0477 (p-value = 0.06944)	0.0463 (p-value = 0.07838)
Creatinine	-0.0123 (p-value = 0.6407)	-0.0098 (p-value = 0.7096)
eGFR	-0.0335 (p-value = 0.2029)	-0.0266 (p-value = 0.3111)
HDL	0.0103 (p-value = 0.6941)	0.0128 (p-value = 0.6261)
LDL	0.0696 (p-value = 0.00801)	0.0596 (p-value = 0.02316)
Cholesterol	0.0866 (p-value = 0.0009637)	0.0749 (p-value = 0.004317)
Triglyceride	0.0561 (p-value = 0.03276)	0.0574 (p-value = 0.02883)
Smoking	0.0315 (p-value = 0.2302)	0.0525 (p-value = 0.04572)
HT combination	0.0412 (p-value = 0.1172)	0.0285 (p-value = 0.2776)
Stress	0.0331 (p-value = 0.2081)	0.0346 (p-value = 0.1874)
Family history	0.0752 (p-value = 0.004193)	0.0720 (p-value = 0.006076)
Metformin used	0.0399 (p-value = 0.1291)	0.0256 (p-value = 0.3301)
Sulfonyl urea used	0.0416 (p-value = 0.1131)	0.0584 (p-value = 0.0261)
Thiazolidinedione used	-0.0034 (p-value = 0.8962)	0.0080 (p-value = 0.7616)
Insulin used	0.0180 (p-value = 0.4932)	0.0294 (p-value = 0.2636)
Medical count	0.0565 (p-value = 0.03162)	0.0712 (p-value = 0.006732)

ตารางที่ 7 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c $\geq 10\%$

ตัวแปรต้น	วิธี Pearson	วิธี Spearman
Age	0.1396 (p-value = 0.05467)	0.1424 (p-value = 0.05002)
Sex	-0.0458 (p-value = 0.53)	-0.0483 (p-value = 0.5079)
BMI	-0.0843 (p-value = 0.2473)	-0.0644 (p-value = 0.3774)
FBS	0.0561 (p-value = 0.4422)	0.0784 (p-value = 0.2821)
BUN	-0.0202 (p-value = 0.7821)	-0.0227 (p-value = 0.7558)
Creatinine	-0.0796 (p-value = 0.275)	-0.0850 (p-value = 0.2435)
eGFR	0.0039 (p-value = 0.9576)	-0.0143 (p-value = 0.8449)
HDL	0.0508 (p-value = 0.4867)	0.0476 (p-value = 0.5141)
LDL	0.0839 (p-value = 0.2499)	0.0450 (p-value = 0.5372)
Cholesterol	0.1258 (p-value = 0.08378)	0.0795 (p-value = 0.2758)
Triglyceride	0.0966 (p-value = 0.1849)	0.1140 (p-value = 0.1174)
Smoking	0.0189 (p-value = 0.7957)	0.0136 (p-value = 0.8524)
HT combination	0.0171 (p-value = 0.8145)	0.0130 (p-value = 0.8587)
Stress	-0.0264 (p-value = 0.718)	-0.0334 (p-value = 0.6473)
Family history	-0.0412 (p-value = 0.5722)	-0.0587 (p-value = 0.4213)
Metformin used	-0.0167 (p-value = 0.8189)	-0.0036 (p-value = 0.9604)
Sulfonyl urea used	0.0481 (p-value = 0.5102)	0.0523 (p-value = 0.4739)
Thiazo used	0.0129 (p-value = 0.8593)	0.0276 (p-value = 0.7053)
Insulin used	0.0313 (p-value = 0.6681)	0.0259 (p-value = 0.7227)
Medical count	0.0612 (p-value = 0.4019)	0.0890 (p-value = 0.222)

จากตารางที่ 7 พบว่าตัวแปรต้นและตัวแปรตามมีความสัมพันธ์ที่น้อยมาก และไม่มีตัวแปรตัวไหนที่มี ความสัมพันธ์ที่มีนัยสำคัญ (p-value < 0.05) อาจเนื่องมาจากกลุ่มตัวอย่างในช่วงนี้มีน้อยมากจึงทำให้ความสัมพันธ์ระหว่างตัวแปรไม่แสดงออกมาอย่างชัดเจน

ตารางที่ 8 ตารางแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c 6.5 – 9.9 %

ตัวแปรต้น	วิธี Pearson	วิธี Spearman
Age	-0.0683 (p-value = 0.0004354)	-0.0643 (p-value = 0.0009336)
Sex	-0.0650 (p-value = 0.0008101)	-0.0704 (p-value = 0.0002868)
BMI	0.0414 (p-value = 0.03317)	0.0316 (p-value = 0.1045)
FBS	0.3480 (p-value < 2.2e-16)	0.3554 (p-value < 2.2e-16)
BUN	0.0237 (p-value = 0.223)	0.0253 (p-value = 0.193)
Creatinine	-0.0766 (p-value = 8.007e-05)	-0.0741 (p-value = 0.0001354)
eGFR	0.0733 (p-value = 0.0001585)	0.0771 (p-value = 7.188e-05)
HDL	-0.0740 (p-value = 0.0001369)	-0.0624 (p-value = 0.001306)
LDL	0.0368 (p-value = 0.05864)	0.0281 (p-value = 0.1487)
Cholesterol	0.0275 (p-value = 0.1568)	0.0216 (p-value = 0.2656)
Triglyceride	0.0338 (p-value = 0.08188)	0.0319 (p-value = 0.1002)
Smoking	0.0375 (p-value = 0.05384)	0.0344 (p-value = 0.07649)
HT combination	-0.0154 (p-value = 0.4271)	-0.0096 (p-value = 0.6225)
Stress	0.0251 (p-value = 0.1964)	0.0274 (p-value = 0.1579)
Family history	0.0772 (p-value = 6.921e-05)	0.0775 (p-value = 6.571e-05)
Metformin used	-0.0441 (p-value = 0.02331)	-0.0342 (p-value = 0.07887)
Sulfonyl urea used	0.0470 (p-value = 0.0156)	0.0598 (p-value = 0.002066)
Thiazo used	0.0380 (p-value = 0.05032)	0.0414 (p-value = 0.03303)
Insulin used	0.2447 (p-value < 2.2e-16)	0.2342 (p-value < 2.2e-16)
Medical count	0.1827 (p-value < 2.2e-16)	0.1918 (p-value < 2.2e-16)

จากตารางที่ 8 จะพบว่าตัวแปรที่มีระดับความสัมพันธ์สูงที่สุดคือค่า FBS 0.3480 รองลงมา คือ การได้รับยา insulin (insulin used) 0.2447, จำนวนชนิดยาที่ได้รับ (medical count) 0.1827 ความสัมพันธ์ 3 ลำดับแรกยังเหมือนกับกลุ่มตัวอย่างที่ยังไม่แบ่งช่วง แต่เมื่อพิจารณาถึงนัยสำคัญ พบว่ามีการเปลี่ยนแปลงไป ซึ่งตัวแปรที่ไม่มีความสัมพันธ์แบบมีนัยสำคัญได้แก่ ค่า BUN, LDL, Cholesterol, Triglyceride, Smoking, HT combination, Stress และ MFM used มีตัวแปรที่ไม่มีนัยสำคัญเพิ่มมากขึ้น

ตารางที่ 9 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ในช่วงข้อมูล HbA1c 6.5 – 9.9 %

รายการ	กลุ่มตัวใช้ตัวแปรทั้งหมด			กลุ่มตัวอย่างใช้เฉพาะตัวแปรที่มีนัยสำคัญ		
	R squared	RMSE	MAPE	R squared	RMSE	MAPE
MLR	0.1701	0.8077	0.0842	0.1676	0.8089	0.0843
SVM	0.1616	0.8333	0.0837	0.1717	0.8278	0.0821
RF	0.2087	0.7888	0.0839	0.1839	0.8024	0.0849
MARS	0.1583	0.8154	0.0855	0.1549	0.8173	0.0857
DNN	0.1729	0.8198	0.0863	0.1811	0.8129	0.0860
XG boost	0.2067	0.7935	0.0835	0.1948	0.7996	0.0846

จากตารางที่ 9 ได้ทดลองสร้างแบบจำลองโดยเปรียบเทียบระหว่างการใช้ตัวแปรทั้งหมด 20 ตัวแปร และกลุ่มตัวอย่างที่ใช้เฉพาะตัวแปรที่มีนัยสำคัญ 12 ตัวแปร โดยมีจำนวนตัวอย่างทั้งหมด 2,648 ราย เมื่อเปรียบเทียบค่า R squared พบว่าแบบจำลองที่ยังมีประสิทธิภาพดีที่สุดคือ Random forest และ XG boost แต่ประสิทธิภาพในการทำนายหรือค่า R squared ลดลงในทุก ๆ แบบจำลอง ตัวอย่างเช่น แบบจำลอง Random forest จากเดิม 37.26 % (ตารางที่ 4) ลดลงเหลือ 20.87% (ตารางที่ 9) แบบจำลอง XG boost จากเดิม 35.60 % (ตารางที่ 4) ลดลงเหลือ 20.67% (ตารางที่ 9) แต่การแบ่งช่วงข้อมูลก็ยังพบข้อดีนั่นคือพบว่าค่า RMSE และค่า MAPE ลดลง

เมื่อเปรียบเทียบการลดจำนวนตัวแปรและใช้เฉพาะตัวแปรต้นที่มีนัยสำคัญมาสร้างแบบจำลอง พบว่าประสิทธิภาพในการทำนายของหลาย ๆ แบบจำลองลดลง เช่น MLR, Random forest, MARS, XG boost แต่แบบจำลอง Support vector machine และ Deep neural network กลับเพิ่มขึ้นแต่มีการเพิ่มเพียงเล็กน้อยมาก

ตารางที่ 10 ตารางแสดงประสิทธิภาพของแบบจำลองสำหรับทำนายแต่ละชนิดในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ซึ่งแบ่งเป็นกลุ่มย่อย 2 กลุ่ม ในช่วงข้อมูล HbA1C 6.5 – 9.9 %

รายการ	กลุ่มย่อยที่ 1			กลุ่มย่อยที่ 2		
	R squared	RMSE	MAPE	R squared	RMSE	MAPE
MLR	0.1616	0.8613	0.0898	0.1549	0.8366	0.0888
SVM	0.1415	0.8891	0.0887	0.1891	0.8208	0.0835
RF	0.2173	0.8303	0.0867	0.1894	0.8190	0.0884
MARS	0.1547	0.8680	0.0899	0.1643	0.8316	0.0879
DNN	0.1237	0.8596	0.0850	0.1049	0.8564	0.0921
XG boost	0.2466	0.7683	0.0784	0.1534	0.7857	0.0809

จากตารางที่ 10 แสดงผลลัพธ์ของการแบ่งกลุ่มย่อยเท่า ๆ กัน พบว่าแบบจำลองที่ได้แสดงถึงความแปรปรวนของประสิทธิภาพในการทำนาย ในข้อมูลกลุ่มย่อยที่ 1 แบบจำลองที่ดีที่สุดคือ Random forest และ XG boost แต่ในกลุ่มย่อยชุดที่ 2 แบบจำลองที่ดีที่สุด คือ Random forest และ Support vector machine

บทที่ 5

สรุปผล อภิปรายผลและข้อเสนอแนะ

การวิจัยนี้เป็นการศึกษาวิจัยแบบกึ่งทดลอง (Quasi-experimental study research) เพื่อสร้างแบบจำลองสำหรับทำนายค่า HbA1c ของผู้ป่วยเบาหวานชนิดที่ 2 มีวัตถุประสงค์หลักเพื่อให้สามารถนำมาประกอบการพิจารณาการรักษาของแพทย์ได้ มีกลุ่มตัวอย่าง 46,727 ราย และตัวแปรที่ศึกษา 23 ตัวแปร เมื่อนำข้อมูลมาผ่านกระบวนการจัดการด้วยกระบวนการเหมืองข้อมูล (Data mining) จะเหลือกลุ่มตัวอย่างเพียง 5,067 ราย จากนั้นนำมาผ่านการคัดเลือกตัวแปรหลากหลายแบบ ได้แก่ หากความสัมพันธ์ด้วยวิธี Pearson และ Spearman, วิธีคัดเลือกตัวแปร Stepwise นอกจากนี้ยังมีการทดลองแบ่งกลุ่มตัวอย่างออกเป็นกลุ่มย่อยอีก 2 ลักษณะคือ พิจารณาตัดตัวอย่างที่เป็นผู้ป่วย CKD และพิจารณาการแบ่งช่วงค่า HbA1c เมื่อนำมาผ่านกระบวนการสร้างแบบจำลองทั้ง 6 ชนิด ได้แก่ Multiple linear regression, Support vector machine, Random forest, Multivariate Adaptive Regression Splines, Deep neural network และ Extreme gradient boost จากนั้นประเมินประสิทธิภาพแบบจำลองด้วยค่า R squared, RMSE, MAPE เพื่อหาแบบจำลองที่ดีที่สุด

สรุปผลการวิจัย

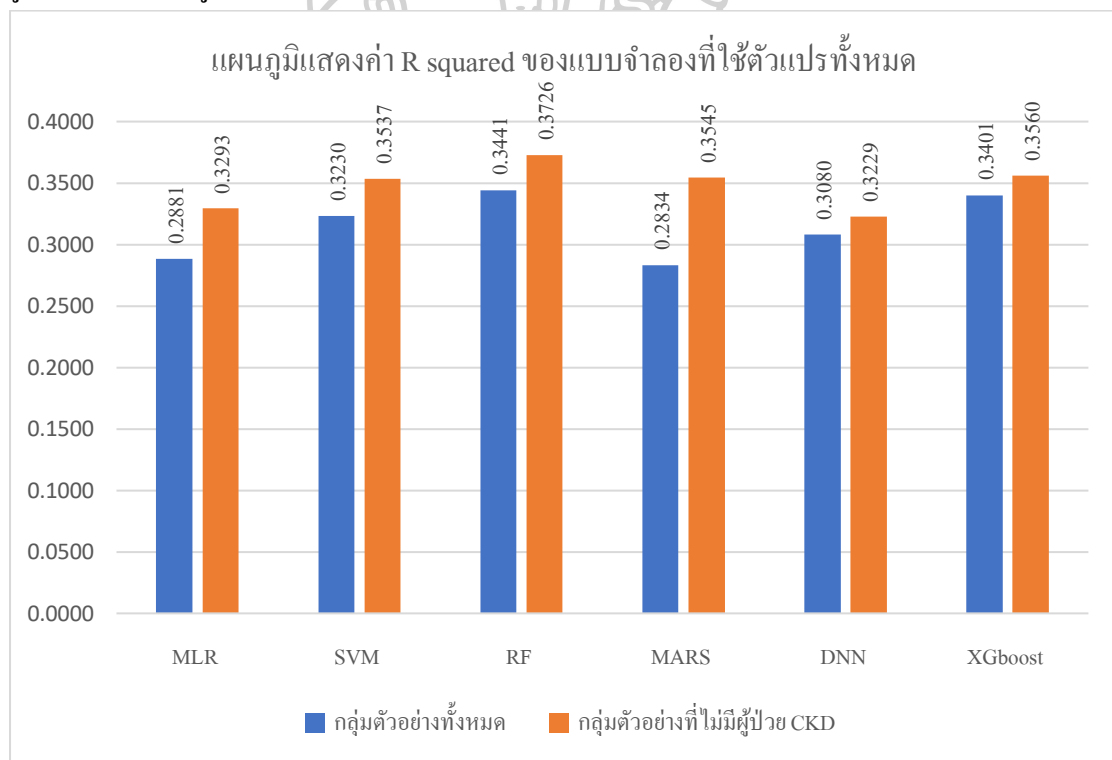
ผลของการคัดเลือกตัวแปรด้วยวิธี Pearson และ Spearman จะพบว่าตัวแปรต้นมีความสัมพันธ์กับตัวแปรตามในระดับที่ค่อนข้างน้อย โดยมีตัวแปรที่มีระดับความสัมพันธ์สูงสุดคือค่า FBS มีความสัมพันธ์เชิงบวก อยู่ที่ 0.4744 และตัวแปรที่ไม่มีความสัมพันธ์อย่างมีนัยสำคัญ กำหนด $p\text{-value} < 0.05$ คือตัวแปร ค่า SBP, ค่า BUN, ระดับการศึกษา และการได้รับยา Metformin

เมื่อเปรียบเทียบความสัมพันธ์ระหว่างกลุ่มตัวอย่าง 2 กลุ่มคือ กลุ่มตัวอย่างทั้งหมด และกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ด้วยวิธี Pearson และ Spearman พบว่า ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD มีระดับความสัมพันธ์ที่มากกว่ากลุ่มตัวอย่างทั้งหมด เมื่อเปรียบเทียบจะพบจำนวนตัวแปรที่มีระดับความสัมพันธ์ที่มากกว่า 12 ตัวแปรจากจำนวนตัวแปรต้นทั้งหมด 22 ตัวแปร โดยเฉพาะตัวแปรที่มีความสัมพันธ์สูง 3 ลำดับแรกได้แก่ ค่า FBS, การได้รับยา insulin (insulin used), จำนวนชนิดยาที่ได้รับ (medical count) และพบว่าตัวแปรที่ไม่มีนัยสำคัญเป็นตัวแปรเดียวกันกับกลุ่มผู้ป่วยทั้งหมด จากข้อมูลการคัดเลือกตัวแปรดังกล่าวข้างต้น จะพบว่าตัวแปรต้นที่สามารถตัดออกไปได้เลยคือค่า SBP และ ระดับการศึกษา เพราะไม่พบว่ามีความสัมพันธ์ที่มีนัยสำคัญไม่ว่าจะด้วยวิธีการทดสอบความสัมพันธ์ใดๆ

ผลของการคัดเลือกตัวแปรด้วยวิธี Stepwise จากกลุ่มตัวอย่าง 2 กลุ่มคือ กลุ่มตัวอย่างทั้งหมด และกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD จะพบว่าสองกลุ่มนี้มีตัวแปรที่ส่วนใหญ่แล้วเหมือนกัน ได้แก่ ตัวแปร Age, FBS, BUN, eGFR, HDL, Triglyceride, Med count, Smoking, HT combination, Family history, MFM used, SU used, Thiazolidinediones used แต่ในกลุ่มตัวอย่างทั้งหมดจะพบตัวแปรเพิ่มคือ ค่า Cholesterol และ Sex ส่วนกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD พบตัวแปรเพิ่มเติมคือ ค่า Creatinine, LDL และ Stress

จากข้อมูลการคัดเลือกตัวแปรดังกล่าวข้างต้น จะพบว่าตัวแปรต้นที่สามารถตัดออกไปได้เลยคือค่า SBP และ ระดับการศึกษา เพราะไม่พบว่ามีความสัมพันธ์ที่มีนัยสำคัญไม่ว่าจะด้วยวิธีการทดสอบความสัมพันธ์ใดๆ ดังนั้นในการทดลองที่ใช้ตัวแปรทั้งหมด จะเหลือตัวแปรต้น 20 ตัวแปร

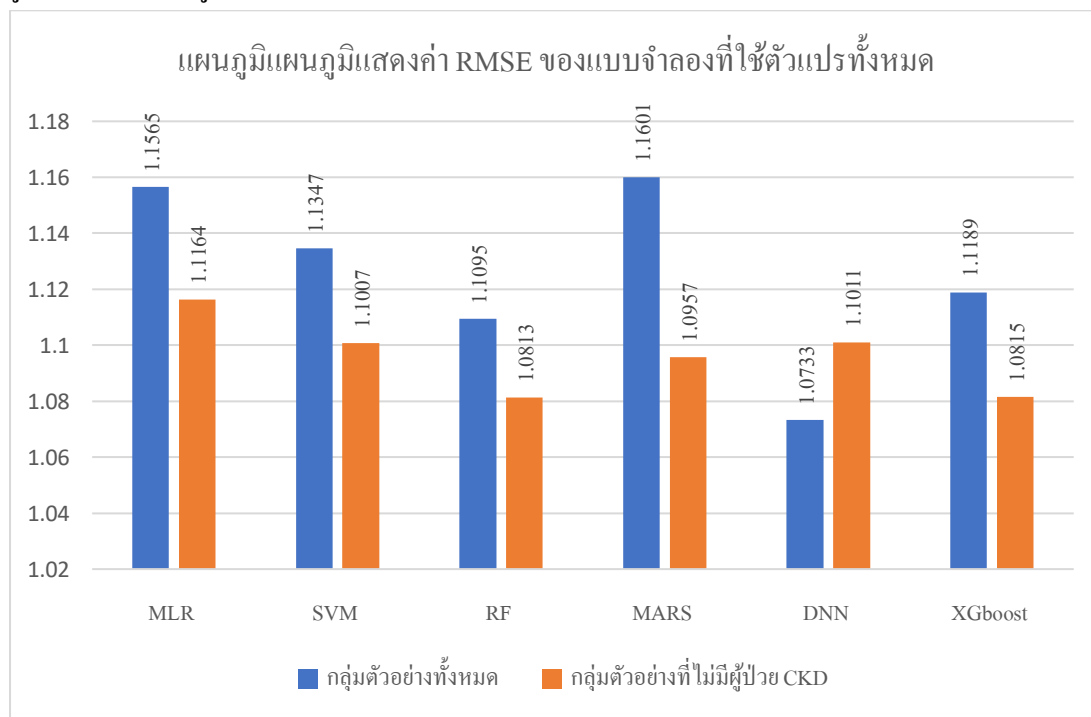
รูปภาพที่ 4 แผนภูมิแสดงค่า R squared ของแบบจำลองที่ใช้ตัวแปรทั้งหมด



จากรูปภาพที่ 4 แสดงการเปรียบเทียบค่า R squared กลุ่มตัวอย่าง 2 กลุ่มคือ กลุ่มตัวอย่างทั้งหมด และกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD โดยใช้ตัวแปรที่ผ่านกระบวนการคัดเลือกตัวแปรทั้งหมด 20 ตัวแปร จะพบว่าในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD แบบจำลองทุกชนิดจะมีค่า R

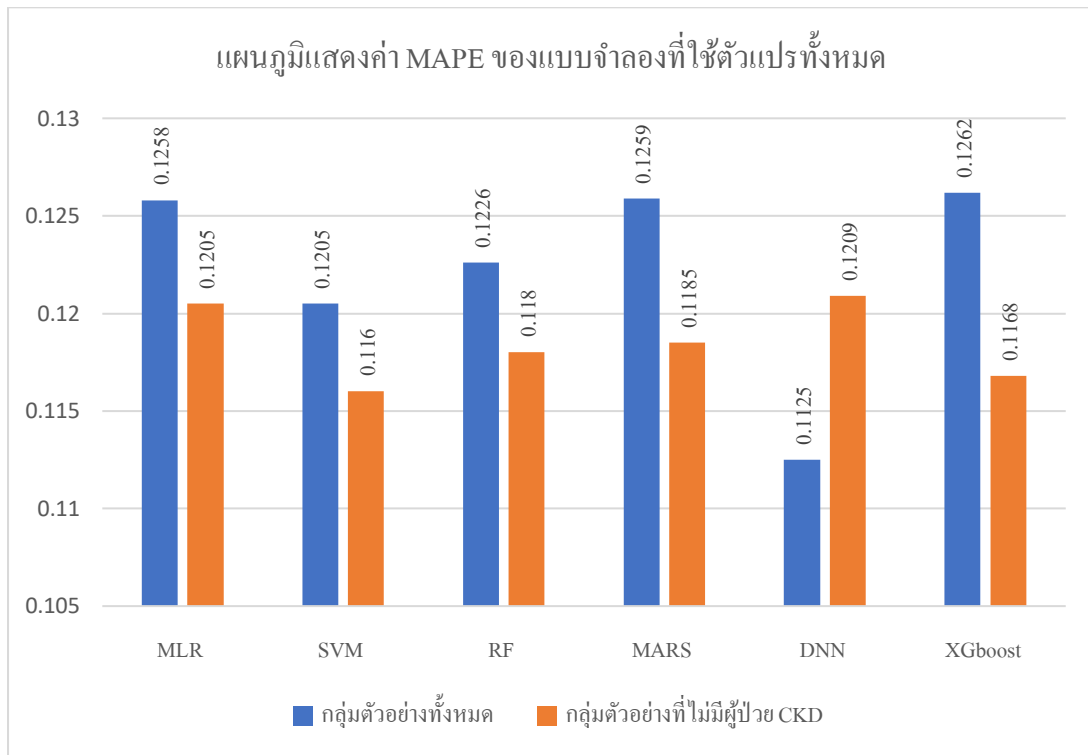
squared ที่มากกว่า โดยในกลุ่มตัวอย่างทั้งหมดแบบจำลองที่มีประสิทธิภาพในการทำนายมากที่สุดคือ Random forest 34.41% และ XGboost 34.01% ส่วนกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD แบบจำลองที่มีประสิทธิภาพในการทำนายมากที่สุดคือ Random forest 37.26% และ XGboost 35.60%

รูปภาพที่ 5 แผนภูมิแสดงค่า RMSE ของแบบจำลองที่ใช้ตัวแปรทั้งหมด



จากรูปภาพที่ 5 เมื่อเปรียบเทียบกันแล้วค่า RMSE ของกลุ่มตัวอย่างทั้งหมด โดยส่วนมากจะมีค่ามากกว่า กลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD โดยเฉพาะแบบจำลองชนิด MARS มีค่า RMSE สูงที่สุด รองลงมาคือ MLR และ SVM ในกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ค่า RMSE จะลดลงอาจเนื่องมาจากแบบจำลองมีประสิทธิภาพมากขึ้น มีค่า R squared สูงขึ้น ยกเว้นแบบจำลองชนิด Deep Neural Network ที่มีค่า RMSE เพิ่มมากขึ้น

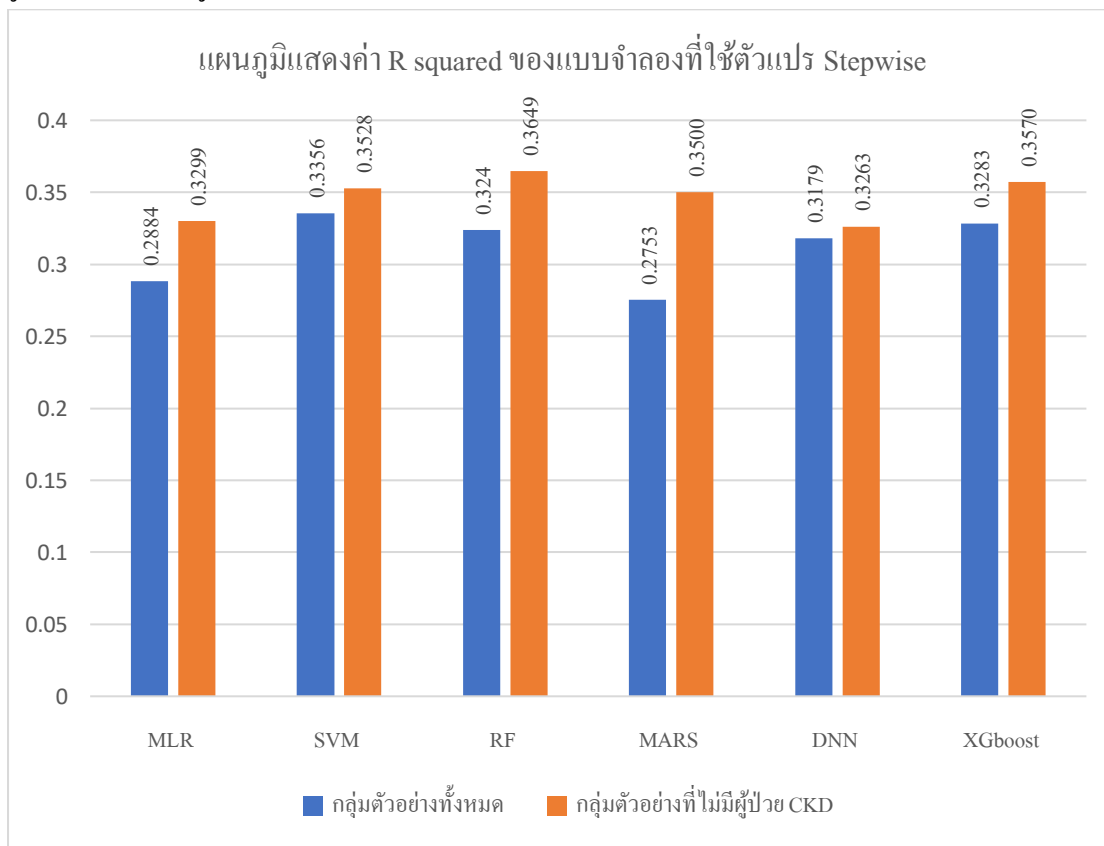
รูปภาพที่ 6 แผนภูมิแสดงค่า MAPE ของแบบจำลองที่ใช้ตัวแปรทั้งหมด



จากรูปภาพที่ 6 แสดงค่า MAPE ที่มีลักษณะเหมือนค่า RMSE นั่นคือกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD ค่า RMSE จะลดลงในทุก ๆ แบบจำลองยกเว้นแบบจำลองชนิด Deep Neural Network ที่มีค่าเพิ่มมากขึ้นสวนทางกับแบบจำลองชนิดอื่น ๆ



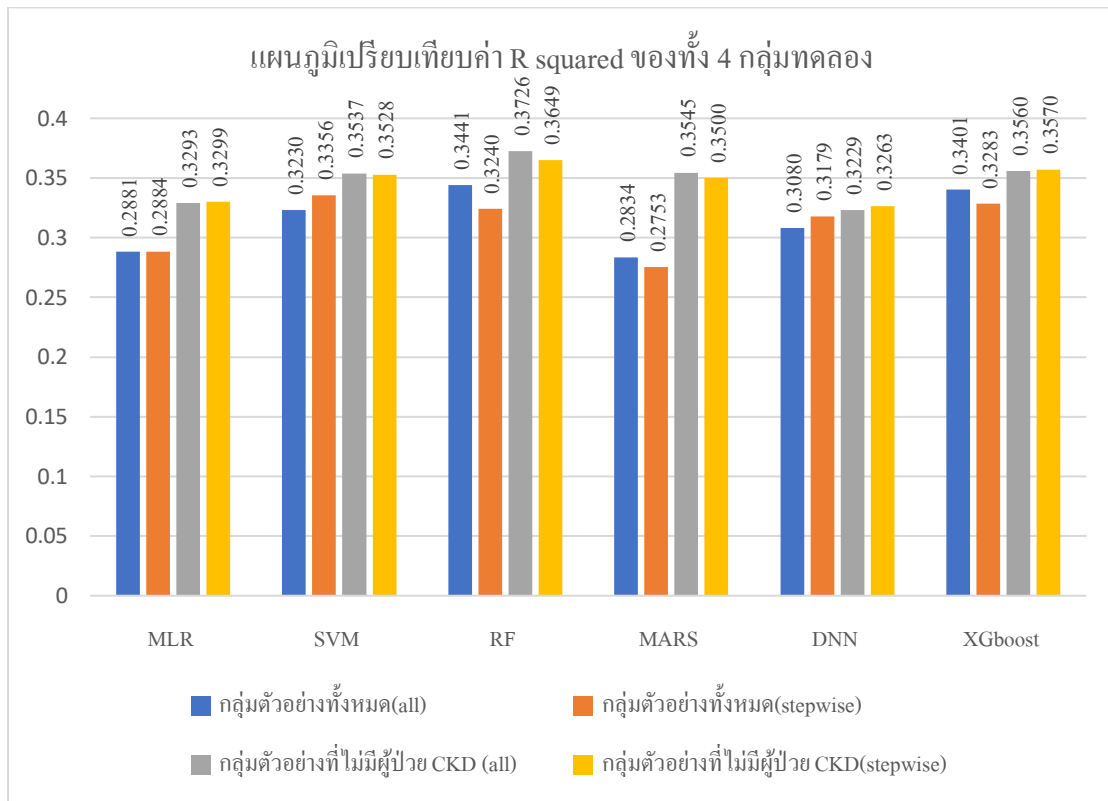
รูปภาพที่ 7 แผนภูมิแสดงค่า R squared ของแบบจำลองที่ใช้ตัวแปร Stepwise



จากรูปที่ 7 เมื่อทดลองสร้างแบบจำลองจากการใช้ตัวแปรที่ผ่านกระบวนการ Stepwise โดยเปรียบเทียบระหว่างกลุ่มตัวอย่าง 2 กลุ่ม พบว่าในกลุ่มตัวอย่างทั้งหมด Support Vector Machine จะได้ค่า R squared ที่มากที่สุด รองลงมาคือ Random Forest และ XGboost ส่วนกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD แบบจำลองที่ดีที่สุดยังเป็น Random Forest, XGboost และ Support Vector Machine อยู่ในลำดับที่ 3

เมื่อพิจารณาภาพรวม กลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD จะมีค่า R squared ที่สูงขึ้นในแบบจำลองทุกชนิด ซึ่งผลลัพธ์ออกมาเหมือนกับกลุ่มการทดลองที่ใช้ตัวแปรทั้งหมด สามารถเปรียบเทียบทั้ง 4 กลุ่มทดลองได้ตามรูปภาพที่ 8

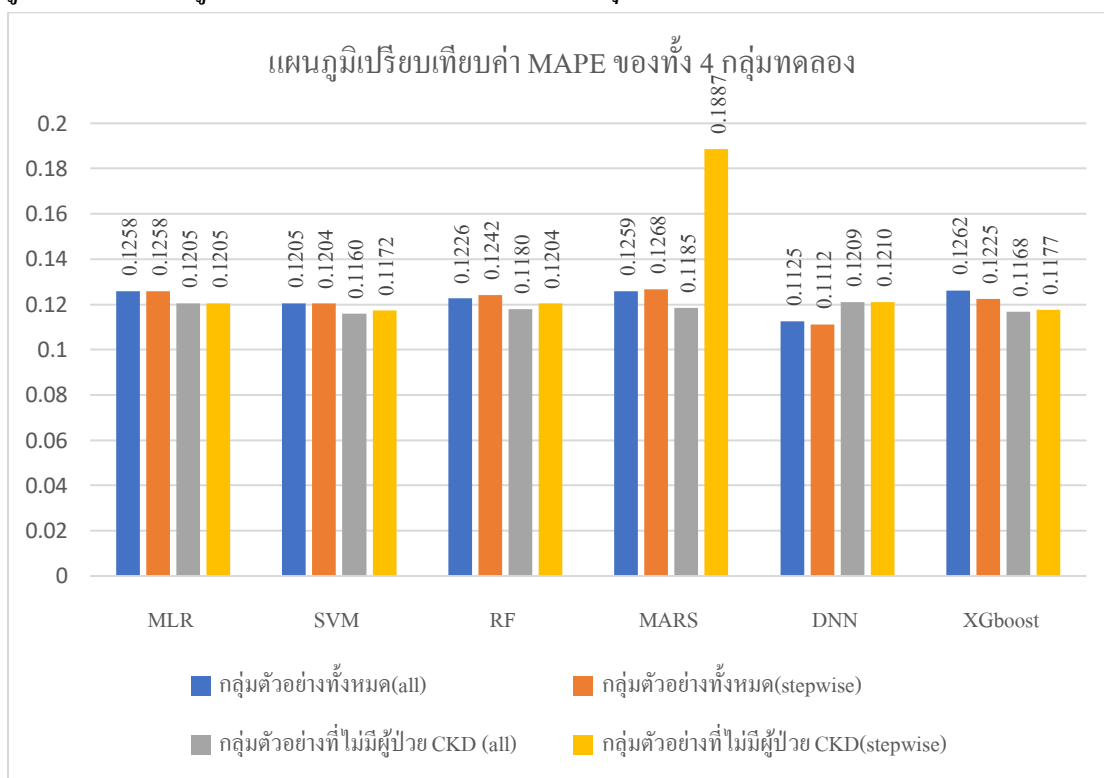
รูปภาพที่ 8 แผนภูมิเปรียบเทียบค่า R^2 ของทั้ง 4 กลุ่มทดลอง



จากรูปที่ 8 แสดงว่ากลุ่มตัวอย่างที่เหมาะสมสำหรับการนำไปสร้างแบบจำลองคือกลุ่มตัวอย่างที่คัดกรองผู้ป่วยที่มีโรคร่วม CKD ออก และในกระบวนการคัดเลือกตัวแปรการใช้วิธี stepwise อาจไม่เหมาะสม เพราะแบบจำลองบางชนิดจะมีประสิทธิภาพลดลง เช่น Random forest, MARS เป็นต้น

เนื่องจากเป็นแบบจำลองชนิด Regression มุ่งเน้นผลลัพธ์ที่เป็นจำนวนตัวเลข และการใช้ข้อมูลจาก real world data ทั่วไปแล้วแบบจำลองที่สร้างได้จะมีประสิทธิภาพในการทำนายที่ต่ำมาก ดังนั้นนอกจากจะพิจารณาค่า R^2 แสดงประสิทธิภาพในการทำนายและค่า RMSE แสดงความผิดพลาดในการทำนายแล้วยังมีการนำค่า MAPE มาใช้ประกอบการพิจารณาผลที่ได้จากแบบจำลองด้วย และในบางครั้งจะสามารถอธิบายผลลัพธ์ได้ง่ายมากยิ่งขึ้น

รูปภาพที่ 9 แผนภูมิเปรียบเทียบค่า MAPE ของทั้ง 4 กลุ่มทดลอง

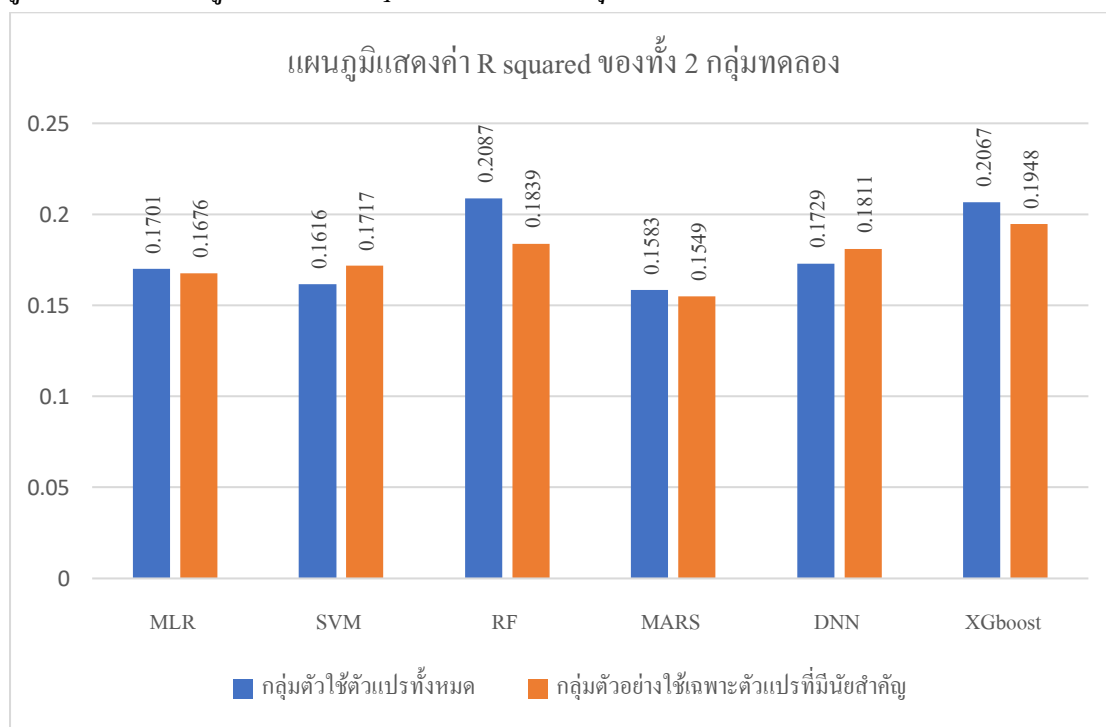


จากรูปที่ 4 แสดงค่า MAPE โดยทั่วไปค่า MAPE จะอยู่ประมาณ 0.12 ยกเว้น ค่า MAPE ของแบบจำลอง MARS ที่สูงมากกว่าแบบจำลองชนิดอื่น

จากการทดลองที่ผ่านมาสามารถสรุปได้ว่าแบบจำลองที่มีประสิทธิภาพดีที่สุดสำหรับประมาณค่า HbA1c คือ Random Forest และ XG boost และค่าที่ประมาณได้จะมีค่าแปรปรวนอยู่ประมาณ +/- 12% จากค่าจริง

จากการทบทวนวรรณกรรมพบว่าหากแบ่งช่วงค่า HbA1c เป็นช่วงที่แคบมากขึ้นอาจทำให้แบบจำลองมีประสิทธิภาพมากขึ้น เมื่อทำการแบ่งกลุ่มตัวอย่างตามช่วงค่า HbA1c ทำให้ทราบว่าในกลุ่มตัวอย่างที่มีค่า HbA1c <6% และ $\geq 10\%$ ความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามลดลงและบางตัวไม่มีนัยสำคัญต่อตัวแปรตามเลย ข้อมูลใน 2 กลุ่มนี้อาจไม่เหมาะสมในการนำมาสร้างแบบจำลอง จึงได้ทดลองต่อในกลุ่มทดลองใหม่ที่มาจากกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD และมีค่า HbA1c อยู่ในช่วง 6.5 - 9.9%

รูปภาพที่ 10 แผนภูมิแสดงค่า R squared ของทั้ง 2 กลุ่มทดลอง



จากรูปภาพที่ 10 จะพบว่าแบบจำลองที่ยังมีประสิทธิภาพสูงสุดคือ Random forest และ XG boost แต่จำนวนตัวแปรที่ลดลงทำให้ค่า R squared ลดลงด้วย นอกจากนี้ค่า MAPE ยังลดลงเหลือประมาณ 8 % จากตารางที่ 9 แสดงว่าค่าประมาณค่า HbA1c ที่ได้ มีความใกล้เคียงค่าจริง +/- 8%

และเมื่อทดลองแบ่งกลุ่มย่อยออกไปอีกตามตารางที่ 10 แสดงให้เห็นว่าแบบจำลองที่ได้มีประสิทธิภาพลดลงเรื่อย ๆ เมื่อขนาดของกลุ่มตัวอย่างลดลง

อภิปรายผล

การวิจัยนี้มีจุดมุ่งหมายในการสร้างแบบจำลองสำหรับทำนายค่า HbA1c ของผู้ป่วยเบาหวานชนิดที่ 2 เพื่อหาชนิดของแบบจำลองที่ดีที่สุดจึงได้ทดลองสร้างแบบจำลองทั้งหมด 6 ชนิด ได้แก่ Multiple linear regression, Support vector machine, Random forest, Multivariate Adaptive Regression Splines, Deep neural network และ Extreme gradient boost จะพบว่าแบบจำลองที่ดีที่สุดคือ Random Forest และ XG boost เนื่องจากเมื่อนำไปทดลองสร้างในสถานะที่แตกต่างกัน ไม่ว่าจะเป็นขนาดกลุ่มตัวอย่างที่แตกต่าง จำนวนตัวแปรต้นที่แตกต่าง Random Forest และ XG boost ยังให้ค่า R squared ที่สูงมากกว่าแบบจำลองอื่น

การคัดกรองกลุ่มตัวอย่างมีความจำเป็น เมื่อทดลองเปรียบเทียบกลุ่มทดลองที่ใช้กลุ่มตัวอย่างทั้งหมดเทียบกับกลุ่มตัวอย่างที่ไม่มีผู้ป่วย CKD กลุ่มหลังเมื่อนำมาสร้างแบบจำลองจะให้ค่า R squared ที่สูงกว่า ค่า RMSE และ MAPE ที่ต่ำกว่า

ขนาดของกลุ่มตัวอย่างสามารถส่งผลต่อประสิทธิภาพของแบบจำลองได้ จะเห็นได้ว่าเมื่อตัดทอนกลุ่มตัวอย่างให้เล็กลงเรื่อย ๆ ค่า R squared ก็จะลดลงตามไปด้วย

การคัดเลือกตัวแปรไม่สามารถใช้แค่วิธีการเดียว จะเห็นได้จาก ตัวแปรต้นที่เมื่อคัดกรองด้วยวิธี Pearson และ Spearman ไม่มีนัยสำคัญ แต่เมื่อนำมาผ่านวิธี stepwise กลับมีการนำมาใช้ เช่น ตัวแปรการได้รับยา Metformin และ ค่า BUN

จำนวนตัวแปรต้นมีผลต่อประสิทธิภาพของแบบจำลองอย่างมาก ในการตัดตัวแปรบางตัวออกไปจะทำให้ประสิทธิภาพของแบบจำลองลดลง

ข้อเสนอแนะ

จากผลลัพธ์ที่ได้จากการศึกษาการประมาณค่า HbA1c ผ่านการสร้างแบบจำลองทั้ง 6 ชนิด ที่มีคุณลักษณะแตกต่างกัน ผู้วิจัยขอเสนอแนะเพื่อปรับปรุงประสิทธิภาพของงานวิจัยต่อไปดังนี้ ในการวิจัยนี้มีจำนวนข้อมูลค่อนข้างน้อย ทำให้จำนวนข้อมูลที่นำมาใช้ในการเรียนรู้สำหรับสร้างแบบจำลองอาจไม่เพียงพอ หากสามารถเพิ่มจำนวนข้อมูลได้มากกว่านี้อาจสามารถสร้างแบบจำลองสำหรับทำนายให้มีประสิทธิภาพมากยิ่งขึ้น

ในการวิจัยนี้ไม่สามารถเก็บตัวแปรได้ครบถ้วน เช่น ไม่สามารถเก็บตัวแปรของระดับรายได้ ระยะเวลาในการป่วยเป็นเบาหวาน ค่า Albumin จากห้องปฏิบัติการ ทำให้ตัวแปรต้นที่สามารถนำมาใช้ประมวลการเรียนรู้ด้วยเครื่องลดลง นอกจากนี้อาจจะต้องกำหนดกลุ่มตัวอย่างที่เฉพาะเจาะจงมากยิ่งขึ้น ไม่มีกลุ่มเสี่ยงที่อาจทำให้ค่า HbA1c มีความแปรปรวน เช่น ผู้ป่วยที่มีภาวะ Anemia, ผู้ป่วยที่ติดเชื้อเรื้อรัง เป็นต้น

การสร้างแบบจำลองชนิด Regression โดยใช้ข้อมูลจาก real world data มีการศึกษาค่อนข้างน้อย ทำให้ขาดข้อมูลอ้างอิงที่หลากหลายและน่าเชื่อถือ ส่วนมากจะพบการศึกษาชนิด Classification ซึ่งอาจจะพิจารณาในการสร้างแบบจำลองชนิด Classification ทดแทน



รายการอ้างอิง

- ADA, A. D. A. (2022). Standards of Medical Care in Diabetes—2022 Abridged for Primary Care Providers. *Clinical Diabetes*, 40(1), 10-38. <https://doi.org/10.2337/cd22-as01>
- Bergental, R. M., Beck, R. W., Close, K. L., Grunberger, G., Sacks, D. B., Kowalski, A., Brown, A. S., Heinemann, L., Aleppo, G., Ryan, D. B., Riddlesworth, T. D., & Cefalu, W. T. (2018). Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring. *Diabetes Care*, 41(11), 2275-2280. <https://doi.org/10.2337/dc18-1581>
- Chaki J., T. G. S., Cidham S.K., Ananda Theertan S. (2020). Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: A systematic review. *J. King Saud Univ.-Comput. Inf. Sci.*
- Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., & Bellazzi, R. (2018). Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol*, 12(2), 295-302. <https://doi.org/10.1177/1932296817706375>
- Del Parigi, A., Tang, W., Liu, D., Lee, C., & Pratley, R. (2019). Machine Learning to Identify Predictors of Glycemic Control in Type 2 Diabetes: An Analysis of Target HbA1c Reduction Using Empagliflozin/Linagliptin Data. *Pharmaceut Med*, 33(3), 209-217. <https://doi.org/10.1007/s40290-019-00281-4>
- Fan, Y., Long, E., Cai, L., Cao, Q., Wu, X., & Tong, R. (2021). Machine Learning Approaches to Predict Risks of Diabetic Complications and Poor Glycemic Control in Nonadherent Type 2 Diabetes. *Front Pharmacol*, 12, 665951. <https://doi.org/10.3389/fphar.2021.665951>
- Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & Garcia-Garcia, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr*, 13(1), 148. <https://doi.org/10.1186/s13098-021-00767-9>
- Gautier, T., Ziegler, L. B., Gerber, M. S., Campos-Náñez, E., & Patek, S. D. (2021). Artificial intelligence and diabetes technology: A review. *Metabolism*, 124, 154872. <https://doi.org/10.1016/j.metabol.2021.154872>
- Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. *Healthcare (Basel)*, 9(12).

<https://doi.org/10.3390/healthcare9121712>

K.C., s. R. (2018). Diabetes Mellitus And Glycosylated Hemoglobin A1c. *Nepalese Medical Journal*, 1, 112. <https://doi.org/10.3126/nmj.v1i2.21744>

Kjellsson, M. C., Cosson, V. F., Mazer, N. A., Frey, N., & Karlsson, M. O. (2013). A model-based approach to predict longitudinal HbA1c, using early phase glucose data from type 2 diabetes mellitus patients after anti-diabetic treatment. *J Clin Pharmacol*, 53(6), 589-600. <https://doi.org/10.1002/jcph.86>

Kodama, S., Fujihara, K., Shiozaki, H., Horikawa, C., Yamada, M. H., Sato, T., Yaguchi, Y., Yamamoto, M., Kitazawa, M., Iwanaga, M., Matsubayashi, Y., & Sone, H. (2021). Ability of Current Machine Learning Algorithms to Predict and Detect Hypoglycemia in Patients With Diabetes Mellitus: Meta-analysis. *JMIR Diabetes*, 6(1), e22458. <https://doi.org/10.2196/22458>

Lu, R., Duan, T., Wang, M., Liu, H., Feng, S., Gong, X., Wang, H., Wang, J., Cui, Z., Liu, Y., Li, C., & Ma, J. (2021). The application of multivariate adaptive regression splines in exploring the influencing factors and predicting the prevalence of HbA1c improvement. *Ann Palliat Med*, 10(2), 1296-1303. <https://doi.org/10.21037/apm-19-406>

Nagaraj, S. B., Sidorenkov, G., van Boven, J. F. M., & Denig, P. (2019). Predicting short- and long-term glycosylated haemoglobin response after insulin initiation in patients with type 2 diabetes mellitus using machine-learning algorithms. *Diabetes Obes Metab*, 21(12), 2704-2711. <https://doi.org/10.1111/dom.13860>

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *J Biomed Inform*, 89, 56-67. <https://doi.org/10.1016/j.jbi.2018.09.001>

Nkoana, M. K., & Khine, A. A. (2020). Correlation between self-monitored mean blood glucose and average plasma glucose estimated from glycosylated haemoglobin in patients attending the diabetes clinic at Dr George Mukhari Academic Hospital, Pretoria, South Africa. *S Afr Med J*, 110(5), 416-421. <https://doi.org/10.7196/SAMI.2020.v110i5.13959>

Ooka, T., Johno, H., Nakamoto, K., Yoda, Y., Yokomichi, H., & Yamagata, Z. (2021). Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. *BMJ Nutrition, Prevention & Health*,

- bmjnp-2020-000200. <https://doi.org/10.1136/bmjnp-2020-000200>
- Radin, M. S. (2014). Pitfalls in hemoglobin A1c measurement: when results may be misleading. *J Gen Intern Med*, 29(2), 388-394. <https://doi.org/10.1007/s11606-013-2595-x>
- Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina (Kaunas)*, 56(9).
<https://doi.org/10.3390/medicina56090455>
- Shruthi Soundara Rajan, A. M., Harish Rangareddy. (2020). Calculated Glycosylated Hemoglobin (HbA1c) Compared with Estimated HbA1c by Nephelometry and Its Correlation to Estimated Average Blood Glucose (eAG). *Galore International Journal of Health Sciences and Research*, 5(4), 65-71.
- Silva, K., Lee, W. K., Forbes, A., Demmer, R. T., Barton, C., & Enticott, J. (2020). Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *Int J Med Inform*, 143, 104268.
<https://doi.org/10.1016/j.ijmedinf.2020.104268>
- Sirithansakul, T. (2021). Evaluation of Fasting Plasma Glucose Control with HbA1c of Patients with Type 2 Diabetes Mellitus and Type 2 Diabetic Patients with Thalassemia Anemia and Abnormal Hemoglobin. *Thai Journal of Public Health and Health Sciences*, 4 No.1(January-April,2021).
- Temsch, W., Luger, A., & Riedl, M. (2008). HbA1c values calculated from blood glucose levels using truncated Fourier series and implementation in standard SQL database language. *Methods Inf Med*, 47(4), 346-355. <https://doi.org/10.3414/me0468>
- WuQiang Fan, H. Z., Nancy Wei, David M. Nathan. (2018). Estimating HbA1c from timed Self-Monitored Blood Glucose values. *Diabetes Research and Clinical Practice*, 141, 56-61.
<https://doi.org/doi.org/10.1016/j.diabres.2018.04.023>.
- Xu, Y., Hirota, Y., Ajjan, R. A., Yamamoto, A., Matsuoka, A., Ogawa, W., & Dunn, T. C. (2021). Accurate prediction of HbA1c by continuous glucose monitoring using a kinetic model with patient-specific parameters for red blood cell lifespan and glucose uptake. *Diab Vasc Dis Res*, 18(3), 14791641211013734. <https://doi.org/10.1177/14791641211013734>
- กรมควบคุมโรค, ก. (2021). รายงานผลการทบทวนผลกระทบของ COVID-19 ต่อระบบบริการโรคไม่

ติดต่อเรื้อง.

สมาคมโรคเบาหวานแห่งประเทศไทย. (2017). แนวทางเวชปฏิบัติสำหรับโรคเบาหวาน 2560.





ประวัติผู้เขียน

ชื่อ-สกุล

ผ่องเพ็ญ คำผิว

วุฒิการศึกษา

พ.ศ. 2552 เกษศาสตรบัณฑิต คณะเกษตรศาสตร์ มหาวิทยาลัยศิลปากร

