การจำแนกโรคเบาหวานจากผลตรวจเลือดด้วยวิธีการทำเหมืองข้อมูล
IDENTIFYING DIABETES FROM BLOOD TEST RESULTS USING DATA MINING METHODS

เชียน หวัง [1*] วุฒิชัย ช่างคิด [2] ศัลย์ ชูภาพ [3] และ วิภาวรรณ บัวทอง [4]
Qian Wang[1]*, Wutthichai Changkhit[2], San Choopap[3] and Wipawan Buathong[4]
สังกัดสาขาวิชาเทคโนโลยีดิจิทัล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏภูเก็ต [1*,2,3,4]
*Corresponding author. E-mail: s6481423101@pkru.ac.th

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองที่มีประสิทธิภาพ และค้นหาคุณลักษณะที่สามารถบ่งชี้การเป็นโรคเบาหวาน ด้วยวิธีการทำเหมืองข้อมูลแบบจำแนกประเภทข้อมูล ชุดข้อมูลที่ใช้ในการวิจัยนำมาจากฐานข้อมูลผู้ป่วยของโรงพยาบาลแห่งหนึ่ง ทั้งหมด 15,000 รายการ 14 แอตทริบิวต์ เมื่อทำความสะอาดและลดมิติข้อมูลด้วยวิธี Wrapper approach แบบ Forward Selection ได้ชุดข้อมูลที่สมบูรณ์ 11,658 รายการ 10 แอตทริ-บิวต์ หลังจากนั้น นำชุดข้อมูลมาประมวลผลโดยโปรแกรม Weka 3.8.5 โดยใช้อัลกอริทึมทั้ง 5 อัลกอริทึม คืออัลกอริทึมต้นไม้ตัดสินใจ อัลกอริทึมแรนดอมฟอเรสต์ อัลกอริทึมเบย์อย่างง่าย อัลกอริทึมเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น และใช้วิธีการแบ่งข้อมูลสำหรับฝึกฝนและทดสอบ แบบ Cross-Validation 10 Folds และ Percentage 70% ผลการวิจัยพบว่าอัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น ให้ค่าประสิทธิภาพของแบบจำลองที่ใช้วิธีการ Cross-Validation 10 Folds สูงที่สุด คือ ค่าความถูกต้อง (Accuracy) เท่ากับ 90.82% ค่าความแม่นยำ (Precision) เท่ากับ 90.50% ค่าระลึก (Recall) เท่ากับ 90.80% และค่าถ่วงดุล (F-Measure) เท่ากับ 90.60%

คำสำคัญ: โรคเบาหวาน ผลตรวจเลือด การทำเหมืองข้อมูล วิธีการจำแนกประเภท

Abstract

This research aims to create an effective model and look for features that can indicate diabetes with a classified data mining method. Used datasets from a patient database of a hospital with a total of 15,000 records and 14 attributes. After cleaning and dimension reduction data with the Wrapper Approach by Forward Selection, we have a total of 11,658 records with 10 attributes. After that, the dataset is processed by Weka 3.8.5. and used 5 algorithms; Decision tree, Random Forest, Naïve Bay, K-Nearest Neighbor and a Multilayer Perceptron algorithm. Used a method of dividing the data for the training and testing model with Cross-validation 10 folds and 70% Percentage. The results showed that the Multilayer Perceptron algorithm gives the highest performance value when

using the Cross-Validation 10 Folds method. Accuracy is 90.82%, Precision is 90.50%, Recall is 90.80%, and F-Measure is 90.60%.

Keywords: Diabetes, Blood test results, Data Mining, Classification

บทนำ

ปัจจุบันโรคเบาหวานเป็นโรคไม่ติดต่อเรื้อรังที่เป็นปัญหาสุขภาพอันดับหนึ่งของโลก ซึ่งมีผู้ป่วยและผู้เสียชีวิตเป็นจำนวนมาก จากการรายงานข้อมูลขององค์การอนามัยโลก พบว่าในปี 2012 ประชากรทั่วโลก เสียชีวิตจากโรคเบาหวาน 1.5 ล้านคน (Gojka Roglic & World Health Organization, 2016) สำหรับสถานการณ์โรคไม่ติดต่อในประเทศไทย พบว่าความชุกของโรคเบาหวานในประชากรอายุ 18 ปีขึ้นไป เพิ่มขึ้นจากร้อยละ 6.9 ในปี พ.ศ.2552 เป็นร้อยละ 8.9 ในปี พ.ศ.2557 (อรรถเกียรติ กาญจนพิบูลวงศ์, ภาณุวัฒน์ คำวังสง่า และสุธิดา แก้วทา, 2563) และจากรายงาน IDF Diabetes Atlas 10th edition 2021 พบว่าประเทศไทยอยู่ลำดับที่ 4 ใน 5 ของประเทศที่มีผู้ป่วยเบาหวานมากที่สุดในภูมิภาคเอเชียตะวันออก (International Diabetes Federation, 2021) การตรวจคัดกรองเบาหวานจะช่วยลดความเสี่ยงของผู้เป็นโรค และผู้ที่มีประวัติครอบครัวเป็นโรค อีกทั้งยังช่วยยับยั้งโรคแทรกซ้อนที่อาจจะตามมากับผู้สูงอายุ ซึ่งในปัจจุบันมีวิธีการตรวจเบาหวานหลายวิธี คือ 1. ผู้ที่มีอาการของโรคชัดเจน คือ มีอาการหิวน้ำบ่อย ปัสสาวะบ่อย น้ำหนักลดลงโดยไม่ทราบสาเหตุ เมื่อตรวจระดับกลูโคสมีค่ามากกว่าหรือเท่ากับ 200 มิลลิกรัม/เดซิลิตร ให้วินิจฉัยว่าเป็นโรค 2. ตรวจระดับกลูโคสตอนเช้าหลังอดอาหารมากกว่า 8 ชั่วโมง หากมีค่ามากกว่าหรือเท่ากับ 126 มิลลิกรัม/เดซิลิตร ให้วินิจฉัยว่าเป็นโรค วิธีนี้เหมาะกับคนทั่วไปที่มาตรวจสุขภาพและไม่มีอาการ 3. ตรวจความทนต่อกลูโคส (75 กรัม Oral Glucose Tolerance Test, OGTT) คือ ดื่มน้ำตาล 75 กรัมถ้าระดับกลูโคสในเวลา 2 ชั่วโมงต่อมา มากกว่าหรือเท่ากับ 200 มิลลิกรัม/เดซิลิตร ให้วินิจฉัยว่าเป็นโรค แต่วิธีนี้มีความคลาดเคลื่อนสูง 4. ตรวจระดับฮีโมโกลบินในเลือด (HbA1C) หากมีค่ามากกว่าหรือเท่ากับร้อยละ 6.5 ให้วินิจฉัยว่าเป็นโรค (สมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี, 2560) ซึ่งวิธีนี้นิยมใช้มากขึ้นในปัจจุบัน เพราะไม่จำเป็นต้องอดอาหาร แต่เป็นการวัดระดับน้ำตาลสะสมเฉลี่ยใน 3 เดือนที่ผ่านมา จากการศึกษาความเสี่ยงของเพศชายที่ไม่ได้เป็นโรคเบาหวาน พบว่าผู้ที่มีระดับน้ำตาลก่อนอาหารอยู่ในช่วง 91-99 มิลลิกรัม/เดซิลิตร ร่วมกับมีระดับไขมันไตรกลีเซอไรด์มากกว่าหรือเท่ากับ 150 มิลลิกรัม/เดซิลิตร จะมีความเสี่ยงต่อการเกิดโรคประมาณ 8 เท่า เมื่อเปรียบเทียบกับบุคคลที่มีระดับน้ำตาลก่อนอาหารในช่วงที่น้อยกว่า 86 มิลลิกรัม/เดซิลิตร และมีระดับไขมันไตรกลีเซอไรด์ที่น้อยกว่า 150 มิลลิกรัม/เดซิลิตร นอกจากนี้ผู้ที่มีระดับน้ำตาลก่อนอาหารอยู่ในช่วง 91-99 มิลลิกรัม/เดซิลิตร ร่วมกับมีดัชนีมวลกายมากกว่าหรือเท่ากับ 30 กิโลกรัม/ตารางเมตร จะมีความเสี่ยงต่อการเกิดโรคประมาณ 8 เท่า เมื่อเปรียบเทียบกับบุคคลที่มีระดับน้ำตาลก่อนอาหารในช่วงที่น้อยกว่า 86 มิลลิกรัม/เดซิลิตร และมีระดับดัชนีมวลกายที่น้อยกว่า 25 กิโลกรัม/ตารางเมตร (สมเกียรติ โพธิสัตย์ และคนอื่น ๆ, 2557, น. 6-3) ดังนั้นบุคคลที่อ้วนหรือมีระดับไขมันไตรกลีเซอไรด์สูง ถึงแม้ว่าจะมีระดับน้ำตาลในเลือดปกติ ก็มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่ดำเนินการกับข้อมูลจำนวนมาก เพื่อหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลมาใช้ได้ (Han & Kamber, 2006) เทคนิคการทำเหมืองข้อมูลมีหลายแบบ ในงานวิจัยฉบับนี้

เลือกใช้แบบการจำแนกประเภท (Classification) เป็นการสร้างโมเดลสำหรับจำแนกประเภทของข้อมูล จากคุณสมบัติ (attribute) เพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด (class) โดยต้องมีข้อมูลบางส่วนใช้สำหรับฝึกสอน และอีกส่วนหนึ่งสำหรับทดสอบซึ่งใช้อัลกอริทึม 5 ชนิดในการสร้างแบบจำลอง ได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ อัลกอริทึมแรนดอมฟอเรสต์ อัลกอริทึมเบย์อย่างง่าย อัลกอริทึมเพื่อนบ้านใกล้สุด อัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น

อัลกอริทึมต้นไม้ตัดสินใจ (Decision tree) เป็นอัลกอริทึมที่ใช้ในงานด้านเหมืองข้อมูลประเภทการจำแนกข้อมูล เป็นวิธีที่ได้รับความนิยมสูง เนื่องจากเป็นอัลกอริทึมไม่ซับซ้อน อุกฤษฎ์ ศรีสุข และจารี ทองคำ ได้ใช้อัลกอริทึมต้นไม้ตัดสินใจ C4.5 สร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์ โดยผลการทดลองมีความถูกต้องในการพยากรณ์ถึง 99.86% (อุกฤษฎ์ ศรีสุข และ จารี ทองคำ, 2564) นอกจากนี้ รุ่งโรจน์ บุญมา และ นิเวศ จิระวิชิตชัย ได้ใช้ต้นไม้ตัดสินใจ ทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวาน ได้ผลการทำนายค่าความถูกต้อง โดยเฉลี่ยอยู่ที่ 68.62% (บุญมา รุ่งโรจน์ และ จิระวิชิตชัย นิเวศ, 2563) การพยากรณ์ของต้นไม้ตัดสินใจ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) หมายถึงต้องทราบกลุ่มที่ชัดเจนก่อนทำการประมวลผล รูปแบบของต้นไม้ตัดสินใจประกอบด้วยโหนดต่าง ๆ เริ่มจากโหนดแรกเรียกว่ารูทโหนด (Root node) จากนั้นแยกออกเป็นโหนดลูก(Decision node) แยกถึงระดับสุดท้ายเรียกว่า โหนดใบ (Leaf node) สำหรับงานวิจัยฉบับนี้ เลือกใช้อัลกอริทึม J48 ในโปรแกรม Weka ซึ่งเป็นอัลกอริทึมที่พัฒนาโดย Ross Quinlan โดย J48 จะสร้างผังต้นไม้ตัดสินใจจากกลุ่มข้อมูลชุดฝึกสอน โดยใช้ค่า Gain Ratio (Pang-Ning Tan, Michael Steinbach, 1981) ในการประมวลผล ดังสมการ

$$\text{Gain Ratio}\,(S,V) = \frac{\text{Gain}\,(S,V)}{\text{SplitInfo}\,(S,V)}$$

$$\text{SplitInfo}\,(S,V) = \sum_{i=1}^{m} -\frac{|S_i|}{S} \times \log_2 \frac{|S_i|}{S}$$

อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) เป็นแบบจำลองที่ใช้พื้นฐานจากอัลกอริทึมต้นไม้ตัดสินใจ (Aurélien Géron, 2019) เป็นการทำนายแบบชุดของต้นไม้ตัดสินใจหลายๆ ต้น (Ensemble of Decision Trees) โดยสร้างจากการสุ่มข้อมูลตัวอย่างแบบเลือกแล้วใส่กลับ (random sampling with replacement) เพื่อนำมาสร้างเป็นแบบจำลองต้นไม้ โดยแต่ละต้นมีลักษณะที่ไม่ซ้ำกัน ข้อดีของ Random Forest คือ ระยะเวลาในการจำแนกผลลัพธ์ที่สั้น เนื่องจากเป็น Decision Tree ที่ข้างในประกอบไปด้วยเงื่อนไข if-Else แต่มีระยะเวลาในการสร้างแบบจำลองที่นาน (ดำรงเดช เดินรีบรัมย์, ฉัตรเกล้า เจริญผล และจริยา จิรานุกูล , 2563)

อัลกอริทึมเบย์อย่างง่าย (Naïve Bayes) เป็นรูปแบบการหาความสัมพันธ์ที่ไม่ซับซ้อนและได้ผลลัพธ์ดี ใช้วิเคราะห์หาความน่าจะเป็นของเหตุการณ์ที่ยังไม่เคยเกิดขึ้น โดยคาดเดาจากเหตุการณ์ที่เคยเกิดขึ้นมาก่อน รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย ได้ใช้อัลกอริทึมเบย์อย่างง่าย ในการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวาน โดยมีผลการทำนายค่าความถูกต้อง โดยเฉลี่ย 75.59% (รุ่งโรจน์ บุญมา, และ นิเวศ จิระวิชิตชัย, 2563) อัลกอริทึมนี้ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem (Roiger & Geatz, 2003) ตัวอย่าง ถ้ากำหนดให้ P(h) เป็นความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ P(h|D) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D จากตัวแปรที่กำหนดไว้ สามารถทำนายเหตุการณ์ได้จากสมการ

$$P(h \mid D) = [P(D \mid h) * P(h)]/P(D)$$

จากสมการข้างต้น สามารถคำนวณความน่าจะเป็นในการจำแนก ได้จากสมการ

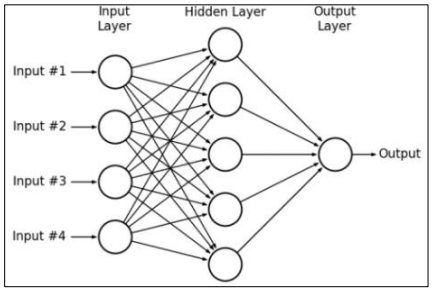
$$P(d \mid h) = P(a_1, \ldots, a_T \mid h) = \prod_t P(a_t \mid h)$$

อัลกอริทึมเพื่อนบ้านใกล้ที่สุด (K-nearest neighbors) เป็นการแบ่งกลุ่มข้อมูล และทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนาย กับข้อมูลที่อยู่ใกล้เคียง เป็นจำนวน K ตัว คำตอบที่ได้มาคือ คลาสที่พบมากที่สุดของข้อมูล ที่เป็นเพื่อนบ้านทั้ง K ตัว เทคนิคนี้มักจะใช้วิธีการวัดระยะห่างแบบ Euclidean ดังสมการ

$$\text{Euclidean Distance} = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2)}$$

โดยที่ x1 คือ แอททริบิวต์ที่ 1 ของข้อมูลชุดที่ 1 และ y1 คือ แอททริบิวต์ที่ 1 ของข้อมูลชุดที่ 2 โดยข้อมูลทั้ง 2 ตัว (x และ y) มีจำนวนแอททริบิวต์เท่ากับ L พงศกร ธีรรัศมี ได้ศึกษาวิธีหาค่า K ที่เหมาะสมในการจำแนกแบบเพื่อนบ้านใกล้ที่สุดกับข้อมูลทางการแพทย์ 5 ชุดข้อมูลคือ ข้อมูลผู้ป่วยโรคหอบหืด โรคหัวใจ โรคมะเร็งเต้านม โรคไทรอยด์ โรคเบาหวาน ซึ่งผลการวิจัยค่า K ที่เหมาะสม คือ 10% จากจำนวนข้อมูล 1, 1, 1, 10% จากจำนวนข้อมูลตามลำดับ และได้ค่าความถูกต้อง คือ 81%, 74%, 99%, 93%, 77% ตามลำดับ (พงศกร ธีรรัศมี, 2558)

อัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น (Multilayer Perceptron) เป็นหนึ่งในประเภทของโครงข่ายประสาทเทียมที่จะต้องเชื่อมต่อนิวรอลมากกว่าหนึ่งตัวเข้าด้วยกัน เพื่อให้เกิดเป็นลักษณะของโครงข่ายที่มีชั้นหรือเลเยอร์ (Layer) ตั้งแต่ 2 ชั้นขึ้นไป สายชล สินสมบูรณ์ทอง ได้ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน พบว่าวิธีโครงข่ายประสาทเทียมหลายชั้น มีค่าความถูกต้องมากที่สุด คือ 95.94% (สินสมบูรณ์ทอง, 2561) โครงข่ายประสาทเทียมหลายชั้น ประกอบด้วยเลเยอร์สามประเภท ได้แก่เลเยอร์อินพุตสำหรับรับข้อมูล, เลเยอร์เอาต์พุตสำหรับการประมวลงาน เช่น การทำนายและการจัดประเภท และเลเยอร์ที่ซ่อนอยู่ระหว่างอินพุตและเอาต์พุตนั้น มีหน้าที่คำนวณ (Abirami & Chitra, 2020) ดังแสดงในภาพที่ 1



ภาพที่ 1 องค์ประกอบของโครงข่ายประสาทเทียมหลายชั้น

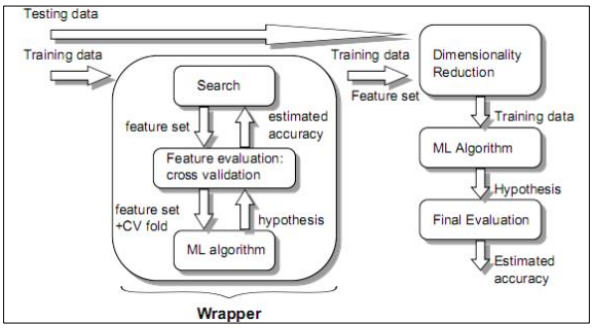
ที่มา: Hassan et al., 2015

การคำนวณที่เกิดขึ้นในทุกเซลล์ประสาทในเอาต์พุตและเลเยอร์ที่ซ่อนอยู่มี ดังนี้

$$o(x) = G(b(2) + W(2)h(x))$$
$$h(x) = \Phi(x) = s(b(1) + W(1)x)$$

จากอัลกอริทึมทั้ง 5 ที่กล่าวมาข้างต้น จะไม่สามารถบอกได้ว่าเทคนิคใดดีที่สุด ดังนั้นจึงควรทดลองสร้างแบบจำลองหลายๆ ครั้ง แล้วนำไปเปรียบเทียบเพื่อดูว่าแบบจำลองใดมีประสิทธิภาพมากที่สุด จึงจะนำไปใช้ประโยชน์ต่อไป

นอกจากนี้ผู้วิจัยยังใช้วิธีการลดมิติข้อมูลด้วยวิธี Wrapper Approach แบบ Forward Selection เป็นวิธีคัดเลือกคุณลักษณะที่สำคัญ (attributes) โดยการคำนวณค่าน้ำหนักกับการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มีการสร้างเซตของคุณลักษณะใหม่ โดยการเพิ่มจำนวนคุณลักษณะจากเซตเดิม การลดมิติข้อมูลด้วยวิธี Wrapper Approach แบบ Forward Selection ทำให้ได้ชุดย่อยของคุณลักษณะที่เหมาะสมที่สุด (Singh & Singh, 2021) ซึ่งเมื่อนำไปใช้ในการประมวลผลการจำแนกประเภท ทำให้ได้ความถูกต้องที่ดีขึ้น เมื่อทำการเปรียบเทียบระหว่างไม่เลือกคุณลักษณะของข้อมูล และเลือกคุณลักษณะของข้อมูล พบว่าการเลือกคุณลักษณะของข้อมูลโดยวิธี Forward Selection ให้ค่าความถูกต้องมากกว่า (รัชพล กลัดชื่น และจรัญ แสนราช, 2561)



ภาพที่ 2 เทคนิคการลดมิติข้อมูลด้วยวิธี Wrapper Approach แบบ Forward Selection
ที่มา: Hall, 1999

จากที่กล่าวมาในข้างต้น งานวิจัยฉบับนี้จึงได้รวบรวมข้อมูลผลตรวจเลือดของผู้ป่วย 2 กลุ่ม คือ ผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นโรคเบาหวานและผู้ป่วยทั่วไปที่ไม่ได้เป็นเบาหวาน นำมาวิเคราะห์และจำแนกโรคเบาหวานด้วยวิธีการทำเหมืองข้อมูลแบบจำแนกประเภท โดยเลือกใช้อัลกอริทึม 5 ประเภท ได้แก่ Decision tree, Random Forest, Naive Bayes, KNN และ Multilayer Perceptron เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยพิจารณาค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึกและค่าความถ่วงดุล เพื่อสร้างแบบจำลองสำหรับจำแนกผู้ป่วยได้อย่างมีประสิทธิภาพ นอกจากนี้ยังเป็นการนำความรู้ด้านเทคโนโลยีดิจิทัลมาประยุกต์ใช้กับด้านการแพทย์ ซึ่งจะช่วยเพิ่มแนวทางในการทำนายและวินิจฉัยการเกิดโรค สามารถลดภาระให้แก่บุคลากรทางการแพทย์ได้

วัตถุประสงค์ของการวิจัย

1. เพื่อค้นหาคุณลักษณะที่สามารถบ่งชี้การเป็นโรคเบาหวาน จากผลตรวจเลือดในห้องปฏิบัติการของโรงพยาบาลแห่งหนึ่ง ด้วยการทำเหมืองข้อมูลโดยใช้วิธีการจำแนกประเภท (Classification)
2. เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการจำแนกข้อมูลของโรคเบาหวาน

วิธีดำเนินการวิจัย

งานวิจัยฉบับนี้ ได้ใช้ข้อมูลจากฐานข้อมูลผู้ป่วยของโรงพยาบาลแห่งหนึ่งในจังหวัดภูเก็ต มีทั้งหมด 15,000 รายการ 14 แอตทริบิวต์ หลังจากทำความสะอาดและลดมิติข้อมูลแล้ว เหลือข้อมูลที่สามารถนำมาใช้งานได้จริง 11,658 รายการ 10 แอตทริบิวต์ ใช้โปรแกรมสำเร็จรูป Weka 3.8.5 ในการทำเหมืองข้อมูล และใช้โปรแกรม

ประเภทสเปรดชีต สำหรับทำความสะอาดข้อมูล ลดมิติข้อมูล และเตรียมข้อมูลก่อนประมวลผลจริง สำหรับงานวิจัยฉบับนี้ ได้เลือกใช้กระบวนการทำเหมืองข้อมูลตามขั้นตอน CRISP-DM โดยมีขั้นตอนทั้ง 6 ขั้นตอนดังนี้

1. การทำความเข้าใจธุรกิจ (Business Understanding) สำหรับงานวิจัยฉบับนี้ เป็นการจำแนกการเป็นโรคเบาหวานจากผลตรวจเลือดในห้องปฏิบัติการ โดยใช้วิธีการจำแนกประเภท โดยนำข้อมูลจากฐานข้อมูลผู้ป่วยของโรงพยาบาลแห่งหนึ่งมาใช้ในการวิจัย

2. การทำความเข้าใจข้อมูล (Data Understanding) เป็นกระบวนทำความเข้าใจข้อมูลที่ต้องการศึกษา สำหรับงานวิจัยฉบับนี้ ได้นำชุดข้อมูลผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นโรคเบาหวานและผู้ป่วยทั่วไป จำนวน 15,000 รายการ (Record) 14 แอตทริบิวต์ (Attribute)

3. การเตรียมข้อมูล (Data Preparation) เป็นกระบวนการเตรียมข้อมูลให้พร้อมก่อนทำการประมวลผลจริง ซึ่งขั้นตอนนี้ใช้เวลานานที่สุด สามารถแบ่งย่อยได้เป็น

3.1 Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออก คือ

- ตัดรายการที่มีเนื้อหาขาดหายไป (missing value) รวม 3,244 รายการ หลังจากตัดรายการที่มีเนื้อหาไม่ครบถ้วนแล้ว เหลือข้อมูลที่ใช้ได้ 11,756 รายการ

- ตัดรายการที่ผิดพลาด (error) หรือมีค่าผิดปกติ (Outliers) ออกจำนวณ 98 รายการ สุดท้ายเหลือข้อมูลที่สามารถประมวลผลกับโปรแกรม Weka 3.8.5 ได้จริง จำนวณ 11,658 รายการ

3.2 ทำการลดมิติข้อมูล (Dimensionality Reduction) เนื่องจากชุดข้อมูลเดิมมี 14 แอตทริบิวต์ ซึ่งบางแอตทริบิวต์ไม่ได้มีผลต่อการวิเคราะห์ ดังนั้นในงานวิจัยฉบับนี้ จึงได้ทำการลดมิติข้อมูล ด้วยวิธี Wrapper approach แบบ Forward Selection จากเดิมที่มี 14 แอตทริบิวต์ คือ Sex, Age, Bw, BMI, Potassium, Hba1c, FBS LDL, Cholesterol, Triglyceride, HDL, Creatinine, GFR และ DM หลังทำการลดมิติข้อมูล เหลือเพียง 10 แอตทริบิวต์ ที่ใช้ในการประมวลผล คือ BMI, HBA1C, FBS, Cholesterol, LDL, HDL,GFR, Triglyceride, Potassium, DM ดังคำอธิบายในตารางที่ 1 ข้างล่างนี้

ตารางที่ 1 รายละเอียดแอตทริบิวต์ที่ใช้ในการประมวลผล

แอตทริบิวต์	**ความหมาย**
BMI	**การหาค่าดัชนีมวลกาย**
HBA1C	ระดับโปรตีนฮีโมโกลบินในเซลล์เม็ดเลือดแดง ที่ถูกจับเกาะด้วยน้ำตาลกลูโคส
FBS	ค่าความเข้มข้นน้ำตาลในเลือด ณ เวลาที่เจาะเลือด ค่าปกติอยู่ที่ 74 – 106 mg/dL
Cholesterol	ไขมันชนิดหนึ่ง โดยค่าปกติภายในเลือดอยู่ที่ < 200 mg/dL
LDL	ไขมันชนิดเลว โดยค่าปกติภายในเลือดอยู่ที่ <130 mg/dL
HDL	ไขมันชนิดดี โดยค่าปกติของเพศชายอยู่ที่ 40-120 mg/dL เพศหญิงอยู่ที่ 50-120 mg/dL
GFR	อัตราการกรองที่หน่วยไตย่อย โดยค่าปกติภายในเลือดอยู่ที่ 90 – 120 ml/min/1.73 m^2
Triglyceride	เป็นไขมันที่ถูกเก็บสะสมในเซลล์ไขมัน โดยค่าปกติภายในเลือดอยู่ที่ < 150 mg/dL
Potassium	แร่ธาตุที่มีส่วนในการส่งสัญญาณไฟฟ้าไปยังเซลล์ โดยค่าปกติอยู่ที่ 3.6-5.2 mmol/L
DM	เป็นการแยกประเภทผู้ป่วยที่เป็นโรคเบาหวาน yes = เป็นเบาหวาน no = ไม่เป็นเบาหวาน

4. การสร้างแบบจำลอง (Modeling) เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่ โดยใช้เทคนิคและอัลกอริทึมที่เหมาะสม สำหรับงานวิจัยฉบับนี้ ได้เลือกใช้เทคนิคการจำแนกประเภท โดยเลือกใช้ 5 อัลกอริทึม ดังนี้ Decision tree J48, Random Forest, Naive Bayes, KNN และ Multilayer Perceptron

5. การวัดประสิทธิภาพของโมเดล (Evaluation) เป็นขั้นตอนประเมินผลว่าผลลัพธ์ที่ได้เหมาะสมหรือไม่ การวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล ที่นิยมใช้โดยทั่วไปจะมีอยู่ 4 ค่า คือ

- ค่าความถูกต้อง (Accuracy) เป็นจำนวนข้อมูลที่ทำนายถูกต้องทุกคลาส ดังสมการ

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- ค่าความแม่นยำ (Precision) เป็นค่าที่ดูสิ่งที่ทำนายออกมาแล้วถูกต้องกี่เปอร์เซ็นต์ ดังสมการ

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- ค่าความระลึก (Recall) เป็นค่าที่บ่งบอกว่า โมเดลที่ทำนายถูกต้องมีกี่ค่า ดังสมการ

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- ค่าความถ่วงดุล (F-measure) เป็นค่าเฉลี่ยของค่าความแม่นยำและค่าความระลึก ดังสมการ

$$\text{F} - \text{measure} = 2\,\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

สำหรับวิธีการแบ่งข้อมูลที่ใช้ในการ Training set และ Testing set จะใช้การแบ่งข้อมูลแบบ 10-folds Cross-Validation และ Percentage 70%

5.1 เทคนิคการทดสอบแบบไขว้ (k-fold Cross-validation Testing)

การทดสอบแบบไขว้ เป็นวิธีการวัดประสิทธิภาพของโมเดล โดยแบ่งข้อมูลออกเป็นหลายส่วน (แทนด้วย k-fold) เช่น 10-fold cross-validation คือ แบ่งข้อมูลออกเป็น 10 ส่วน แต่ละส่วนมีจำนวนข้อมูลเท่า ๆ กัน เลือกข้อมูล 9 ส่วนสำหรับฝึกสอนข้อมูล อีก 1 ส่วนที่เหลือสำหรับทดสอบประสิทธิภาพ จากนั้นวนการทำงาน ให้ข้อมูลทุกส่วนได้มีโอกาสเป็นชุดทดสอบ

5.2 เทคนิคการทดสอบแบบแยกชุด (Percentage Testing)

การทดสอบแบบแยกชุด เป็นการแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนแรกคือชุดสำหรับเรียนรู้เพื่อสร้างแบบจำลอง เรียกว่า ชุดฝึกสอน (Training Set) ส่วนที่สองใช้สำหรับทดสอบแบบจำลอง เรียกว่า ชุดทดสอบ (Testing Set) เช่น 70:30 หมายถึง แบ่งข้อมูลทั้งหมด ให้เป็นชุดฝึกสอน 70% แบ่งเป็นชุดทดสอบ 30% ข้อดีของเทคนิคนี้คือใช้เวลาน้อย เหมาะกับชุดข้อมูลขนาดใหญ่

6. การนำโมเดลไปใช้งานจริง (Deployment) สามารถนำข้อมูลที่ไม่เคยพบมาก่อน (Unseen Data) มาวิเคราะห์ หรือทำนายเพื่อสามารถจำแนกการเป็นโรคเบาหวานได้

สรุปผลการวิจัย

การทำเหมืองข้อมูลโดยใช้เทคนิคจำแนกประเภทข้อมูล จากข้อมูลเริ่มต้น 15,000 รายการ 14 แอตทริบิวต์ เมื่อทำความสะอาดข้อมูลแล้วเหลือข้อมูลที่ใช้งานได้จริง 11,658 รายการ จากนั้นได้ทำการลดมิติข้อมูลด้วยวิธี Wrapper approach แบบ Forward Selection ด้วยวิธี Cross-Validation 10 Folds พบว่าประสิทธิภาพของข้อมูลก่อนลดมิติข้อมูล ได้ค่าความถูกต้องอยู่ที่ 87.65% หลังจากทำการลดมิติข้อมูลเหลือคุณลักษณะที่สามารถบ่งชี้การเป็นโรคเบาหวานได้จำนวน 10 แอตทริบิวต์ คือ BMI, HBA1C, FBS, Cholesterol, LDL, HDL, GFR, Triglyceride, Potassium และ DM ได้ให้ค่าความถูกต้องของข้อมูลเพิ่มขึ้นเป็น 88.21% ดังตารางที่ 3 เมื่อได้คุณลักษณะหรือแอตทริบิวต์จากการลดมิติข้อมูล ผู้วิจัยจึงใช้อัลกอริทึมทั้ง 5 ชนิด ได้แก่ Decision tree ,Random Forest, Naive Bayes, K-NN และ Multilayer Perceptron เพื่อสร้างแบบจำลองและทำการวัดประสิทธิภาพของแบบจำลองด้วยวิธีการแบ่งข้อมูลแบบ Cross-Validation 10 Folds และ Percentage 70% ได้ผลการทดลองดังตารางที่ 4

ตารางที่ 3 ขั้นตอนและผลการลดมิติข้อมูล

ขั้นตอนที่	วิธีการ	แอตทริบิวต์ที่เลือกใช้	Accuracy
1	ทำการวัดประสิทธิภาพของแบบจำลอง ก่อนลดมิติข้อมูล	ใช้ทั้ง 14 แอตทริบิวต์	87.65%
2	ทำการลดมิติข้อมูล เริ่มจาก 1 แอตทริบิวต์ที่มีผลเกี่ยวข้องกับน้ำตาลในเลือด เพื่อดูค่าความถูกต้องเป็นตัวตั้งหลัก	HBA1C FBS	87.73% 86.42%
3	เพิ่มเป็น 2 แอตทริบิวต์ เพื่อดูค่าความถูกต้องและพบว่าค่าความถูกต้องเพิ่มมากขึ้น	HBA1C + FBS	87.86%
4	เพิ่มแอตทริบิวต์ตัวอื่น ๆ ที่ไม่ใช่สารเคมีในเลือด พบว่า แอตทริบิวต์ HBA1C +FBS + BMI ให้ค่าความถูกต้องได้มากที่สุด	HBA1C +FBS + BMI	87.92%
5	จากนั้นเพิ่มทีละ 1 แอตทริบิวต์ ที่ไม่ใช่สารเคมีในเลือด คือ Sex, Age, Bw และพบว่าค่าความถูกต้องไม่เพิ่มขึ้น จึงตัดออกไปไม่ใช้งานอีก	BMI + HBA1C +FBS+ Sex / Age / Bw	87.86%
6	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 4 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C +FBS+ Cholesterol	88.04%
7	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 5 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C +FBS+ Cholesterol + LDL	88.07%
8	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 6 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C +FBS + Cholesterol+ LDL+HDL	88.17%

ขั้นตอนที่	วิธีการ	แอตทริบิวต์ที่เลือกใช้	Accuracy
9	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 7 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C +FBS + Cholesterol+ LDL+HDL+GFR	88.20%
10	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 8 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C + FBS + Cholesterol + LDL+ HDL+ GFR + Triglyceride	88.20%
11	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า 9 แอตทริบิวต์นี้ ให้ค่าความถูกต้องมากที่สุด	BMI + HBA1C + FBS + Cholesterol + LDL + HDL + GFR + Triglyceride + Potassium	88.21%
12	หลังจากนั้น เพิ่มอีกทีละ 1 แอตทริบิวต์ พบว่า ความถูกต้องไม่เพิ่มมากขึ้น จึงหยุดกระบวนการ	BMI + HBA1C + FBS + Cholesterol + LDL + HDL + GFR + Triglyceride + Potassium	88.21%
13	สุดท้ายเพิ่มแอตทริบิวต์ DM ที่เป็นการแบ่งประเภทว่าเป็นโรคเบาหวานหรือไม่เป็นโรค รวมทั้งหมดเป็น 10 แอตทริบิวต์สำหรับการทำวิจัยฉบับนี้	BMI + HBA1C + FBS + Cholesterol + LDL + HDL + GFR + Triglyceride + Potassium + DM	88.21%

ตารางที่ 4 ผลการทดลอง

Methods	Performance	Decision tree J 48	Random Forest	Naive Bayes	KNN Lazy IBK	Multilayer Perceptron
Cross-Validation 10 Folds	Accuracy	88.20%	88.98%	87.55%	84.97%	**90.82%**
	Precision	87.70%	88.60%	87.00%	84.30%	**90.50%**
	Recall	88.20%	89.00%	87.60%	85.00%	**90.80%**
	F-Measure	87.80%	88.80%	87.10%	84.50%	**90.60%**
Percentage 70%	Accuracy	88.19%	88.65%	87.19%	84.96%	**90.53%**
	Precision	87.70%	88.20%	86.50%	84.20%	**90.20%**
	Recall	88.20%	88.60%	87.20%	85.00%	**90.50%**
	F-Measure	87.80%	88.30%	86.60%	84.40%	**90.10%**

จากตารางที่ 4 สามารถแสดงให้เห็นว่า อัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น (Multilayer Perceptron) เป็นแบบจำลองที่มีประสิทธิภาพมากที่สุด เมื่อวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล โดยการวิธี Cross-Validation 10 Folds พบว่ามีค่าความถูกต้อง 90.82%, ค่าความแม่นยำ 90.50%, ค่าความระลึก 90.80% และ

ค่าความถ่วงดุล 90.60% และเมื่อใช้วิธี Percentage 70% ได้ค่าความถูกต้อง 90.53%, ค่าความแม่นยำ 90.20%, ค่าความระลึก 90.50% และค่าความถ่วงดุล 90.10% แบบจำลองที่ใช้อัลกอริทึมป่าสุ่ม ให้ประสิทธิภาพดีรองลงมา โดยการวิธี Cross-Validation 10 Folds พบว่ามีค่าความถูกต้อง 88.98%, ค่าความแม่นยำ 88.60%, ค่าความระลึก 89.00% และค่าความถ่วงดุล 88.80% และเมื่อใช้วิธี Percentage 70% ได้ค่าความถูกต้อง 88.65%, ค่าความแม่นยำ 88.20%, ค่าความระลึก 88.60% และค่าความถ่วงดุล 88.30% สำหรับแบบจำลองที่ใช้อัลกอริทึม KNN Lazy iBK เป็นแบบจำลองที่มีประสิทธิภาพน้อยที่สุด เมื่อวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล โดยการวิธี Cross-Validation 10 Folds โดยวัดค่าความถูกต้อง 84.97%, ค่าความแม่นยำ 84.30%, ค่าความระลึก 85.00% และค่าความถ่วงดุล 84.50% และเมื่อใช้วิธี Percentage 70% ได้ค่าความถูกต้อง 84.96%, ค่าความแม่นยำ 84.20%, ค่าความระลึก 85.00% และค่าความถ่วงดุล 84.40%

อภิปรายผลการวิจัย

การลดมิติข้อมูล ของงานวิจัยครั้งนี้ ช่วยเพิ่มประสิทธิภาพในการประมวลผล ช่วยทำให้ค่าความถูกต้องเพิ่มมากขึ้น จากการใช้อัลกอริทึมแรนดอมฟอเรสต์ และประเมินประสิทธิภาพแบบจำลองด้วยวิธี Cross-Validation 10 Folds หลังจากลดมิติข้อมูลแล้ว ค่าความถูกต้องเพิ่มขึ้นจาก 87.65% เป็น 88.21% และเมื่อทดสอบใช้อัลกอริทึมอื่น ๆ ก็พบว่ามีค่าความถูกต้องเพิ่มมากขึ้นเช่นเดียวกัน สอดคล้องกับงานวิจัยของอัจจิมา มณฑาพันธุ์ ที่ศึกษาเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่สำคัญด้วย 7 เทคนิค พบว่าการใช้เทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมจะช่วยเพิ่มความถูกต้องของการพยากรณ์ได้มากยิ่งขึ้น (อัจจิมา มณฑาพันธุ์, 2562)

นอกจากการลดมิติข้อมูลงานวิจัยนี้ ยังทำการจำแนกโรคเบาหวานจากผลตรวจเลือด ได้เปรียบเทียบการทำงานของเทคนิคการจำแนกประเภท ซึ่งใช้อัลกอริทึม 5 อัลกอริทึม คือ อัลกอริทึมต้นไม้ตัดสินใจ อัลกอริทึมแรนดอมฟอเรสต์ อัลกอริทึมเบย์อย่างง่าย อัลกอริทึมเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น จากผลการทดลองพบว่า แบบจำลองที่ใช้อัลกอริทึมโครงข่ายประสาทเทียมหลายชั้น เป็นอัลกอริทึมที่ให้ค่าความถูกต้องของการจำแนกข้อมูลมากที่สุด ทั้งวิธีแบบ Cross-Validation 10 Folds และ Percentage 70% ด้วยจำนวน 90.82% และ 90.535% ตามลำดับ แสดงว่าเป็นอัลกอริทึมโครงข่ายประสาทเทียมมีประสิทธิภาพมากที่สุดกับชุดข้อมูลที่นำมาทำการวิจัยในครั้งนี้ ซึ่งสอดคล้องกับงานวิจัยของรัชนีวรรณ ไพศาลวรเกียรติ ได้ศึกษาการเปรียบเทียบตัวแบบการถดถอยลอจิสติกและเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเป็นโรคเบาหวาน โดยใช้เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียม ผลการศึกษาพบว่าเทคนิคโครงข่ายประสาทเทียมมีประสิทธิภาพในการพยากรณ์ดีที่สุด (รัชนีวรรณ ไพศาลวรเกียรติ, 2564) และงานวิจัยของปพนน์ศรณ์ สิ่วสำแดงเดช ที่ศึกษาเกี่ยวกับการจำแนกเบาหวานโดยใช้เทคนิคการโหวตรวม ซึ่งได้ประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผู้ป่วยโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคโหวตรวม และเทคนิคป่าสุ่ม ซึ่งผลการวิจัยพบว่าเทคนิคป่าสุ่ม ให้ค่าความถูกต้องในการทำนายผลดีที่สุด (ปพนน์ศรณ์ สิ่วสำแดงเดช, 2565) เนื่องจากงานวิจัยไม่มีการใช้อัลกอริทึมโครงข่ายประสาทเทียม และสอดคล้องกับงานวิจัยของสายชล สินสมบูรณ์ทอง ที่ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานโดยวิธีการจำแนกที่นำมาเปรียบเทียบมี 7 วิธี และพบว่าวิธีโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้องมากที่สุด คือ 95.94% (สายชล สินสมบูรณ์ทอง, 2561) สำหรับงานวิจัยฉบับนี้ ถึงแม้จะได้ข้อสรุปสอดคล้องกับงานของสายชล สินสมบูรณ์ทอง ที่ใช้อัลกอริทึมโครงข่ายประสาทเทียมแบบหลายชั้นแล้วให้ประสิทธิภาพดีที่สุด แต่เนื่องจาก

ข้อมูลของผู้ป่วยที่เป็นโรคเบาหวาน และผู้ป่วยที่ไม่ได้เป็นโรคเบาหวาน ของการวิจัยครั้งนี้มีจำนวนที่แตกต่างกันค่อนข้างมาก ทำให้ข้อมูลไม่สมดุล (Imbalanced data) ส่งผลให้ประสิทธิภาพในการประมวลผลต่ำกว่าที่คาดคิด

จากผลการวิจัยการจำแนกโรคเบาหวานจากผลตรวจเลือดด้วยวิธีการทำเหมืองข้อมูล สามารถนำแบบจำลองที่ได้ไปใช้ประโยชน์ในโรงพยาบาล ที่มีผู้ป่วยความเสี่ยงสูงเป็นโรคเบาหวานจำนวนมากได้ นอกจากนี้ยังสามารถนำไปพัฒนาเป็นแอปพลิเคชันเพื่อใช้จำแนกการเกิดโรคเบาหวานได้ อย่างไรก็ตามแบบจำลองที่สร้างขึ้นในการวิจัยครั้งนี้มีความเหมาะสมกับชุดข้อมูลผู้ป่วยที่ได้รับการวินิจฉัยว่าเป็นโรคเบาหวานและผู้ป่วยทั่วไป ของโรงพยาบาลวชิระภูเก็ตเท่านั้น อาจไม่เหมาะสมกับชุดข้อมูลอื่น เนื่องจากคุณลักษณะของแต่ละชุดข้อมูลมีความแตกต่างกัน จึงไม่สามารถนำผลลัพธ์ของแบบจำลองนี้ไปใช้สรุปผลกับชุดข้อมูลอื่น ๆ ได้

ข้อเสนอแนะ

1. เนื่องจากข้อมูลเดิมของผู้ป่วยโรคเบาหวาน และผู้ป่วยที่ไม่ได้เป็นโรค มีจำนวนที่แตกต่างกันค่อนข้างมาก ทำให้ข้อมูลไม่สมดุล (Imbalanced data) อาจจะส่งผลต่อประสิทธิภาพในการประมวลผล สามารถแก้ไขปัญหาดังกล่าวได้โดยใช้เทคนิค Undersampling หรือ Oversampling

2. ควรสร้างเป็นระบบผู้เชี่ยวชาญโดยการพัฒนาแอปพลิเคชันสำหรับจำแนกการเกิดโรคเบาหวาน เพื่อนำโมเดลที่ได้ไปประยุกต์ใช้งานจริง ซึ่งจะช่วยเพิ่มความแม่นยำในการจำแนกผู้ป่วยที่เป็นและไม่เป็นโรคเบาหวาน เพิ่มความรวดเร็วในการให้การรักษา ลดการทำงานของเจ้าหน้าที่ แพทย์ พยาบาล และเจ้าหน้าที่ที่เกี่ยวข้องได้

3. อาจใช้วิธีการทำเหมืองข้อมูลด้วยเทคนิคกฎความสัมพันธ์ (Association Rule) เพื่อหากฎความสัมพันธ์ของการเป็นโรคเบาหวาน

กิตติกรรมประกาศ

ในงานวิจัยฉบับนี้ สำเร็จลุล่วงได้อย่างสมบูรณ์ ต้องขอขอบคุณเจ้าหน้าที่จากโรงพยาบาลวชิระภูเก็ต ที่เอื้อเฟื้อข้อมูลจากฐานข้อมูลผู้ป่วย เพื่อให้สามารถดำเนินการวิจัยได้สำเร็จลุล่วง

เอกสารอ้างอิง

ดำรงเดช เดินรีบรัมย์, ฉัตรเกล้า เจริญผล และจริยา จิรานุกูล. (2563). การเปรียบเทียบประสิทธิภาพโครงสร้างเหมืองข้อมูลเพื่อจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์. *Journal of Science and Technology Mahasarakham University*, 39(3).

รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย. (2563). การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล. *PKRU SciTech Journal*, 3(2), 11–19.

ปพนน์ศรณ์ สิ่วสำแดงเดช. (2565). *การจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี.* วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต (สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์). มหาวิทยาลัยมหาสารคาม, มหาสารคาม.

พงศกร ธีรรัศมี. (2558). *วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูล ทางการแพทย์.* วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต (สาขาวิชาวิศวกรรมคอมพิวเตอร์). มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

รัชนีวรรณ ไพศาลวรเกียรติ. (2564). *การเปรียบเทียบตัวแบบการถดถอยลอจิสติกและเทคนิคเหมืองข้อมูลสำหรับการพยากรณ์การเป็นโรคเบาหวาน*. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต (สาขาวิชาสถิติ). มหาวิทยาลัยนเรศวร, พิษณุโลก.

รัชพล กลัดชื่น และจรัญ แสนราช. (2561). การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา. *Research Journal Rajamangala University of Technology Thanyaburi*, 17(1).

สมเกียรติ โพธิสัตย์, สถิตย์ นิรมิตมหาปัญญา, ชัยชาญ ดีโรจน์วงศ์, วีระศักดิ์ ศรินนภากร, นภา ศิริวิวัฒนากุล, สิทธิชัย อาชายินดี, และธนพร รัตนสุวรรณ. (2557). *โรคเบาหวาน (Diabetes Mellitus).* Thailand Medical Services Profile 2011-2014 (การแพทย์ไทย 2554-2557) First Edition. กรุงเทพฯ: กรมการแพทย์ กระทรวงสาธารณสุข.

สมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี. (2560). *แนวทางเวชปฏิบัติสำหรับโรคเบาหวาน 2560* (3rd ed.). บริษัท ร่มเย็น มีเดีย จำกัด.

สายชล สินสมบูรณ์ทอง. (2561). การเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน. *วารสารวิทยาศาสตร์และเทคโนโลยี*, 26(2).

อรรถเกียรติ กาญจนพิบูลวงศ์, ภาณุวัฒน์ คำวังสง่า และสุธิดา แก้วทา. (2563). *รายงานสถานการณ์โรค NCDs เบาหวาน ความดันโลหิตสูง และปัจจัยเสี่ยงที่เกี่ยวข้อง พ.ศ.2562*. กรุงเทพฯ: สำนักพิมพ์อักษรกราฟฟิกแอนด์ดีไซน์.

อัจจิมา มณฑาพันธุ์. (2562). การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม. *Royal Thai Air Force Medical Gazette*, 65(2), 49–56.

อุกฤษฎ์ ศรีสุข และจารี ทองคำ. (2564). การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค. *Journal of Science and Technology Mahasarakham University*, 40(2), 157–163.

Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. *Advances in Computers*, 117(1), 339–368.

Aurélien Géron. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. *O'Reilly Media*, 851.

Gojka Roglic, & World Health Organization. (2016). *Global report on diabetes*. World Health Organization.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. The University of Waikato.

Han, J., & Kamber, Micheline. (2006). *Data Mining Concepts and Techniques (2nd Edition)*.

Hassan, H., Negm, A., Zahran, M., & Saavedra, O. (2015). *Assessment of Artificial Neural Network for bathymetry estimation using High Resolution Satellite imagery in Shallow Lakes: Case Study El Burullus Lake*.

International Diabetes Federation. (2021). *IDF Diabetes Atlas 10th edition*.

Pang-Ning Tan, Michael Steinbach, V. K. (1981). *Introduction to Data Mining*. https://doi.org/10.1016/0022-4405(81)90007-8

Roiger, R., & Geatz, M. (2003). *Data Mining : A Tutorial Based Primer by Michael Geatz*. 5.

Singh, N., & Singh, P. (2021). A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemometrics and Intelligent Laboratory Systems*, *217*, 104396.