การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิควิธีแบบร่วมกันตัดสินใจและวิธีเลือกคุณลักษณะเด่นไปข้างหน้า

DATA CLASSIFICATION FOR DIABETES RISK DIAGNOSIS USING MAJORITY VOTING ENSEMBLE METHOD AND FORWARD FEATURE SELECTION METHOD

นพรัตน์ นนท์ศิริ[1,*], พิศณุ ชัยจิตวณิชกุล[2] และ กริช สมกันธา[1]
Nopparat Nonsiri[1,*], Pitsanu Chaichitwanidchakol[2] and Krit Somkantha[1]

[1] สาขาวิทยาการข้อมูลและเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี
[2] สาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี
[1] Data Science and Information Technology, Faculty of Science, Udon Thani Rajabhat University
[2] Computer Science and Information Technology, Faculty of Science, Udon Thani Rajabhat University

Received: 9 March 2022 Accepted: 24 June 2022

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาขั้นตอนวิธีในการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน กรณีข้อมูลผู้ป่วยโรคเบาหวานโรงพยาบาลสมเด็จพระยุพราชบ้านดุง เป็นข้อมูลที่เกิดจากการทบทวนเวชระเบียนผู้ป่วยโรคเบาหวานย้อนหลังปี 2557-2561 ซึ่งลักษณะข้อมูลดังกล่าวเป็นข้อมูลที่มีมิติสูง เนื่องจากคุณลักษณะของข้อมูลนั้นมีหลายคุณลักษณะ และบางคุณลักษณะไม่มีความสัมพันธ์ต่อการจำแนกข้อมูล ดังนั้นจำเป็นต้องมีการเลือกคุณลักษณะเบื้องต้น เพื่อลดความซ้ำซ้อนของข้อมูลและเพิ่มประสิทธิภาพการจำแนกความถูกต้องของคลาส (Class) ในการแก้ปัญหาเหล่านี้ผู้วิจัยได้ใช้วิธีเลือกคุณลักษณะเด่นไปข้างหน้า (Forward Selection) และวิธีร่วมกันตัดสินใจจากต้นไม้ตัดสินใจ 3 โมเดล เพื่อเลือกคุณสมบัติที่เหมาะสม (Voting Tree) วัดประสิทธิภาพคุณลักษณะด้วยวิธีครอสวาลิเดชั่น (Cross Validation) จำแนกข้อมูลด้วยขั้นตอนวิธีร่วมกันตัดสินใจ (Voting Ensemble), วิธีเกรเดียนท์บูตทรีส์ (Gradient Boosted), วิธีต้นไม้ตัดสินใจ (Decision Tree), วิธีแรนดอม

* ผู้ประสานงาน: นพรัตน์ นนท์ศิริ
อีเมล์: Nopparat.nonsiri@gmail.com

ฟอเรสต์ (Random Forest), วิธีนาอีฟเบย์ (Naïve Bayes), วิธีซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine), วิธีเพื่อนบ้านที่ใกล้เคียงกันที่สุด (K-Nearest Neighbor) วัดประสิทธิภาพแบบจำลอง (Accuracy) ด้วยวิธีครอสวาลิเดชั่น จากการทดสอบความถูกต้องในการจำแนกข้อมูล ผลการเปรียบเทียบพบว่า วิธีร่วมกันตัดสินใจให้ผลลัพธ์ที่ดีกว่าการใช้เทคนิคแบบโมเดลเดี่ยว (Single Model) ทั้งนี้เพราะเมื่อนำตัวจำแนกข้อมูลที่หลากหลายมาช่วยกันตัดสินใจด้วยโหวตเสียงข้างมากจะช่วยในการลดปัญหาการเกิดความโน้มเอียงของข้อมูล (Bias) และการเลือกใช้ตัวจำแนกที่ดีแต่ละตัวช่วยกันเสริมประสิทธิภาพในการจำแนกข้อมูล ทำให้โมเดลที่ได้มีประสิทธิภาพสูงขึ้น นอกจากนั้นยังพบอีกว่าการเลือกใช้คุณลักษณะที่เหมาะสมด้วยวิธีเลือกคุณลักษณะเด่นไปข้างหน้าที่ผ่านวิธีร่วมกันตัดสินใจทำให้โมเดลมีประสิทธิภาพในการจำแนกเพิ่มมากยิ่งขึ้น เหมาะสมที่จะนำโมเดลดังกล่าวนำไปใช้เป็นแนวทางในการสนับสนุนการตัดสินใจทางการแพทย์ในการวินิจฉัยการเป็นโรคเบาหวานอย่างมีประสิทธิภาพ

คำสำคัญ: วิธีแบบร่วมกันตัดสินใจ, การคัดเลือกคุณลักษณะที่เหมาะสม, ความเสี่ยงการเป็นโรคเบาหวาน

Abstract

The purpose of this research was to find an algorithm to classify data for diagnosing diabetes risk. Diabetes Patient Information, Somdej Phrayuparaj Ban Dung Hospital. The data were from the review of medical records of patients with diabetes in the past 2014-to 2018. The nature of such information was high-dimensional information. Because there were many attributes and some attributes were not related to the classification of data. Therefore, a preliminary selection of features was required to reduce data redundancy and optimize the classification accuracy of classes (Class) to solve these problems.

The researcher used the Forward Selection method and how to make decisions together from 3 model decision trees to select the appropriate properties (Voting Tree) to measure performance with Cross-Validation, Voting

Ensemble, Gradient Boosted, Decision Tree method. Tree Method, Random Forest Method, Naïve Bayes Method, Support Vector Machine Method, Nearest Neighbor Method (K-Nearest Neighbor). Measure model performance (Accuracy) by cross-validation method by testing the accuracy of data classification. The comparison results found that a collaborative decision-making approach yields better results than a single model technique.

This was because when using a variety of classifiers to make decisions with a majority vote; help to reduce bias, and choosing a good classifier enhances the efficiency of data classification, make the model more efficient. It also found that the selection of suitable attributes through a co-op decision-making approach made the model more efficient in its classification. It is appropriate to use this model as a guideline for effective medical decision support in the diagnosis of diabetes.

Keywords: Majority Voting Ensemble, Diabetes Risk Diagnosis, Features Selection

บทนำ

จากข้อมูลของสหพันธ์เบาหวานนานาชาติ พบผู้ป่วยโรคเบาหวานทั่วโลกราว 425 ล้านคนในปี 2560 และคาดการณ์ว่าจะมีจำนวนผู้ป่วยด้วยโรคนี้มากถึง 520 ล้านคนในปี 2578 (Li et al., 2017) สำหรับสถานการณ์โรคเบาหวานในประเทศไทยพบว่า คนไทยช่วงอายุ 20-79 ปี เป็นโรคเบาหวานร้อยละ 8.3 หรือหมายความว่าใน 100 คน จะพบคนที่ป่วยเป็นโรคเบาหวานประมาณ 8 คน และจำนวนมากกว่าครึ่งไม่ทราบว่าตนเองเป็นโรคเบาหวาน ซึ่งผู้ที่อยู่ในกลุ่มเสี่ยงต่อภาวะการเป็นเบาหวาน สามารถพัฒนาการเกิดโรคเบาหวานประเภทที่ 2 ได้ และพบว่าผู้ป่วยโรคเบาหวานมีภาวะแทรกซ้อน เนื่องจากไตเสื่อมสูงสุดถึงร้อยละ 43.9 ต้อกระจกร้อยละ 42.8 และจอประสาทตาเสื่อมร้อยละ 30.7 และพบมีภาวะแทรกซ้อนจากโรคหัวใจขาดเลือดและโรคหลอดเลือดสมองร้อยละ 8.1 และ 4.4 ตามลำดับ โดยโอกาสของคนจะเป็นเบาหวานจะมีค่าน้ำตาลเกินค่าปกติ คือ 126 มิลลิกรัมต่อเดซิลิตร จะถือว่าเป็นโรคเบาหวาน ซึ่งโรคนี้จะทำให้ภูมิคุ้มกันของร่างกายลดลง (WHO & IDF, 2006)

โรงพยาบาลสมเด็จพระยุพราช เป็นโรงพยาบาลชุมชนประจำอำเภอสังกัดกระทรวงสาธารณสุข มีขีดความสามารถระดับปฐมภูมิ (Primary Care) และระดับทุติยภูมิ (Secondary Care) ซึ่งปัจจุบันกำลังประสบปัญหาโรคเรื้อรังที่เกี่ยวข้องกับการไม่ปฏิบัติตนตามพฤติกรรมสุขภาพที่เหมาะสมของคนในพื้นที่ เช่น รับประทานอาหารที่มีประโยชน์อย่างเหมาะสม ออกกำลังกายแบบแอโรบิค จากสภาวการณ์ปัจจุบันนี้เองก่อเกิดปัจจัยเสี่ยงด้านสุขภาพมากมาย ทำให้ประชาชนในท้องถิ่นมีแนวโน้มเจ็บป่วยด้วยโรคเรื้อรังมากขึ้นนั่นก็คือโรคเบาหวานซึ่งเป็นหนึ่งในโรคที่สำคัญที่ทำให้เกิดภาวะแทรกซ้อนของโรคอื่น อีกทั้งทางโรงพยาบาลยังขาดแคลนบุคลากรทางการแพทย์โดยเฉพาะทีมบริการเฉพาะโรคเมื่อโรงพยาบาลออกบริการชุมชน

ดังนั้นคณะผู้วิจัยจึงเห็นถึงปัญหาและความสำคัญในการศึกษาสถานการณ์สุขภาพของผู้ป่วยโรคเบาหวานในโรงพยาบาลสมเด็จพระยุพราชบ้านดุง โดยมีวัตถุประสงค์เพื่อหารูปแบบที่เหมาะสมในการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานและเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี ในการจำแนกข้อมูลความเสี่ยงการเป็นโรคเบาหวานด้วยวิธีแบบร่วมกันตัดสินใจและการคัดเลือกคุณลักษณะที่เหมาะสม เพื่อที่จะนำข้อมูลดังกล่าวไปพัฒนาระบบวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน ให้สามารถตรวจสอบความเสี่ยงในการเกิดโรคเบื้องต้น แล้วนำผลไปใช้กำหนดแนวทางส่งเสริมสุขภาวะที่ดีของกลุ่มผู้ที่มีแนวโน้มที่จะป่วยเป็นโรคเบาหวานในอนาคต ตลอดทั้งการวางแผนรองรับการรักษาโรคเบาหวาน ซึ่งนับเป็นเรื่องที่สำคัญที่จะต้องเร่งรัด ดำเนินการ และจะต้องให้ความสนใจในสาเหตุหรือปัจจัยที่มีความสัมพันธ์ต่อการเกิดโรค หากผู้ป่วยเกิดพัฒนาเป็นโรคเบาหวานประเภทที่ 2 จะทำ ให้มีอัตราการเสียชีวิตเพิ่มขึ้น ถึงร้อยละ 90 เนื่องจากเกิดอาการแทรกซ้อนของโรคอื่น ๆ (Kazerouni et al., 2020)

วิธีการดำเนินวิจัย

ในปัจจุบันการเรียนรู้ด้วยเครื่องมีหลายรูปแบบคือแบบผู้สอน (Supervised learning) แบบไม่มีผู้สอน (Unsupervised learning) และแบบการเรียนรู้เกิดมาจากการปฏิสัมพันธ์ (Reinforcement learning) ซึ่งผู้วิจัยได้ศึกษาเทคนิคเพื่อการจำแนกกลุ่มของข้อมูลประกอบด้วย วิธีร่วมกันตัดสินใจ, วิธีเกรเดียนท์บูตทรีส์, วิธีต้นไม้ตัดสินใจ, วิธีแรนดอมฟอเรสต์, วิธีนาอีฟเบย์, วิธีซัพพอร์ทเวกเตอร์แมชชีนและวิธีความใกล้เคียงกันที่สุด เนื่องจากเทคนิค

ดังกล่าว เป็นเทคนิคที่นิยมและเหมาะกับการจำแนกข้อมูลที่เป็นหมวดหมู่ข้อมูลต่าง ๆ ที่ถูกแนบอยู่ในแต่ละเรคคอร์ดของชุดข้อมูล โดยการเรียนรู้แบบมีผู้สอนจะแตกต่างกับการเรียนรู้แบบไม่มีผู้สอนที่จะไม่ทราบถึงหมวดหมู่ของข้อมูล ตัวอย่างเช่น ในการวิเคราะห์ข้อมูลโรคเบาหวานที่ไม่มีข้อมูลที่บ่งบอกว่าเป็นโรคเบาหวาน ก็จะนำข้อมูลปัจจัยไปวิเคราะห์หาความเสี่ยงว่าเป็นโรคเบาหวานหรือไม่ เป็นต้น และข้อมูลที่ผู้วิจัยได้ทำการรวบรวมนั้นมีหมวดหมู่ของข้อมูลที่บ่งบอกว่าเป็นโรคเบาหวานอย่างชัดเจน ผู้วิจัยจึงได้เลือกใช้เทคนิควิธีดังกล่าวข้างต้น โดยมีรายละเอียดดังนี้

1. วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine) เป็นวิธีที่ใช้จำแนกค่าคุณลักษณะของ 2 กลุ่มโดยจะสร้างเส้นแบ่ง (Plane) ที่ เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่ม ออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดย เส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้างออกไปจนกว่าจะสัมผัส กับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุดโดยอาศัยการปรับค่าสมการด้วยวิธีการเคอร์เนล (Kazerouni et al., 2020) ในการหารูปแบบและความสัมพันธ์ เคอร์เนลฟังก์ชันนั้น มีอยู่เป็นจำนวนมากที่รู้จักกันดี เช่น Polynomial, RBF หรือ Sigmoid โดยผู้วิจัยได้เลือกใช้เคอร์เนลรวมแบบดอท หาได้จากสมการดังนี้

$$(x_{i,}y_{i,}),\ \ldots,(x_{n,}y_{n,})\ \text{เมื่อ}\ x\ \in\ R^{m},\ y \in \{+1,-1\} \quad (1)$$

$$(w^{t}\ *x)+b \quad (2)$$

$$(w^{t}*x)+b>0\ \text{ถ้า}\ y_{i}=+1$$

$$(w^{t}*x)+b<0\ \text{ถ้า}\ y_{i}=\ -1$$

$(x_{i,}y_{i,}),\ \ldots,(x_{n,}y_{n,})$ แทน ข้อมูลกลุ่มตัวอย่าง

w^{t} แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature

b แทน ค่าโน้มเอียง (bias)

n แทน จำนวนข้อมูลตัวอย่าง

m แทน จำนวนมิติข้อมูล

y แทน ผลลัพธ์กลุ่มข้อมูลมีค่า +1 หรือ -1

2. วิธีเพื่อนบ้านที่ใกล้เคียงกันที่สุด (K-Nearest Neighbor) เป็นวิธีการหนึ่งสำหรับแก้ปัญหา การประมาณค่าที่ไม่ใช่พารามิเตอร์สำหรับการจำแนกกลุ่ม (Kazerouni et al., 2020) ซึ่งเหมาะกับข้อมูลที่เป็นรูปร่างที่ดี หลักการก็คือจะวัดค่าของข้อมูลที่ใกล้ที่สุด เพื่อหาจุดได้เท่ากับจำนวน k จนกว่าจะเจอข้อมูลจำนวน k 5 ตัว จึงจะหยุด จากนั้นก็ทำการจำแนกว่ามันอยู่ใกล้กับข้อมูลจุดไหนมากสุดและหาค่าได้จากสมการดังนี้

$$Distance = \sqrt{(w_1 - x_1)^2 + (w_2 - x_2)^2 + (w_i - x_i)^2} \quad (3)$$

โดย Distance คือผลรวมระยะทางของ x_i

w_i คือจำนวนของแอตทริบิวต์ข้อมูล Test ทั้งหมดของชุดข้อมูล

x_i คือจำนวนแอตทริบิวต์ข้อมูล Train ทั้งหมดของชุดข้อมูล

3. วิธีต้นไม้ตัดสินใจ (Decision tree) เป็นการจำแนกค่าคุณลักษะของข้อมูลโดยจะประกอบด้วยบัพ (node) และ กิ่ง (link) ที่ต่อกับบัพ บัพที่ปลายสุดจะเรียก ว่าบัพใบ (leaf) ต้นไม้ตัดสินใจจะทำโดยสร้างบัพทีละบัพเพื่อตรวจสอบคุณสมบัติของตัวอย่าง แล้วแยกตัวอย่างลงตามค่าของกิ่ง ทำจนกระทั่งตัวอย่างในใบแต่ละใบอยู่ในประเภทเดียวกันทั้งหมดโดย s เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น c_i แทนกลุ่มในลำดับ ที่ i โดย ที่ i มีค่าระหว่าง 1 ถึง n, s_i แทนจำนวน ข้อมูลสมาชิกของ s และอยู่ในกลุ่ม c_i หาได้จากสมการที่ (4), (5) และ (6) (Zou et al., 2018)

$$I(s_1, s_{2,\dots}, s_n) = -\sum_{i=1}^{n} \frac{s^1}{s} \log_2 \frac{s^1}{s} \quad (4)$$

$$E(A) = \sum_{j=1}^{n} \frac{s_{1j}+\cdots+s_{nj}}{s} I(s_{1j}, s_{2j}, \dots s_{nj}) \quad (5)$$

$$Gain(A) = I(s_{1j}, s_{2j}, \dots, s_{nj}) - E(A) \quad (6)$$

4. วิธีนาอีฟเบย์ (Naïve Bayes) เป็นตัวจำแนกที่เหมาะกับกรณีของเซ็ตตัวอย่างมีจำนวนมากและคุณสมบัติ (attribute) ของตัวอย่างไม่ขึ้นต่อกัน (Nagaratnam et al., 2020) มีการนำตัวจำแนกประเภทเบย์ไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความและพบว่าใช้งานได้ดี ไม่ต่างจากการจำแนกประเภทวิธีอื่น ๆ หาได้จากสมการ (7) และ (8)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (7)$$

$$P(c|X) = P(x_1|c) \times \ldots \times P(x_n|c) \times P(c) \quad (8)$$

c คือ คลาสของข้อมูล

x คือ แอตทริบิวต์

p คือ ความน่าจะเป็นของข้อมูล

$P(c|x)$ คือความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c

$P(x|c)$ คือ ความน่าจะเป็นที่ข้อมูลที่มีคลาส c และมีแอตทริบิวต์ x

$P(c)$ คือ จำนวนคลาสที่อาจจะเกิดขึ้นหารด้วยจำนวนคลาสทั้งหมดของคลาส c

$P(x)$ คือ จำนวน แอตทริบิวต์ ทั้งหมด

5. วิธีการร่วมกันตัดสินใจคือการนำเอาตัวจำแนกข้อมูลหลาย ๆ ตัวมาร่วมกันตัดสินใจโดยการโหวตเสียงข้างมาก (Voting Ensemble) จะเลือกค่านำหนักในการลงคะแนน 2 ใน 3 ของโมเดลเพื่อตัดสินผลลัพธ์ สามารถกำหนดค่า f_k โดยที่ $(k = 1, 2, \ldots, K)$ ซึ่ง K คือ จำนวนของตัวจำแนกข้อมูลและ $f_k(x) = c,\ c \in \{+1, -1\}$ โดยที่ฟังก์ชันในการตัดสินใจจะตัดสินใจผลลัพธ์ที่ได้ว่าอยู่ในคลาสบวกหรือคลาสลบ จากนั้นนำค่าฟังก์ชัน $f_e(x)$ ที่ได้จากตัวจำแนกข้อมูลมาตัดสินใจร่วมกันโดยดูที่เสียงข้างมากว่าของโมเดลตัดสินใจเป็นคลาสบวก หรือคลาสลบโดยมีสมการดังสมการที่ (9)

$$f_e(x) = \arg\ max\ \sum_{k:f_e(x)=c} 1 \quad (9)$$

6. วิธีแรนดอมฟอเรสต์ (Random Forest) เป็นโมเดลอีกประเภทหนึ่งของเครื่องจักรเรียนรู้ (Machine Learning) ถูกพัฒนาขึ้นจากต้นไม้ตัดสินใจ ต่างกันที่วิธีแรนดอมฟอเรสต์เป็นการเพิ่มจำนวน ต้นไม้ตัดสินใจหลาย ๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้นแม่นยำมากขึ้น ซึ่งเป็นโมเดลที่ได้รับความนิยมไปอย่างมากในการใช้กับเครื่องจักรเรียนรู้ กำหนดให้ $RFfi_i$ คือคุณลักษณะที่คำนวณได้จากต้นไม้ทั้งหมดในแบบจำลอง Random Forest $normf_{i_{i_j}}$ คือคุณสมบัติที่ทำให้เป็นมาตรฐานสำหรับ i ใน $tree_j$ และ T คือจำนวนต้นไม้ทั้งหมด โดยมีสมการดังสมการที่ (10) (Lan & Pan, 2019)

$$RFfi_i = \frac{\sum_{j \in all\ tree} norm f_{i_j}}{T} \quad (10)$$

7. วิธีเกรเดียนท์บูตทรีส์ (Gradient Boosted) คือวิธีที่มีพื้นฐานมาจากต้นไม้ตัดสินใจ ซึ่งเป็นการปรับปรุงประสิทธิภาพของแบบจำลองใหม่โดยการสุ่มสร้างต้นไม้ตัดสินใจหลายร้อยโมเดลและประเมินผลแต่ละโมเดลจนกว่าจะได้ต้นไม้ตัดสินใจที่สมบูรณ์ โดย n คือจำนวนของน้ำหนัก S_i น้ำหนักของเซต x_i, y_i เซตของการจำแนก หาได้จากสมการที่ (11) และ (12) (Dutta et al., 2020)

$$S_i = \{(x_i, y_i)\}^n \quad (11)$$

$$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x) \quad (12)$$

8. การวัดประสิทธิภาพแบบจำลองโดยใช้วิธีครอสวาลิเดชั่น (Cross validation) คือทำการแบ่งข้อมูลออกเป็นส่วน ๆ โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากันและนำแต่ละส่วนไปทดสอบ จากนั้นเก็บค่าความถูกต้องของโมเดลแต่ละรอบไว้แล้วนำมาพิจารณาค่าเฉลี่ยทุกกลุ่มข้อมูล (Dutta et al., 2020) คำนวณได้จากสมการดังนี้

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

$$Precision = \frac{TP}{TP+TN} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

$$F - Measure = \frac{Precision * Recall}{Precision+Recall} \quad (16)$$

$TP =$ อัตราความถูกต้องเชิงบวก
$TN =$ อัตราความถูกต้องเชิงลบ
$FP =$ อัตราความผิดพลาดเชิงบวก
$FN =$ อัตราความผิดพลาดเชิงลบ

9. วิธีเลือกคุณลักษณะเด่นไปข้างหน้า (Forward Selection) เป็นวิธีการที่นำคุณลักษณะแต่ละตัวมาหาความถูกต้องด้วยขั้นตอนวิธีการจำแนกข้อมูลและทำการทดสอบว่า

คุณลักษณะที่เข้ามานั้นมีค่าความถูกต้องเท่าใด จากนั้นจะเก็บคุณลักษณะที่เหมาะสมไว้ โดยในบทความนี้ผู้วิจัยใช้ ต้นไม้ตัดสินใจ 3 โมเดลช่วยกันตัดสินใจเพื่อเลือกคุณลักษณะที่ดีมาใช้ในการจำแนกข้อมูล (Priyank et al., 2020)

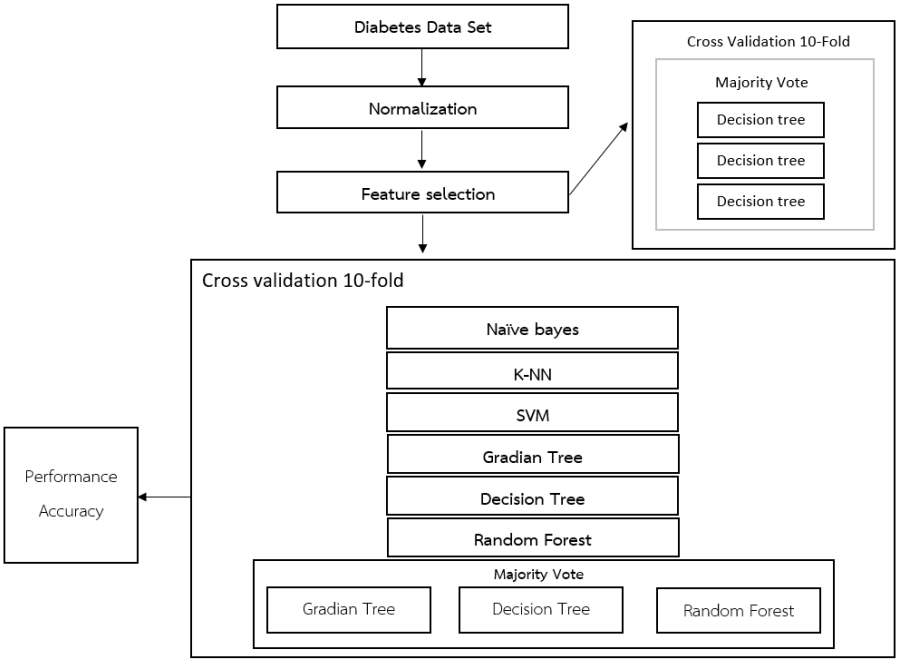
10. ขั้นตอนการวิจัย

10.1 การเก็บรวบรวมข้อมูลการวิจัยนี้ เป็นการทบทวนเวชระเบียนผู้ป่วยโรคเบาหวาน จากโรงพยาบาลสมเด็จพระยุพราช ตั้งแต่ปี 2557-2561 จำนวน 10,875 รายการ ประกอบด้วย 20 คุณลักษณะ ได้ทำความสะอาดข้อมูล กรณีที่ข้อมูลไม่สมบูรณ์ (Incomplete data) ให้ได้ข้อมูลที่สมบูรณ์คือ 1,435 รายการ 15 คุณลักษณะ แสดงดังตารางที่ 1 แบ่งเป็นข้อมูลผู้ป่วยที่เป็นโรคเบาหวาน 715 คน และ 720 คน คือกลุ่มที่ร่างกายปกติ เมื่อนำมาผ่านการคัดเลือกคุณสมบัติที่เหมาะสม เหลือคุณลักษณะที่เหมาะสมเพียง 6 คุณลักษณะที่จะนำไปใช้ในการจำแนกข้อมูล คือ FBS, HBA1C, Creatinine, FH, EGFR, Drinking โดยค่าความถูกต้องของชุดข้อมูลคือ 94.91%

ตารางที่ 1 คุณลักษณะของข้อมูลผู้ป่วยโรคเบาหวาน

คุณลักษณะ	ความหมาย
BPS	ความดันโลหิตตัวบน
BPD	ความดันโลหิตตัวล่าง
BW	น้ำหนัก
Height	ส่วนสูง
FBS	ค่าระดับน้ำตาลในเลือด
BMI	ดัชนีมวลกาย
TG	ไตรกลีเซอไรด์
HDL	ไขมันดี
EGFR	อัตราการกรองของเสียของไต
Creatinine	การทำงานของไต
HBA1C	น้ำตาลสะสมในเลือด
FH	กรรมพันธุ์ที่มีโรคเบาหวาน
Waist	รอบเอว
Smoking	บุหรี่
Drinking	สุรา
Outcome	ผลลัพธ์ 0 = ปกติ, 1= เป็นเบาหวาน

10.2 การสร้างแบบจำลองงานวิจัยนี้ผู้วิจัยได้ใช้เครื่องมือวิเคราะห์ข้อมูลคือโปรแกรม Rapid miner v.9.6 และ Jupyter Notebook v.6.3.0 ขั้นตอนวิธีการจำแนกข้อมูลที่ใช้ประกอบด้วย วิธีร่วมกันตัดสินใจโดยโหวตเสียงข้างมากจากต้นไม้ตัดสินใจ, วิธีเกรเดียนท์บูตทรีส์, วิธีต้นไม้ตัดสินใจ, วิธีแรนดอมฟอเรสต์, วิธีนาอีฟเบย์, วิธีซัพพอร์ทเวกเตอร์แมชชีน, วิธีความใกล้เคียงกันที่สุด และวิธีการคัดเลือกไปข้างหน้าเพื่อคัดเลือกคุณลักษณะที่เหมาะสมวัดประสิทธิภาพด้วยวิธีครอสวาลิเดชั่น การวัดประสิทธิภาพด้วยวิธีนี้จะทำการแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากันแล้วนำข้อมูลทีละส่วนเข้าทดสอบในแบบจำลองในแต่ละรอบและเก็บค่าเฉลี่ยไว้ ทำแบบนี้ไปเรื่อย ๆ จนกว่าจะครบ จากนั้นก็เอาค่าเฉลี่ยในแต่ละรอบมาหาค่าเฉลี่ยทั้งหมด ก็จะได้ค่าความถูกต้องในการทำนายของแบบจำลองแต่ละวิธี ภาพรวมในการสร้างแบบจำลองแสดงได้ดังรูปที่ 1



รูปที่ 1 ภาพรวมในการสร้างแบบจำลองจำแนกข้อมูล

ผลการวิจัยและอภิปรายผล

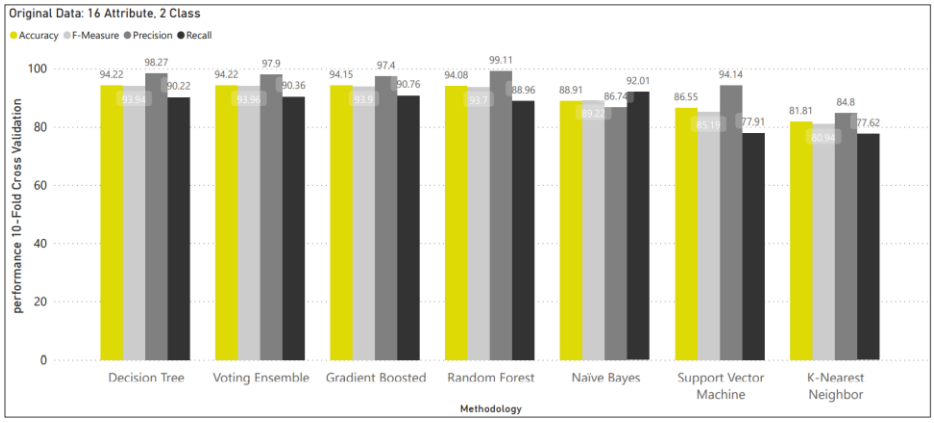
ผลการทดลองบทความฉบับนี้ได้วิเคราะห์ถึงค่าความถูกต้องของแบบจำลองจำแนกความเสี่ยงในการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูล โดยค่าความถูกต้องของการจำแนกข้อมูลที่ยังไม่ผ่านการเลือกคุณลักษณะแสดงได้ดังตารางที่ 2 และข้อมูลที่ผ่านการเลือกคุณลักษณะด้วยวิธีเลือกไปข้างหน้าแสดงได้ดังตารางที่ 3

ตารางที่ 2 ผลการวัดประสิทธิภาพของข้อมูล Original Data

Original Data: 16 Attribute, 2 Class				
Method	**Precision**	**Recall**	**Accuracy**	**F-Measure**
Voting Ensemble	97.9	90.36	94.22	93.96
Gradient Boosted	97.4	90.76	94.15	93.9
Decision Tree	98.27	90.22	94.22	93.94
Random Forest	99.11	88.96	94.08	93.7
Naïve Bayes	86.74	92.01	88.91	89.22
SVM	94.14	77.91	86.55	85.19
K-Nearest Neighbor	84.8	77.62	81.81	80.94

จากผลลัพธ์ที่ได้จากตารางที่ 2 เมื่อพิจารณาจากค่าความถูกต้อง พบว่าวิธีร่วมกันตัดสินใจมีประสิทธิภาพในการจำแนกข้อมูลมากที่สุด โดยมีค่า Accuracy 94.22%, Precision 97.90%, Recall 90.36% และ F-Measure 93.96% รองลงมาคือวิธีต้นไม้ตัดสินใจมีค่า Accuracy 94.22%, Precision 89.18%, Recall 92.45% F-Measure 93.94% วิธีเกรเดียนท์บูตทรีส์ มีค่า Accuracy 94.15%, Precision 97.40%, Recall 90.76% F-Measure 93.90% วิธีแรนดอมฟอเรสต์ มีค่า Accuracy 94.08%, Precision 99.11%, Recall 88.96% F-Measure 93.70% วิธีนาอีฟเบย์ มีค่า Accuracy 88.91%, Precision 86.74%, Recall 92.01% F-Measure 89.22% วิธีความใกล้เคียงกันที่สุดและวิธีซัพพอร์ทเวกเตอร์แมชชีนมีค่า Accuracy 81.81%, Precision 84.14%, Recall 77.62%, F-Measure 80.94% และมีค่า Accuracy 86.55%, Precision 94.14%, Recall 77.91% และ F-Measure 85.19% ตามลำดับ จากผลการวิจัยดังตารางที่ 2 พบว่าชุดข้อมูลที่ยังไม่

ทำการคัดเลือกคุณลักษณะ วิธีร่วมกันตัดสินใจโหวตเสียงข้างมากมีประสิทธิภาพความแม่นยำสูงกว่าขั้นตอนวิธีอื่น ๆ ดังรูปที่ 2



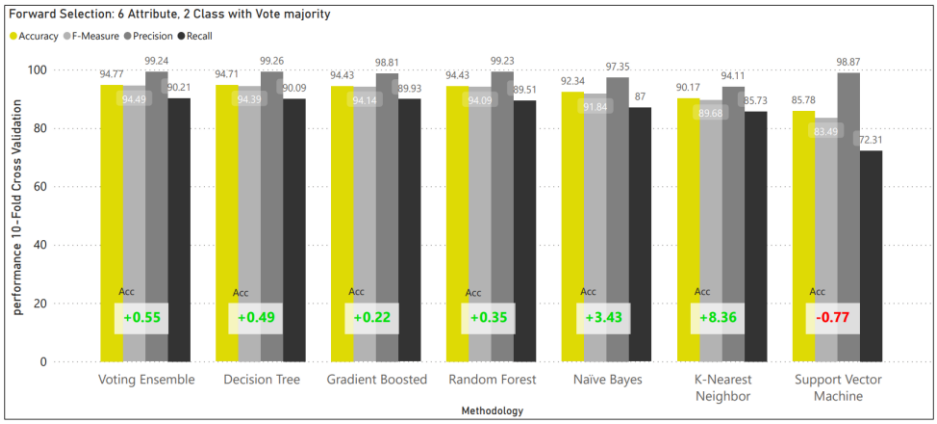
รูปที่ 2 ประสิทธิภาพของโมเดลด้วยวิธีครอสวาลิเดชั่นของข้อมูล Original Data

ตารางที่ 3 ผลการวัดประสิทธิภาพของโมเดลของข้อมูลที่ผ่านคัดเลือกคุณลักษณะ

Forward Selection: 6 Attribute, 2 Class with Vote majority				
Method	**Precision**	**Recall**	**Accuracy**	**F-Measure**
Voting Ensemble	99.24	90.21	94.77	94.49
Gradient Boosted	98.81	89.93	94.43	94.14
Decision Tree	99.26	90.09	94.71	94.39
Random Forest	99.23	89.51	94.43	94.09
Naïve Bayes	97.35	87	92.34	91.84
SVM	98.87	72.31	85.78	83.49
K-Nearest Neighbor	94.11	85.73	90.17	89.68

จากตารางที่ 3 เมื่อพิจารณาจากค่าความถูกต้องวิธีร่วมกันตัดสินใจมีค่า Accuracy 94.77%, Precision 99.24%, Recall 90.21% และ F-Measure 94.49% วิธีเกรเดียนท์บูตทรีส์ มีค่า Accuracy 94.43%, Precision 98.81%, Recall 89.93% F-Measure

94.14% วิธีต้นไม้ตัดสินใจมีค่า Accuracy 94.71%, Precision 99.26%, Recall 90.09%, F-Measure 94.39% วิธีแรนดอมฟอเรสต์ มีค่า Accuracy 94.43%, Precision 99.23%, Recall 89.51%, F-Measure 94.09% วิธีนาอีฟเบย์ มีค่า Accuracy 92.34%, Precision 97.35%, Recall 87.00%, F-Measure 91.84% วิธีความใกล้เคียงกันที่สุดและวิธีซัพพอร์ทเวกเตอร์แมชชีนมีค่า Accuracy 90.17%, Precision 94.11%, Recall 85.73%, F-Measure 89.68% และมีค่า Accuracy 85.78%, Precision 98.87%, Recall 72.31% และ F-Measure 83.49% ตามลำดับ เมื่อเปรียบเทียบกับการจำแนกข้อมูลที่ยังไม่เลือกคุณลักษณะพบว่าวิธีร่วมกันตัดสินใจมีค่า Accuracy เพิ่มขึ้น 0.55%, วิธีต้นไม้ตัดสินใจเพิ่มขึ้น 0.49%, วิธีเกรเดียนท์บูตทรีส์เพิ่มขึ้น 0.22%, วิธีแรนดอมฟอเรสต์เพิ่มขึ้น 0.35%, วิธีความใกล้เคียงกันที่สุดเพิ่มขึ้น 8.36%, วิธีนาอีฟเบย์เพิ่มขึ้น 3.43% และวิธีซัพพอร์ทเวกเตอร์มชชีนลดลง 0.77% ซึ่งค่าที่ลดลงนั้นเกิดจากเคอเนลที่เรานำเสนอไป อาจทำงานได้ไม่ดีถ้าความสัมพันธ์ของข้อมูลนั้นมีการเพิ่มหรือลดความซับซ้อนจนแบ่งด้วยเส้นตรงไม่ได้ (ควรเลือกใช้เคอเนลที่เหมาะสมกับข้อมูล) ทำให้มีค่าความถูกต้องของโมเดลลดลง แสดงผลการวัดประสิทธิภาพของโมเดลดังรูปที่ 3



รูปที่ 3 ประสิทธิภาพของโมเดลด้วยวิธีครอสวาลิเดชั่นของข้อมูลที่ผ่านคัดเลือกคุณลักษณะ

สรุปผลการวิจัย

บทความนี้มีวัตถุประสงค์เพื่อหาขั้นตอนวิธีที่เหมาะสมในการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานและเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีในการจำแนกข้อมูลด้วยการคัดเลือกคุณลักษณะเด่นไปข้างหน้า จำแนกข้อมูลด้วยวิธีร่วมกันตัดสินใจโดยโหวตเสียงข้างมากจากต้นไม้ตัดสินใจ, วิธีเกรเดียนท์บูตทรีส์, วิธีต้นไม้ตัดสินใจ, วิธีแรนดอมฟอเรสต์, วิธีนาอีฟเบย์, วิธีซัพพอร์ทเวกเตอร์แมชชีน, วิธีความใกล้เคียงกันที่สุด และใช้วิธีเลือกคุณลักษณะเด่นไปข้างหน้า วัดค่าความถูกต้องโดยใช้วิธีครอสวาลิเดชั่น จากผลการวิจัยสามารถสรุปผลได้สองส่วนดังนี้ ส่วนแรกคือการคัดเลือกคุณสมบัติที่เหมาะสมด้วยวิธีการเลือกไปข้างหน้า มีค่าความถูกของการจำแนกคุณลักษณะ 94.91% ลดคุณลักษณะจากข้อมูลเดิม 15 คุณลักษณะเหลือเพียง 6 คุณลักษณะ ในการคัดเลือกคุณลักษณะใช้วิธีต้นไม้ตัดสินใจ 3 โมเดลในการโหวตเลือกคุณลักษณะที่เหมาะสม ซึ่งเหตุผลในการเลือกเฉพาะต้นไม้ตัดสินใจ เนื่องจากข้อเสียของการโหวตเสียงข้างมากคือ ถ้าหากโมเดล 2 ใน 3 ทำนายผลลัพธ์เป็นคลาสบวกหรือ คลาสลบก็จะส่งผลต่อค่าความถูกต้องของการจำแนกข้อมูล และการเลือกใช้ตัวจำแนกที่ดีแต่ละตัวช่วยกันเสริมประสิทธิภาพในการจำแนกข้อมูลทำให้มีค่าความถูกต้องสูงขึ้น ส่วนที่สองคือการนำข้อมูลที่ผ่านการคัดเลือกคุณลักษณะมาจำแนกด้วยขั้นตอนวิธีต่าง ๆ เมื่อพิจารณาจากค่าความถูกต้องวิธีร่วมกันตัดสินใจมีค่า Accuracy 94.77% วิธีเกรเดียนท์บูตทรีส์ มีค่า Accuracy 94.43% วิธีต้นไม้ตัดสินใจมีค่า Accuracy 94.71% วิธีแรนดอมฟอเรสต์ มีค่า Accuracy 94.43% วิธีนาอีฟเบย์ มีค่า Accuracy 92.34% วิธีความใกล้เคียงกันที่สุดและวิธีซัพพอร์ทเวกเตอร์แมชชีนมีค่า Accuracy 90.17% และมีค่า Accuracy 85.78% ตามลำดับ จะเห็นได้ว่าวิธีร่วมกันตัดสินใจมีค่าความถูกต้องมากที่สุดเมื่อเทียบกับขั้นตอนวิธีอื่น ๆ ซึ่งข้อดีของวิธีนี้คือการนำตัวจำแนกข้อมูลที่หลากหลายมาช่วยกันตัดสินใจด้วยการโหวตเสียงข้างมากซึ่งจะช่วยในการลดปัญหาการเกิดความโน้มเอียงของข้อมูล นอกจากนี้ยังพบว่าการเลือกใช้คุณลักษณะที่เหมาะสมด้วยวิธีเลือกคุณลักษณะเด่นไปข้างหน้าทำให้โมเดลมีประสิทธิภาพในการจำแนกเพิ่มมากยิ่งขึ้น ผู้วิจัยจึงได้นำผลลัพธ์ที่ได้จากวิธีร่วมกันตัดสินใจไปใช้ในการศึกษาเพื่อพัฒนาระบบวินิจฉัยโรคเบาหวานและเป็นแนวทางในการสนับสนุนการตัดสินใจทางการแพทย์ให้มีประสิทธิภาพมากขึ้น

เอกสารอ้างอิง

Li, X., Zhao, Z., Gao, C., Rao, L., Hao, P., Jian, D., Li, W., Tang, H., & Li M., (2017). The diagnostic value of whole blood lncRNA ENST00000550337. 1 for prediabetes and type 2 diabetes mellitus. *Experimental and Clinical Endocrinology & Diabetes*, *125*(6), 377–383.

WHO & IDF. (2006). *Diabetes.mellitus.California*.[online]. Retrieved August 26, 2021, from Available: https://www.who.int/diabetes/publications/diagnosis_diabetes2006/en.

Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long noncoding RNAs expression: a comparison of four data mining approaches. *BMC Bioinformatics*, 21, 372- 385.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genet*ics, 9, 515-525.

Nagaratnam, A., Deepika, B., Sharoon, T., & Ajay, CH. (2020). Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques. *International Journal of Scientific and Research Publications*, 10(5), 984-987.

Lan, H., & Pan, Y. (2019). *A Crowdsourcing quality prediction model based on random forests*. In: Proceedings of 18th International Conference on Computer and Information Science (ICIS), 17-19 June 2019, Beijing, China. 315-319.

Dutta, J., Yong Woon K., & Dalia, D. (2020). *Comparison of gradient boosting and extreme boosting ensemble methods for webpage classification*. In: Proceedings of Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 26 – 27 November 2020, Bangalore, India. 77-82.

Priyanka, S., Srabani, P., & Sarmistha, N. (2020*). A Correlation - Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis*. In: Proceedings of IEEE International Conference on Computing, Power and Communication Technologies (GUCON), 2-4 October, 2020, Greater Noida, India. 69-72.