วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 33 ฉบับที่ 2 เม.ย.–มิ.ย. 2566
The Journal of KMUTNB., Vol. 33, No. 2, Apr.–Jun. 2023

บทความวิจัย

การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล

นพรัตน์ นนท์ศิริ
สาขาวิทยาการข้อมูลและเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี
ราตรี มนัสศิลา
โรงพยาบาลสมเด็จพระยุพราชบ้านดุง
กริช สมกันธา*
กลุ่มวิจัยปัญญาประดิษฐ์และวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 3152 4445 อีเมล: dr_krit@udru.ac.th DOI: 10.14416/j.kmutnb.2022.10.004
รับเมื่อ 19 กุมภาพันธ์ 2564 แก้ไขเมื่อ 28 เมษายน 2564 ตอบรับเมื่อ 16 มิถุนายน 2564 เผยแพร่ออนไลน์ 7 ตุลาคม 2565

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล 4 วิธี ซึ่งประกอบด้วย วิธีนาอีฟเบย์ (Naïve Bayes) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีความใกล้เคียงกันที่สุด (K-Nearest Neighbor) และวิธีต้นไม้ตัดสินใจ (Decision Tree) โดยใช้ข้อมูลของผู้ป่วยโรคเบาหวานโรงพยาบาลสมเด็จพระยุพราชบ้านดุง สร้างชุดตัวแบบและชุดทดสอบตัวแบบ เป็นข้อมูลที่เกิดจากการทบทวนเวชระเบียนผู้ป่วยโรคเบาหวานย้อนหลัง จำนวน 1,435 ชุดข้อมูล 16 คุณลักษณะ จากนั้นทำการหาค่าความถูกต้องของแบบจำลอง (Accuracy) โดยใช้วิธี 10- Fold Cross Validation ผลการเปรียบเทียบพบว่า วิธีต้นไม้ตัดสินใจให้ค่าประสิทธิภาพสูงสุด โดยมีค่าความถูกต้อง 93.73% วิธีนาอีฟเบย์ค่าความถูกต้อง 88.92% วิธีความใกล้เคียงกันที่สุดและวิธีซัพพอร์ตเวกเตอร์แมชชีนค่าความถูกต้อง 86.97% และ 86.13% ตามลำดับ จะพบว่า วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการสร้างแบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน เนื่องจากเป็นวิธีที่ไม่มีการแจกแจงหรือไม่ใช้พารามิเตอร์ซึ่งไม่ได้ขึ้นอยู่กับสมมุติฐานการแจกแจงความน่าจะเป็น อีกทั้งสามารถจัดการกับข้อมูลที่มีมิติสูงได้อย่างแม่นยำ เหมาะสมที่จะนำแบบจำลองไปพัฒนาระบบจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน เพื่อเป็นแนวทางในการสนับสนุนการตัดสินใจทางการแพทย์ในการวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานต่อไป

คำสำคัญ: การทำเหมืองข้อมูล โรคเบาหวาน การจำแนกประเภทผู้ป่วยโรคเบาหวาน

การอ้างอิงบทความ: นพรัตน์ นนท์ศิริ, ราตรี มนัสศิลา และ กริช สมกันธา, "การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 33, ฉบับที่ 2, หน้า 538–547, เม.ย.–มิ.ย. 2566.

Research Article

Data Classifying to Diagnose Diabetes Risk Using Data Mining Techniques

Nopparat Nonsiri
Department of Data Science, Faculty of Science, Udon Thani Rajabhat University, Udon Thani, Thailand
Ratree Manassila
Somdej Phra Yupparat Ban Dung Hospital, Udon Thani, Thailand
Krit Somkanta*
Artificial Intelligence and Data Science Research Group, Faculty of Science, Udon Thani Rajabhat University, Udon Thani, Thailand

* Corresponding Author, Tel. 08 3152 4445, E–mail: dr_krit@udru.ac.th
DOI: 10.14416/j.kmutnb.2022.10.004
Received 19 February 2021; Revised 28 April 2021; Accepted 16 June 2021; Published online: 7 October 2022

Abstract

This research aims to create a data classification model for diagnosing diabetes risk by using four data mining techniques, which are Naïve Bayes Method, Support Vector Machine Method, K-Nearest Neighbor Method, and Decision Tree Method. The study employed data on diabetic patients from Somdej Phra Yuparat Hospital, Ban Dung to create a model and a model test kit. The data was derived from a retrospective review of diabetes medical records of 1,435 data sets with 16 attributes. Then the accuracy of the model was determined using the 10-fold cross validation method. The decision tree method yielded the highest efficiency with 93.73% accuracy, Naïve Bay method of 88.92% accuracy, closest approximation, and support vector machine method accuracy values of 86.97% and 86.13% respectively. It was found that the decision tree method was the most efficient in modeling compared to the comparative approach. This is because it is a non-distribution or nonparametric method which does not depend on the probability distribution hypothesis. It can also handle high-dimensional data with precision. It is appropriate to use the model to develop a classification system for diagnosing diabetes risk and as a guideline to support medical decision-making in the diagnosis of diabetes risk.

Keywords: Data Mining, Diabetes, Diabetes Classification

Please cite this article as: N. Nonsiri, R. Manassila, and K. Somkanta, "Data classifying to diagnose diabetes risk using data mining techniques," *The Journal of KMUTNB*, vol. 33, no. 2, pp. 538–547, Apr.–Jun. 2023 (in Thai).

1. บทนำ

จากข้อมูลของสหพันธ์เบาหวานนานาชาติ พบผู้ป่วยโรคเบาหวานทั่วโลกราว 425 ล้านคน ใน พ.ศ. 2560 และคาดการณ์ว่าจะมีจำนวนผู้ป่วยด้วยโรคนี้มากถึง 520 ล้านคน ใน พ.ศ. 2578 [1] สำหรับสถานการณ์โรคเบาหวานในประเทศไทยพบว่า คนไทยช่วงอายุ 20–79 ปี เป็นโรคเบาหวานร้อยละ 8.3 หรือหมายความว่าใน 100 คน จะพบคนที่ป่วยเป็นโรคเบาหวานประมาณ 8 คน และจำนวนมากกว่าครึ่งไม่ทราบว่าตนเองเป็นโรคเบาหวาน ซึ่งผู้ที่อยู่ในกลุ่มเสี่ยงต่อภาวะการเป็นเบาหวาน สามารถพัฒนาการเกิดโรคเบาหวานประเภทที่ 2 ได้ และพบว่า ผู้ป่วยโรคเบาหวานมีภาวะแทรกซ้อนเนื่องจากไตเสื่อมสูงสุดถึงร้อยละ 43.9 ต้อกระจกร้อยละ 42.8 และจอประสาทตาเสื่อมร้อยละ 30.7 และพบมีภาวะแทรกซ้อนจากโรคหัวใจขาดเลือดและโรคหลอดเลือดสมองร้อยละ 8.1 และ 4.4 ตามลำดับ [2] ปัจจัยที่เกี่ยวข้องกับภาวะแทรกซ้อนของโรคเบาหวาน ได้แก่ ระยะเวลาการเป็นโรคเบาหวาน อายุ ดัชนีมวลกาย การสูบบุหรี่ ผลกระทบของการสูบบุหรี่ทำให้การดื้ออินซูลิน ก่อเกิดเป็นกลุ่มอาการเมแทบอลิก มีภาวะแทรกซ้อนของหลอดเลือดเล็กและหลอดเลือดใหญ่ [3] และมีการศึกษาความเสี่ยงของการเกิดโรคเบาหวานด้วยลักษณะความเสี่ยงด้วยชุดข้อมูลต่างๆ ในการวิเคราะห์ความเสี่ยงการเป็นโรคเบาหวาน [4], [5] และนำชุดข้อมูลนั้นมาผ่านกระบวนการของเครื่องจักรการเรียนรู้

โรงพยาบาลสมเด็จพระยุพราช เป็นโรงพยาบาลชุมชนประจำอำเภอสังกัดกระทรวงสาธารณสุข มีขีดความสามารถระดับปฐมภูมิ (Primary Care) และระดับทุติยภูมิ (Secondary Care) ซึ่งปัจจุบันกำลังประสบปัญหาโรคเรื้อรังที่เกี่ยวข้องกับการไม่ปฏิบัติตนตามพฤติกรรมสุขภาพที่เหมาะสมของคนในพื้นที่ เช่น รับประทานอาหารที่มีประโยชน์อย่างเหมาะสม ออกกำลังกายแบบแอโรบิค จากสภาวการณ์ปัจจุบันนี้เองก่อเกิดปัจจัยเสี่ยงด้านสุขภาพมากมาย ทำให้ประชาชนในท้องถิ่นมีแนวโน้มเจ็บป่วยด้วยโรคเรื้อรังมากขึ้น นั่นก็คือโรคเบาหวานซึ่งเป็นหนึ่งในโรคที่สำคัญที่ทำให้เกิดภาวะแทรกซ้อนของโรคอื่น อีกทั้งทางโรงพยาบาลยังขาดแคลนบุคลากรทางการแพทย์ โดยเฉพาะทีมบริการเฉพาะโรคเมื่อโรงพยาบาลออกบริการชุมชน

ดังนั้นคณะผู้วิจัยจึงเห็นถึงปัญหาและความสำคัญในการศึกษาสถานการณ์สุขภาพของผู้ป่วยโรคเบาหวานในโรงพยาบาลสมเด็จพระยุพราชบ้านดุง มีวัตถุประสงค์เพื่อสร้างแบบจำลองการจำแนกข้อมูล เพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน และเปรียบเทียบประสิทธิภาพของแบบจำลอง เพื่อที่จะนำแบบจำลองที่ดีที่สุดไปพัฒนาระบบวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน ให้สามารถตรวจสอบความเสี่ยงในการเกิดโรคเบื้องต้น แล้วนำผลไปใช้กำหนดแนวทางส่งเสริมสุขภาวะที่ดีของกลุ่มผู้ที่มีแนวโน้มที่จะป่วยเป็นโรคเบาหวานในอนาคต ตลอดทั้งการวางแผนรองรับการรักษาโรคเบาหวาน ซึ่งนับเป็นเรื่องที่สำคัญที่จะต้องเร่งรัดดำเนินการ และจะต้องให้ความสนใจในสาเหตุหรือปัจจัยที่มีความสัมพันธ์ต่อการเกิดโรค หากผู้ป่วยเกิดพัฒนาเป็นโรคเบาหวานประเภทที่ 2 จะทำให้มีอัตราการเสียชีวิตเพิ่มขึ้นถึงร้อยละ 90 เนื่องจากเกิดอาการแทรกซ้อนของโรคอื่นๆ [6]

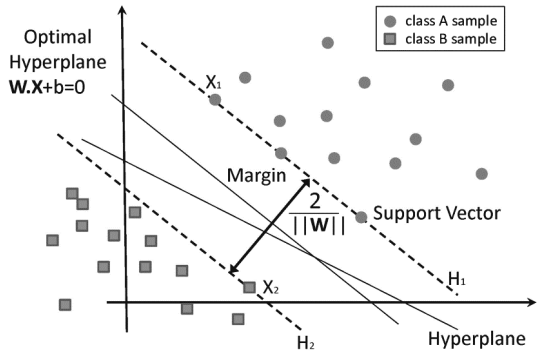
2. วัสดุ อุปกรณ์และวิธีวิจัย

2.1 การทำเหมืองข้อมูล (Data Mining)

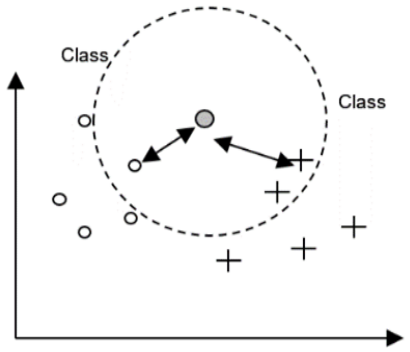
การทำเหมืองข้อมูล คือ การวิเคราะห์ข้อมูลเพื่อหารูปแบบ (Patterns) หรือความสัมพันธ์ของข้อมูลสารสนเทศ ซึ่งปัจจุบันมีการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลมาใช้ในด้านการแพทย์ เพื่อหาสาเหตุหรือปัจจัยที่มีความสัมพันธ์ต่อความเสี่ยงของโรคต่างๆ เช่น การพยากรณ์ผู้ป่วยโรคเบาหวาน การจำแนกผู้ป่วยโรคไต เพื่อนำข้อมูลมาใช้ในการบริหารจัดการและควบคุมดูแล และการวางแผน เพื่อบริหารจัดการดูแลด้านสุขภาวะของประชากร

2.2 เทคนิคเหมืองข้อมูล

เทคนิคเหมืองข้อมูลในปัจจุบันมีหลายรูปแบบทั้งเทคนิคแบบผู้สอน (Supervised Learning) แบบไม่มีผู้สอน (Unsupervised Learning) แบบการเรียนรู้เกิดมาจากการปฏิสัมพันธ์ (Reinforcement Learning) ซึ่งผู้วิจัยได้ศึกษาเทคนิคที่ใช้ทำเหมืองข้อมูลเพื่อการจำแนกกลุ่มของข้อมูล ได้แก่ วิธีนาอีฟเบย์ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีความ



รูปที่ 1 ขั้นตอนการทำงานของซัพพอร์ตเวกเตอร์แมชชีน



รูปที่ 2 ขั้นตอนการทำงานของวิธีที่ใกล้เคียงกันที่สุด

ใกล้เคียงกันที่สุด และวิธีต้นไม้ตัดสินใจ เนื่องจากเทคนิคดังกล่าวเป็นเทคนิคที่นิยม และเหมาะกับการจำแนกข้อมูลที่เป็นหมวดหมู่ข้อมูลต่างๆ ที่ถูกแนบอยู่ในแต่ละระเบียน (Record) ของชุดข้อมูล โดยการเรียนรู้แบบมีผู้สอนจะแตกต่างกับการเรียนรู้แบบไม่มีผู้สอนที่จะไม่ทราบถึงหมวดหมู่ของข้อมูลตัวอย่างเช่น ในการวิเคราะห์ข้อมูลโรคเบาหวานที่ไม่มีข้อมูลที่บ่งบอกว่าเป็นโรคเบาหวาน ก็จะนำข้อมูลปัจจัยไปวิเคราะห์หาความเสี่ยงว่าเป็นโรคเบาหวานหรือไม่ และข้อมูลที่ผู้วิจัยได้ทำการรวบรวมนั้น มีหมวดหมู่ของข้อมูลที่บ่งบอกว่าเป็นโรคเบาหวานอย่างชัดเจน ผู้วิจัยจึงได้เลือกใช้เทคนิควิธีดังกล่าวข้างต้น โดยมีรายละเอียดดังนี้

1) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นวิธีที่ใช้จำแนกค่าคุณลักษณะของ 2 กลุ่ม โดยจะสร้างเส้นแบ่ง (Plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่ม ออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (Margin) ออกไปทั้งสองข้างออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุดโดยอาศัยการปรับค่าสมการด้วยวิธีการเคอร์เนล (แสดงดังรูปที่ 1) [7] ในการหารูปแบบและความสัมพันธ์ เพื่อให้ได้ผลลัพธ์ที่มีต้องแม่นยำสูง เคอร์เนลฟังก์ชันมีอยู่เป็นจำนวนมากที่รู้จักกันดี เช่น Polynomial, RBF หรือ Sigmoid โดยผู้วิจัยได้เลือกใช้เคอร์เนลรวมแบบดอท

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

$$(w^{t}\times x)+b$$

$$(w^{t}\times x)+b>0 \text{ ถ้า } y_{i}=+1 \quad (2)$$

$$\text{และ } (w^{t}*x)+b<0 \text{ ถ้า } y_{i}=-1 \quad (3)$$

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

w^{t} แทน ค่าน้ำหนักที่เชื่อมโยงจาก Feature

b แทน ค่าโน้มเอียง (Bias)

n แทน จำนวนข้อมูลตัวอย่าง

m แทน จำนวนมิติข้อมูล

y แทน ผลลัพธ์กลุ่มข้อมูลมีค่า +1 หรือ –1

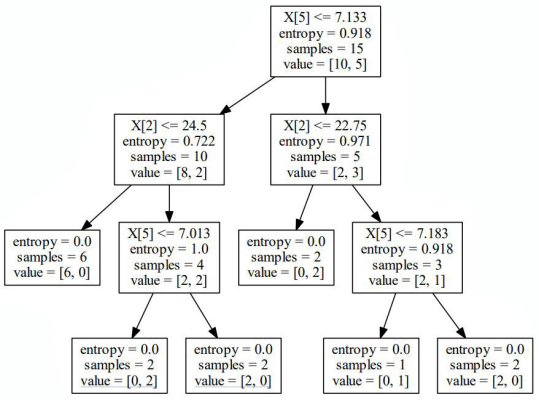
2) วิธีความใกล้เคียงกันที่สุด (K-Nearest Neighbor) เป็นวิธีการหนึ่งสำหรับแก้ปัญหา การประมาณค่าที่ไม่ใช่พารามิเตอร์สำหรับการจำแนกกลุ่ม [7] ซึ่งเหมาะกับข้อมูลที่เป็นรูปร่างที่ดี หลักการคือจะวัดค่าของข้อมูลที่ใกล้ที่สุดเพื่อหาจุดได้เท่ากับจำนวน k ตัว ที่ต้องการ เช่น k = 5 ก็จะขยายเคอร์เนล ไปจนกว่าจะเจอข้อมูลจำนวน k 5 ตัว จึงจะหยุด (แสดงดังรูปที่ 2) จากนั้นก็ทำการจำแนกว่ามันอยู่ใกล้กับข้อมูลจุดไหนมากสุด และหาค่าได้จากสมการดังนี้

$$F_{n}(x/w_{i})=\frac{P(x\ Fall\ in\ R(x)\ with\ volume\ V_{n})}{V_{n}} \quad (4)$$

$$F_{n}(x/w_{i})=\frac{k_{n}/n}{V_{n}} \quad (5)$$

$$k_{n}=\sqrt{n}$$

$$F_{n}(x/w_{i})=\frac{1/\sqrt{n}}{V_{n}}\approx f(x) \quad (6)$$



รูปที่ 3 ขั้นตอนการทำงานของวิธีต้นไม้ตัดสินใจ

$$V_n = \frac{1}{\sqrt{n} f(x)} \tag{7}$$

k_n แทน จำนวนของข้อมูลที่ต้องการที่อยู่ในหน้าต่าง

n แทน จำนวนข้อมูลทั้งหมด

v_n คือ ขนาดของหน้าต่าง ที่ทำการขยายไปจาก Bay's จะได้สมการดังต่อไปนี้

$$P(w_n / x) = \frac{F(x / w_i)}{F(x)} = \frac{F(x / w_i)}{\sum_{j=1}^{c} F(x / w_i)} = \frac{(k_i / n) / V}{\sum_{j=1}^{c} (k_i / n) / V} = \frac{k_i}{\sum_{j=1}^{c} k_i} P(w_n / x) = \frac{k_i}{k} \tag{8}$$

k คือ จำนวนของข้อมูลทั้งหมดของคลาส i

k_i คือ จำนวนของข้อมูลทั้งหมดทุกคลาส

3) วิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นการจำแนกค่าคุณลักษะของข้อมูล โดยจะประกอบด้วยบัพ (Node) และกิ่ง (Link) ที่ต่อกับบัพ บัพที่ปลายสุดจะเรียกว่าบัพใบ (Leaf) ต้นไม้ตัดสินใจจะทำโดยสร้างบัพทีละบัพเพื่อตรวจสอบคุณสมบัติของตัวอย่าง แล้วแยกตัวอย่างลงตามค่าของกิ่ง ทำจนกระทั่งตัวอย่างในใบแต่ละใบอยู่ในประเภทเดียวกันทั้งหมด แสดงดังรูปที่ 3 [8]

$$I(s_1, s_{2,\ldots}, s_n) = -\sum_{i=1}^{n} \frac{s^1}{s} \log_2 \frac{s^1}{s} \tag{9}$$

Algorithm: Naïve-Bayes

- **Naive_Bayes_Learn(examples)**
  FOR EACH target value v DO
  $\overline{P}(v_j) \leftarrow$ estimate $P(v_j)$
  FOR EACH attribute value a of each attribute DO
  $\overline{P}(a_i \mid v_j) \leftarrow$ estimate $P(a_i \mid v_j)$
- **Classify_New_Example(x)**
  $v_{NB} = \arg\max_{v_j \in V} \overline{P}(v_j) \times \prod_{i=1}^{n} \overline{P}(a_i \mid v_j)$

รูปที่ 4 ขั้นตอนการทำงานของวิธีนาอีฟเบย์

s เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s ระเบียน

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

c_i แทนกลุ่มในลำดับ ที่ i โดย ที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวน ข้อมูลสมาชิกของ s และอยู่ในกลุ่ม c_i

$$E(A) = \sum_{j=1}^{n} \frac{s_{1j+\ldots+} s_{nj}}{s} I(s_{1j}, s_{2j}, \ldots s_{nj}) \tag{10}$$

$$Gain(A) = I(s_{1j}, s_{2j}, \ldots, s_{nj}) - E(A) \tag{11}$$

4) วิธีนาอีฟเบย์ (Naïve Bayes) เป็นตัวจำแนกที่เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมาก และคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน [9] มีการนำตัวจำแนกประเภทเบย์ไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความ (Text Classification) และพบว่า ใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีอื่นๆ มีขั้นตอนการทำงานแสดงดังรูปที่ 4

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c) \times \ldots \times P(x_n \mid c) \times P(c) \tag{12}$$

c คือ คลาสของข้อมูล

x คือ แอตทริบิวต์

p คือ ความน่าจะเป็นของข้อมูล

$P(c|x)$ คือ ความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c

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

$P(c)$ คือ จำนวนคลาสที่อาจจะเกิดขึ้นหารด้วยจำนวนคลาสทั้งหมดของคลาส c

$P(x)$ คือ จำนวนแอตทริบิวต์ทั้งหมด

5) การวัดประสิทธิภาพแบบจำลองโดยใช้วิธี K-Fold Cross Validation วิธีนี้เป็นวิธีที่นิยมในการทำงานวิจัยเพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลอง เนื่องจากผลที่ได้มีความน่าเชื่อถือมาก การวัดประสิทธิภาพด้วยวิธี Cross-validation คือ ทำการแบ่งข้อมูลออกเป็นส่วนๆ โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน เพื่อทดสอบค่าความถูกต้องของโมเดลโดยพิจารณาทุกกลุ่มข้อมูล คำนวณได้จากสมการ

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

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

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

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

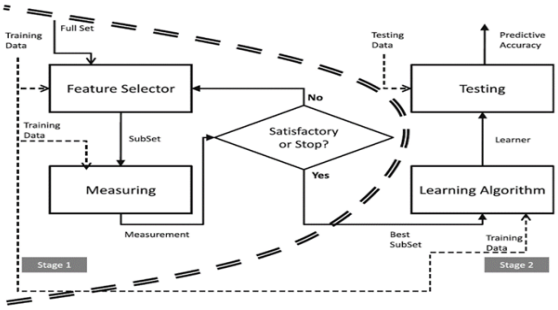
TP = อัตราความถูกต้องเชิงบวก
TN = อัตราความถูกต้องเชิงลบ
FP = อัตราความผิดพลาดเชิงบวก
FN = อัตราความผิดพลาดเชิงลบ

2.3 วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงประยุกต์ (Applied Research) ซึ่งมีการนำข้อมูลทุติยภูมิของผู้ป่วยโรคเบาหวานมาประยุกต์ใช้งานกับเทคนิคการทำเหมืองข้อมูลซึ่งจะสร้างแบบจำลองในการทำนายคุณลักษณะความเสี่ยงของผู้ที่จะเป็นโรคเบาหวาน เป็นการศึกษาแบบ Cross-sectional Study มีรายละเอียดในการศึกษาวิจัยดังต่อไปนี้

2.4 การเตรียมข้อมูล

การเก็บรวบรวมข้อมูลการวิจัยนี้ เป็นการทบทวนเวชระเบียนผู้ป่วยโรคเบาหวานจากโรงพยาบาลสมเด็จพระยุพราช ตั้งแต่ พ.ศ. 2557–2561 จำนวน 10,875 รายการ ประกอบด้วย 20 คุณลักษณะ ได้ทำความสะอาดข้อมูลกรณีที่ข้อมูลไม่สมบูรณ์ (Incomplete Data) ให้เหลือเฉพาะข้อมูลที่มีคุณลักษณะความเสี่ยงที่จะเกิดโรค เมื่อทำความสะอาดข้อมูลแล้ว เหลือข้อมูลที่สมบูรณ์คือ 1,435 รายการ 16 คุณลักษณะ ประกอบด้วย bps, bpd, bw, height, fbs, bmi, tg, hdl, creatinine,hba1c, fh, waist, smoking_type_id, drinking_type_id, egfr, outcome แบ่งเป็นข้อมูลผู้ป่วยที่เป็นโรคเบาหวาน 715 คน และ 720 คน คือกลุ่มที่ร่างกายปกติ และก่อนการเก็บรวบรวมข้อมูลจะเลือกเวชระเบียนของผู้ป่วยจากฐานข้อมูลผู้ป่วยที่มาตรวจรักษาตามนัด โดยเก็บข้อมูลไปตามลำดับผู้ป่วยที่มาตามนัดจนครบตามจำนวนที่กำหนดในแต่ละครั้งโดยไม่มีการข้ามเวชระเบียน ยกเว้นผู้ป่วยไม่มาตามนัด และไม่สามารถติดตามผู้ป่วยให้มารักษาต่อเนื่องได้ ทั้งนี้ผู้วิจัยได้มีการขอจริยธรรมการวิจัยในมนุษย์และได้ปฏิบัติตามกฎระเบียบ ข้อบังคับ แนวทางมาตรฐานวิชาชีพ และวางแผนการควบคุมคุณภาพของการเก็บข้อมูลอย่างเป็นระบบ เพื่อให้ได้ข้อมูลที่ถูกต้องตามความเป็นจริงมากที่สุด โดยมีขั้นตอนการดำเนินการคือ ภายหลังการจัดเก็บข้อมูลจากประวัติผู้ป่วยในโรงพยาบาล จะให้เจ้าหน้าที่อีกท่านหนึ่งซึ่งเป็นพยาบาลชำนาญการ กรณีผู้ป่วยโรคเบาหวานและความดันโลหิตสูง และเป็นผู้วิจัยร่วมดำเนินการตรวจสอบข้อมูลว่าเป็นไปตามกระบวนการวิจัยที่ได้ระบุไว้อย่างละเอียดแล้วหรือไม่ หากพบข้อสงสัย หรือข้อผิดพลาดจะทำการซักถามไปยังผู้บันทึกข้อมูลเพื่อแก้ไขให้ตรงตามความเป็นจริงที่ปรากฏในเวชระเบียนของผู้ป่วย ทั้งนี้ เพื่อให้ได้แบบจำลองระบบจำแนกผู้ป่วยโรคเบาหวาน และเลือกเทคนิคเหมืองข้อมูลที่ดีที่สุดมาพัฒนาระบบ เพื่อนำมาใช้งานจริงในคลินิกผู้ป่วยโรคเบาหวานและความดันโลหิตสูงในโรงพยาบาลสมเด็จพระยุพราชบ้านดุง ผู้วิจัยจึงได้มีการควบคุมคุณภาพข้อมูล และความถูกต้องของข้อมูลอย่างเคร่งครัด โดยคุณลักษณะของข้อมูลผู้ป่วยโรคเบาหวานสามารถแสดงดังตารางที่ 1



รูปที่ 5 ขั้นตอนการสร้างแบบจำลอง

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

2.5 การสร้างแบบจำลอง

ในการสร้างแบบจำลองงานวิจัยนี้ผู้วิจัยได้ใช้เครื่องมือวิเคราะห์ข้อมูลคือ โปรแกรม RapidMiner v.9.6 และเทคนิควิธีการจำแนกกลุ่มข้อมูลที่ใช้ประกอบด้วย วิธีเนอีฟเบย์ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีความใกล้เคียงกันที่สุด ต้นไม้ตัดสินใจ และการสร้างแบบจำลอง แสดงดังรูปที่ 5

2.6 การวิเคราะห์ประสิทธิภาพ

วัดประสิทธิภาพแบบจำลองโดยใช้วิธี 10-Fold Cross Validation ในการทดสอบประสิทธิภาพของโมเดล การวัดประสิทธิภาพด้วยวิธีนี้จะทำการแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน [10] แล้วนำข้อมูลทีละส่วนเข้าทดสอบในแบบจำลองในแต่ละรอบและเก็บค่าเฉลี่ยไว้ ทำแบบนี้ไปเรื่อยๆ จนกว่าจะครบ จากนั้นก็เอาค่าเฉลี่ยในแต่ละรอบมาหาค่าเฉลี่ยทั้งหมด ก็จะได้ค่าความถูกต้องในการทำนายของแบบจำลองแต่ละเทคนิค แสดงดังตารางที่ 2 และการเปรียบเทียบค่าความถูกต้องแบบจำลองของผู้ป่วยโรคเบาหวาน แสดงดังตารางที่ 3

3. ผลการทดลอง

ผลการทดลองงานวิจัยฉบับนี้ได้วิเคราะห์ถึงค่าความถูกต้องของแบบจำลองพยากรณ์ความเสี่ยงในการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลสามารถแสดงค่าความถูกต้องของการพยากรณ์ได้ดังตารางที่ 2

ตารางที่ 2 ผลการวัดประสิทธิภาพของโมเดลด้วย 10-Fold Cross Validation ในแต่ละรอบ

Number of Folds	KNN	SVM	Decision Tree	Naïve Bayes
1	0.8541	0.8472	0.9236	0.8680
2	0.8601	0.8601	0.9300	0.8951
3	0.9097	0.8611	0.9444	0.8750
4	0.8391	0.8531	0.9090	0.8461
5	0.8888	0.9027	0.9375	0.8472
6	0.8181	0.7832	0.9090	0.8741
7	0.9300	0.9300	0.9790	0.9300
8	0.8958	0.8611	0.9375	0.9027
9	0.8741	0.8601	0.9580	0.9510
10	0.8263	0.8541	0.9444	0.9027
Accuracy	**0.8696**	**0.8613**	**0.9372**	**0.8892**

จากตารางที่ 2 จะเห็นได้ว่าผลการทดสอบประสิทธิภาพของโมเดล แต่ละส่วนจะมีค่าความถูกต้องในแต่ละรอบ เมื่อทดสอบครบจำนวน 10 พับ (Fold) จึงจะเฉลี่ยผลค่าความถูกต้องของแบบจำลองแต่ละวิธี [11]

ตารางที่ 3 การเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองความเสี่ยงการเป็นโรคเบาหวาน

	Precision	Recall	Accuracy	F-Measure
Decision Tree	94.04%	89.93%	93.73%	93.44%
Naïve Bayes	89.18%	92.45%	88.92%	89.32%
SVM	87.98%	78.18%	86.13%	85.67%
KNN	87.93%	81.81%	86.97%	81.81%

จากตารางที่ 3 เมื่อพิจารณาจากค่าความถูกต้อง พบว่า วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการจำแนกข้อมูลมากที่สุด โดยมีค่า Accuracy 93.73%, Precision 94.04%, Recall 89.93% และ F-Measure 93.44% รองลงมาคือวิธีนาอีฟเบย์ มีค่า Accuracy 88.92%, Precision 89.18%, Recall 92.45% และ F-Measure 89.32% วิธีความใกล้เคียงกันที่สุด และวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่า Accuracy 86.97%, Precision 87.93%, Recall 81.81%, F-Measure 85.67% และมีค่า Accuracy 86.13%, Precision 87.98%, Recall 78.18% และ F-Measure 81.81% ตามลำดับ จากผลการวิจัยดังตารางที่ 2 พบว่า วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการสร้างแบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน จะได้โครงสร้างต้นไม้ตัดสินใจ แสดงดังรูปที่ 6 และกฎการจำแนกด้วยต้นไม้ตัดสินใจ 11 กฎ แสดงดังรูปที่ 7

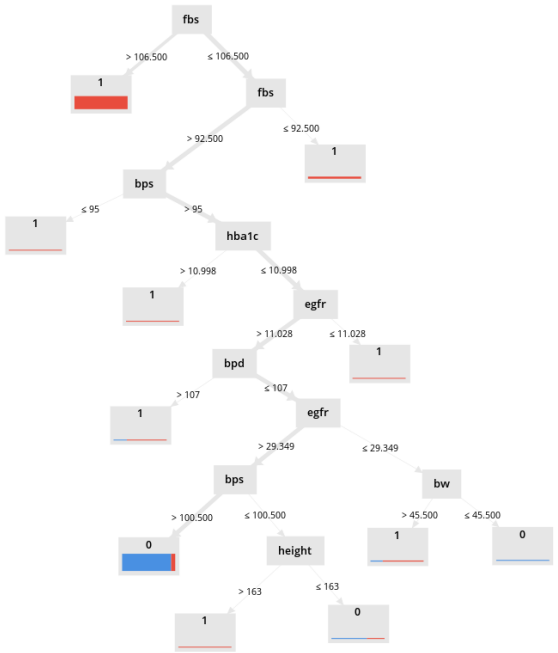
จากโครงสร้างต้นไม้ตัดสินใจที่แสดงดังรูปที่ 7 สามารถสรุปเป็นกฎได้ ดังนี้

ยกตัวอย่างกฎการจำแนกด้วยต้นไม้ตัดสินใจ เช่น

กฎข้อที่ 1 ถ้า fbs (ระดับน้ำตาลในเลือด) มากกว่า 106.500 ผลลัพธ์คือเป็นโรคเบาหวาน

กฎข้อที่ 2 ถ้า fbs น้อยกว่าหรือเท่ากับ 106.500 ให้ไปดูต่อว่า fbs มากกว่า 92.500 ถ้ามากกว่า ให้ไปดู ค่า bps (ความดันโลหิตตัวบน) น้อยกว่าหรือเท่ากับ 95 ผลลัพธ์ คือเป็นโรคเบาหวาน

กฎข้อที่ 3 ถ้า bps มากกว่าหรือเท่ากับ 95 ให้ไปดู hba1c (น้ำตาลสะสมในเลือด) มากกว่า 10.998 ผลลัพธ์คือเป็นโรคเบาหวาน



รูปที่ 6 แบบจำลองการจำแนกผู้ป่วยเบาหวานด้วยต้นไม้ตัดสินใจ

```
fbs > 106.500: 1 {0=0, 1=572}
fbs ≤ 106.500
|   fbs > 92.500
|   |   bps > 95
|   |   |   hba1c > 10.998: 1 {0=0, 1=2}
|   |   |   hba1c ≤ 10.998
|   |   |   |   egfr > 11.028
|   |   |   |   |   bpd > 107: 1 {0=1, 1=3}
|   |   |   |   |   bpd ≤ 107
|   |   |   |   |   |   egfr > 29.349
|   |   |   |   |   |   |   bps > 100.500: 0 {0=707, 1=59}
|   |   |   |   |   |   |   bps ≤ 100.500
|   |   |   |   |   |   |   |   height > 163: 1 {0=0, 1=2}
|   |   |   |   |   |   |   |   height ≤ 163: 0 {0=4, 1=2}
|   |   |   |   |   |   egfr ≤ 29.349
|   |   |   |   |   |   |   bw > 45.500: 1 {0=3, 1=10}
|   |   |   |   |   |   |   bw ≤ 45.500: 0 {0=5, 1=0}
|   |   |   |   egfr ≤ 11.028: 1 {0=0, 1=2}
|   |   bps ≤ 95: 1 {0=0, 1=2}
|   fbs ≤ 92.500: 1 {0=0, 1=61}
```

รูปที่ 7 กฎการจำแนกด้วยต้นไม้ตัดสินใจ

กฎข้อที่ 4 ถ้า hba1c น้อยกว่าหรือเท่ากับ 10.998 ให้ไปดู egfr (อัตราการกรองของเสียของไต) มากกว่า 11.028 ถ้าใช่ไปดู bpd (ความดันโลหิตตัวล่าง) มากกว่า 107 ถ้าใช่ ผลลัพธ์คือเป็นโรคเบาหวาน

กฎข้อที่ 5 ถ้า egfr น้อยกว่าหรือเท่ากับ 11.028 ผลลัพธ์

คือเป็นโรคเบาหวาน

กฎข้อที่ 6 ถ้า egfr มากกว่า 29.349 ไปดู bps ถ้ามากกว่า 100.500 ผลลัพธ์คือเป็นไม่เป็นโรคเบาหวาน เป็นต้น

4. อภิปรายและสรุปผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและเปรียบเทียบประสิทธิภาพของการจำแนกข้อมูลด้วยเทคนิคเหมืองข้อมูล 4 วิธี จากนั้นทำการหาค่าความถูกต้องของแบบจำลอง (Accuracy) โดยใช้วิธี10-Fold Cross Validation เพื่อหาแบบจำลองที่เหมาะสมที่สุดในการจำแนกข้อมูลผู้ป่วยโรคเบาหวาน จากผลการวิจัยสามารถสรุปได้ดังนี้ วิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง คือ 93.73% วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้อง คือ 86.13% วิธีนาอีฟเบย์มีค่าความถูกต้อง คือ 88.92% และวิธีความใกล้เคียงกันมากที่สุด 86.97% อีกทั้งผู้วิจัยได้ทดลองลดคุณลักษณะให้น้อยกว่า 16 คุณลักษณะ ให้เหลือเพียงคุณลักษณะที่มีในผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจ และทดลองกับเทคนิคต่างๆ อีกครั้งหนึ่งพบว่า ค่าความถูกต้องของแบบจำลองลดลง สรุปได้ว่า วิธีการจำแนกกลุ่มที่มีประสิทธิภาพในการจำแนกดีที่สุดสำหรับข้อมูลผู้ป่วยโรคเบาหวานโรงพยาบาลสมเด็จพระยุพราชบ้านดุงคือ วิธีต้นไม้ตัดสินใจ เนื่องจากเป็นวิธีที่ไม่มีการแจกแจงหรือไม่ใช้พารามิเตอร์ ซึ่งไม่ได้ขึ้นอยู่กับสมมุติฐานการแจกแจงความน่าจะเป็น สามารถจัดการกับข้อมูลที่มีมิติสูงได้อย่างแม่นยำ ทั้งนี้ปัจจัยที่เกี่ยวข้องกับคุณลักษณะความเสี่ยงที่นำมาใช้ในการสร้างแบบจำลองมีการใช้ตัวแปรที่มีความสัมพันธ์ในการเกิดโรคเบาหวาน เช่น bps, bpd, fbs, hdl, Creatinine, egfr จากการทบทวนวรรณกรรม [12], [13] พบว่า ปัจจัยเหล่านี้ไม่ได้ถูกนำมาเพื่อประมวลผลในเครื่องจักรการเรียนรู้ แต่ผู้เชี่ยวชาญยืนยันว่ามีผลกระทบทำให้เกิดโรคเบาหวาน [14] ยกตัวอย่างเช่น ตัวแปร egfr หากมีค่าน้อย หมายความว่าอัตราการกรองของเสียของไตเริ่มจะทำงานหนักเป็นสาเหตุอีกอย่างที่ทำให้โรคเกิดเบาหวานและความดันโลหิตสูง หรือ SLE (ภูมิคุ้มกันผิดปกติ) แม้ตัวแปรนี้และตัวแปรอื่นๆ ข้างต้นที่ได้กล่าวมา ไม่ได้มีน้ำหนักมากเท่าตัวแปร hba1c หรือ fbs (ซึ่งส่วนใหญ่จะใช้ตัวแปรนี้ ในการทำนายโรคเบาหวาน) แต่ก็มีค่าน้ำหนักในระดับมาก มีความสัมพันธ์ต่อการเกิดโรคเบาหวานอย่างชัดเจน ผู้วิจัยจึงได้นำผลลัพธ์ที่ได้จากแบบจำลองของกฎต้นไม้ตัดสินใจไปใช้ในการพัฒนาระบบจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน เพื่อเป็นแนวทางในการสนับสนุนการตัดสินใจทางการแพทย์ในส่วนต่อไป

เอกสารอ้างอิง

[1] X. Li, Z. Zhao, C. Gao, L. Rao, P. Hao, D. Jian, W. Li, H. Tang, and M. Li, "The diagnostic value of whole blood lncRNA ENST00000550337. 1 for prediabetes and type 2 diabetes mellitus," *Experimental and Clinical Endocrinology & Diabetes*, vol. 125, no. 6, pp. 377–383, 2017.

[2] WHO and IDF. (2006, November). *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia; Report of a WHO/IDF consultation*. [Online]. Available: https://www.who.int/diabetes/publications/diagnosis_diabetes2006/en

[3] A. Petersmann, M. Nauck, D. Müller-Wieland, W. Kerner, U.A. Müller, R. Landgraf, G. Freckmann, and. L. Heinemann, "Definition, classification and diagnosis of diabetes mellitus," *Exp Clin Endocrinol Diabetes*, vol. 126, pp. 406–410, July 2018.

[4] T. Daghistani and R. Alshammari, "Diagnosis of diabetes by applying data mining classification techniques," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 329–332, July 2016.

[5] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model

based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.

[6] J. Tuomilehto, J. Lindström, J. G. Eriksson, T. T. Valle, H. Hämäläinen, P. Ilanne-Parikka, S. Keinänen-Kiukaanniemi, M. Laakso, A. Louheranta, and M. Rastas, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," *New England Journal of Medicine*, vol. 344, no. 18, pp. 1343–1350, 2001.

[7] K. Faranak, "Type2 diabetes mellitus prediction using data mining algorithms based on the long noncoding RNAs expression: A comparison of four data mining approaches," *BMC Bioinformatics*, vol. 21, no. 1, pp. 372–386, 2020.

[8] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front Genet*, vol. 9, pp. 515–525, 2018.

[9] A. Kemal and S. Baha, "Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms," *International Journal of Modern Education and Computer Science*, vol. 10, no. 6, pp. 10–16, 2018.

[10] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–99, 2013.

[11] V. Vijayan and A. Ravikumar, "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus," *International Journal of Computer Applications*, vol. 95, no. 17, pp. 12–16, 2014.

[12] B. Kakillioglu, R. Sharma, and V. Jindal, "Diabetes determination using retraining neural network," presented at the International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018.

[13] Y. Hayashi and S. Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48 graft combined with sampling selection techniques for the diagnosis of type2 diabetes mellitus in the Pima Indian dataset," *Informatics in Medicine Unlocked*, vol. 2, pp. 92–104, 2016.

[14] W. Bethany, Casey M. Rebholz, S. Yingyin, A. K. Lee, C. Josef, S. Elizabeth, and M. E. Grams, "Diabetes and trajectories of estimated glomerular filtration rate: A prospective cohort analysis of the atherosclerosis risk in communities study," *Diabetes Care*, vol. 41, pp. 1646–1653, 2018.